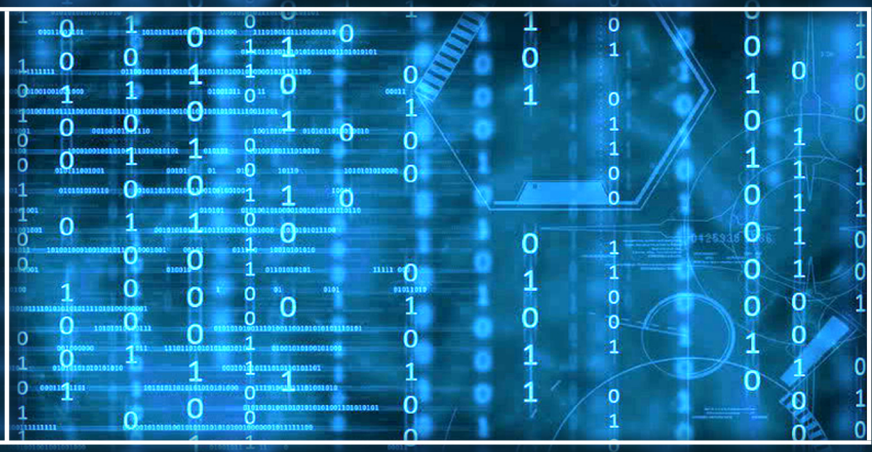
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 12 Issue 9
September 2021



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

SAI
www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Kohei Arai
Editor-in-Chief
IJACSA
Volume 12 Issue 9 September 2021
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Editor-in-Chief

Dr. Kohei Arai - Saga University
Domains of Research: Technology Trends, Computer Vision, Decision Making, Information Retrieval, Networking, Simulation

---

Associate Editors

Alaa Sheta
Southern Connecticut State University
Domain of Research: Artificial Neural Networks, Computer Vision, Image Processing, Neural Networks, Neuro-Fuzzy Systems

Domenico Ciuonzo
University of Naples, Federico II, Italy
Domain of Research: Artificial Intelligence, Communication, Security, Big Data, Cloud Computing, Computer Networks, Internet of Things

Dorota Kaminska
Lodz University of Technology
Domain of Research: Artificial Intelligence, Virtual Reality

Elena Scutelnicu
"Dunarea de Jos" University of Galati
Domain of Research: e-Learning, e-Learning Tools, Simulation

In Soo Lee
Kyungpook National University
Domain of Research: Intelligent Systems, Artificial Neural Networks, Computational Intelligence, Neural Networks, Perception and Learning

Krassen Stefanov
Professor at Sofia University St. Kliment Ohridski
Domain of Research: e-Learning, Agents and Multi-agent Systems, Artificial Intelligence, e-Learning Tools, Educational Systems Design

Renato De Leone
Università di Camerino
Domain of Research: Mathematical Programming, Large-Scale Parallel Optimization, Transportation problems, Classification problems, Linear and Integer Programming

Xiao-Zhi Gao
University of Eastern Finland
Domain of Research: Artificial Intelligence, Genetic Algorithms

CONTENTS

Information Flow Control for Serverless Systems

Rishabh Chawla
The Pennsylvania State University
May 2020

***Abstract*—Security for Serverless Systems is looked at from two perspectives, the server-level security managed by the infrastructure company and the Application level Security managed by the tenants.The Trusted computing base for cloud systems is enormous as it encompasses all the functions running on a system. Authentication for systems is mostly done using ACL. Most Serverless Systems share data and thus, ACL isn't sufficient. IFC using appropriate label design can enforce continuously throughout the application. IFC can be used to increase confidence between functions with other functions and cloud provider and also mitigate security vulnerabilities making the system safer. A survey of the present IFC implementations for Serverless Systems is presented and system designs which are relevant to Serverless Systems and could be added to Serverless Systems Architecture and, an idea of an IFC model that could be effectively applied in a decentralised model like serverless systems.**

***Keywords*—Information flow control; serverless systems; language based security; cloud computing**

I. Introduction

Serverless Systems are systems where functions owned by tenants are executed when the functions are triggered, on platforms managed by the cloud provider. The infrastructure, security and updates of these systems along with the hardware are managed by the cloud provider and the tenant only manages their function and it's security. Security is looked at from two perspectives, the server-level security managed by the infrastructure company and the Application level Security managed by the tenants. The reasons for the boom of serverless computing are elastic scalability, ease of deployment, and flexible pay-per-use pricing. Trusted computing base(TCB) consists of all the parts of the system (like hardware, software, libraries, firmware), all the components which could leave the system vulnerable and jeopardize the security of the whole system. The TCB for cloud systems is enormous.

A. Server-Level Security

Serverless Systems are cost-effective resource sharing platforms where the tenants only pay for the time their function/service is executing/working on the machines and thus, the machine setup time for a function i.e the microVM/container creation and startup time has to be minimal, as that time is paid for by the cloud provider, which does not leave a lot of scope for setting up security measures specific for functions by the cloud provider. The cloud provider's platform manages function placement and scheduling, automatically spawning new function instances on demand. This also means that multiple functions are run on the same server, the functions owned by different teams/companies with no security guarantees to each other, which leaves the possibility for side-channel attacks, and attacks specific to those applications/services which might leave the host machine vulnerable. Traditional security practices are unable to achieve the flexibility, generality and efficiency expected by cloud providers and tenants [1].

B. Application Level Security

Users express their applications as collections of functions triggered in response to user requests or calls by other functions. With serverless systems the use of third party services has increased which in turn increases the risk of data vulnerability during communication, the security of the third party services, the storage of keys used for communication with the services. [2] Applications need to consider security from the perspective that they are vulnerable to exploits of the third party apps, infrastructure, other tenants sharing the system, among others. Serverless Systems removes the burden of managing Server Level Security for the Application Development Teams as most tenants believe the cloud provider they are using, though measures could be taken to increase this confidence.

C. Paper Outline

Section 2 describes the Background on some of the different parts used in serverless architecture and basic idea of attacks specific to serverless systems. Sections 3,4,5 are the motivation to use IFC on serverless systems and explain the need for IFC and the advantages it could bring. Section 6 explains some IFC ideas which could be applied to different parts of the serverless architecture. Section 7 describes some serverless, cloud and general IFC implementations. The cloud and general implementations are on system parts which are part of the serverless architecture and could be modified to work on serverless architecture. Sections 8 and 9 explain the advantages and remaining questions after adding DIFC to serverless architecture. Section 10 is my idea of how all the mentioned ideas and implementations could be implemented together to setup a serverless DIFC system. Section 11 explains some Future Research ideas/direction.

II. Background

Many functions are run on a single bare-metal machine in Serverless systems, to improve security a virtual machine or container is used to execute the function, so two functions are basically running on virtualized environments rather than on the bare-metal machine itself and thus, separated by an extra layer of abstraction and thus, increasing security and making it harder for the functions to affect or read each other's information. We explain some exploits and mitigation techniques used in serverless systems.

A. Containers

Functions are hosted inside containers as containers encapsulate all the underlying software required for the application to run and are useful when applications need to run on different environments/machines. Containers fall short as they use the host operating system kernel, which means that there is a fundamental trade-off between security and code compatibility as, Container implementors can choose to improve security by limiting syscalls, at the cost of breaking codes which require the restricted calls [3].

B. MicroVM

MicroVM are minimilistic virtual machines. The idea behind microVM's were to protect against privilege escalation, information disclosure, covert channels and others. With microVMs which take less than a second to boot up, the security measures are enhanced. Adding advanced security features affects the performance of the system and might not be implemented by tenants in lesser security demanding environments [3]. To give an idea, Firecracker checks that the host kernel has mitigation enabled for Kernel Page-Table Isolation, Indirect Branch Prediction Barriers, Indirect Branch Restricted Speculation and cache flush mitigation against L1 Terminal Fault among others. [3]

C. Scheduling or Warm Containers

Each function invocation should ideally take place in a fresh environment, such as a container that is immediately destroyed after it's execution but to reduce the cost of setting up an entire runtime environment for each function execution, warm containers are cached and reused for future invocations of the same function within a pre-configured timeout window [4]. Opaque platform policies and scheduling algorithm details obscure this practice, making it difficult for customers to account for such issues during application development. Attackers can get their function on the same machine if enough functions are deployed [5].

1) Device Drivers: VM's use paravirtualised device drivers which interact directly with the VM host via an agreed channel. The alternative to this is a way slower virtual hardware using the native device drivers. Cloudburst describes the vulnerability in VM display functions of VMware Workstation that could be exploited by a video file to take over the operating system [6].

D. Hypervisors

Hypervisors are used to create virtual machines on bare-metal machines. KVM is a virtualization module for linux and is a type 1 hypervisor. It is used in Firecracker. Virtunoid is a privilege escalation exploit on KVM made in 2011 because of a missing check on the KVM emulation of PCI device hotplugging, which is used for devices which don't support being unplugged but when unplugged left a corrupt state and dangling pointers [7]. These kind of vulnerabilities are being mitigated by Hypervisor verification [8].

E. Attacks on Serverless Systems

Serverless systems contain functions which generally last seconds, thus, it is harder to attack them but there are exploits made for this kind of system. Rapidly ex-filtrate stolen data [9], cross-tenant side-channels [5] are some attacks for this type of system. Persistent function compromise is possible by malware in an in-memory partition of the system, is another example of an attack for this system. Attackers can also take advantage of the cloud providers warm container reuse policy to cache a compromised copy of the function that persists across invocations [9]. Logging and debugging support in serverless platforms lacks the ability to monitor a serverless application as a whole and therefore struggles to trace sophisticated attacks, for example an attack that depends on two executions of the function. [10]

III. Present Security Solutions for Serverless Systems

Present security solutions include language run-time libraries which are used to secure a single function according to developer defined policies as part of the source code. Static analysis of function source code could be used to detect violations of the principle of least privilege [11] and checking function dependencies against vulnerability databases [12]. Function developers rarely consider and secure interactions between functions, giving rise to emergent attack vectors such as API-based data exfiltration. There are products that model function behavior using machine learning to detect anomalous behaviors or wrap function event handler wrappers to inspect specific activities [13]. There are also run-time protections which include machine learning based detection of anomalous function behaviors [13] to prevent event-data injection prevention by inspecting incoming function invocation requests using existing penetration testing techniques like sqlmap. Runtime semi-automated troubleshooting based on log data, to ease reasoning about function behavior is present to make auditing easier [14].

IV. Lack in Present Security Solutions for Serverless Systems

A lot of pre-compiled third-party objects and proprietary closed-source functions don't provide source code access which is required by many of the present security techniques. The present security solutions are largely function-centric and their efficacy depends on the correctness of policies written by the function developers, complete access to source code and configuration files, and the compatibility of the tool with the functions specific language runtime, platforms, and event sources. Existing monitoring techniques offer limited observability into the interactions between functions and most of these monitoring services are limited to strict specified/available conditions. [13] Static check tools aren't able to detect implicit flows. Cross invocation attacks [13] aren't considered by present security solutions which occur between containers and also among reuse of containers (warm starts). A major and serverless specific problem is lack of proper function isolation [10]. Event injection attacks may target the function source code which might also leak other secrets stored in the container [15]. Azure Functions had an exploitable placement vulnerability, which led to the exploit to run arbitrary binary

code in containers making them vulnerable to many kinds of side-channel attacks [5].

V. Need for Discretionary Access Control

Major authentication check in today's world is done using Access Control List(ACL), also called Role-Based Access Control(RBAC). ACL has some limitations and vulnerabilities which can be fixed using IFC. It may be possible to bypass ACL checks, especially in web-based systems [16]. ACL does not implement any further control once the data has been authorized at entry point or discrete check point by checking the users allowed permissions. The application is trusted not to leak the data after the check. As there is a lot of data sharing among applications there is a need for controlling data flows between applications which may also send data ahead and these checks can be done using IFC [17]. Data can propagate or influence system behaviour indirectly in ways that aren't disclosed, which access control barriers at discrete points in code do not detect, while IFC using appropriate label design can enforce continuously throughout the application [18] IFC [19] could be used to add security policies to data and use these policies at run-time to control where user data flows. Since IFC security is linked to the data that it protects, both tenants and cloud providers can agree on the security policy, in a manner that does not require them to depend and rely on the particulars of the cloud software stack or application stack in order to effect enforcement [1]. IFC [20] provides a means to control and monitor data flow continuously, according to policy which could restrict that the data be restricted to a certain location in favour of laws [1]. IFC mechanisms can help enforce non-interference policies mitigating the fact that another system running on the same machine may observe the public outputs. IFC supports isolation of individual users data, and inter-tenant isolation [1]. IFC protects information by a global security policy that cannot be overridden by a misconfigured application. The policy explicitly and concisely captures constraints on end-to-end information flow through the system, majorly protecting the system by system calls and restricting the data flowing outside the network. The IFC system enforces the policy even for buggy or malicious applications, thus removing application code and configuration from the TCB of the cloud [21]. This case is valid when the right policy check parameters are set inside the application as used in the cloud infrastructure. A generic model to detect type could be made which could make this model stronger. Specifying an effective security policy is a difficult problem, failure to adequately restrict flows violates the principle of least privilege and leaves the system vulnerable but defining overly-restrictive rules prevents the correct operation of the system, thus increasing the development and testing time and requires checking all expected flows [22].

VI. IFC System Design with Serverless

This section contains general ideas which people have mentioned in regards to Cloud Computing/Serverless Systems and given a general idea of how IFC could be useful in solving them.

A. Warm Starts

Container creation accounts for a major chunk of time in the response time for a function after it was called/triggered and cost for container creation time is not paid by the tenant and is covered by the cloud provider and thus, cloud providers use warm starts which is reusing the container which was recently used to run the function, so that on another function call of that same function in a certain time limit, the container is reused rather than creating another container, so that the response time for the function is reduced and the cost for the container creation doesn't need to be paid. The cloud provider kills a function after a certain time limit as the tenant only pays for the time when the function is being used and not when it is idle and waiting for a request. It is expected that a serverless function activation handles a single request on behalf of a specific user and only accesses secrets related to this request. Each invocation starts from a clean state and does not get contaminated with sensitive data from previous invocations. Any state shared across invocations must be kept in a global data store. Warm startup is done using the method that after the initial invocation is complete, but before the actual function process starts, the process is forked and the function is run on a child process of the same process and purged after it's completed and this process is repeated again when the function is called again and thus, another child process runs it. This way the address space is in the child space and will not affect other processes that are run on or from it [21]. This way could be vulnerable when two child processes are running at the same time, as there isn't a lot of separation in that case, but in serverless system this isn't done. The child process would have to strictly be limited to it's address space as a this could be used for cross invocation attacks. Another way is tainting the sensitive data which is used for file access, and deleting all the tainted data after the function execution ends before the next function is executed on the same container [13]. We could also taint all the changes made to the filesystem during the function execution and revert them using something like a git svn or snapshot but that would have a higher overhead, so one could use tracking on all the changes on the filesystem which uses more processing power and could increase the execution time of the function, the time could be reduced by using even more processing power. One of these two methods could be used based on the trade-off of the time it takes after the execution of the function compared to the other one taking extra processing during function execution.

B. Termination

A container is created everytime a function is called/triggered and thus, for serverless systems the termination of a function can be as many times the function is called which is generally a lot. IFC Systems which leak information through the termination channel, where one bit of information can be observed by observing the termination or non-termination of the program. The parallel nature of the serverless environment amplifies this weakness, allowing the attacker to construct a high-bandwidth information channel, effectively defeating the purpose of IFC [21]. The termination channel present in most existing IFC systems can be arbitrarily amplified via multiple concurrent requests, requiring a stronger termination-sensitive non-interference guarantee, which can be achieved using a combination of static labeling of serverless processes and dynamic faceted labeling of persistent data [21]. We can use the security property termination-sensitive non-interference (TNSI) to

eliminate this channel [23]. SLam Calculus is used in the TNSI security property to achieve termination insensitive IFC model [24]. A way of achieving this in serverless systems is a combination of static program labeling with dynamic labeling of the data store, based on a faceted store semantics. Static program labeling restricts the sensitivity of data a serverless function can observe ahead of time, and is used to eliminate the termination channel. Dynamic data labeling is important to secure unmodified applications that do not statically partition the data store into security compartments, while the faceted store semantics eliminates implicit storage channels also [21].

C. Sticky Policies

In serverless systems functions are short lasting and can be run on any machine at any time, and a lot of them could be running at the same time triggered by different users. Data could be required for these functions to run which they get from databases or filestores (S3 and dynamo for AWS). At a higher level, sticky policies could be used to achieve end-to-end control over data. In sticky policy systems, data is encrypted along with the policy on that data. To obtain the decryption key from a Trusted Authority (TA)(In this case the database of filestore), a process must agree to enforce the policy. This agreement may be considered part of forming a contractual link between the data owner and the process decrypting the data [25]. A logging system of the decryption could represent a starting point of data flows and can be used for tracking.

D. Continuous Checks

Serverless systems applications are made using a combination of functions where the data flows from one function to another, and may also flow to third party functions for various reasons like verification and so on. In a system like this where data flow continues and the limitations of the data flow inside the intranet or outside isn't known, Continuous data checks could be used to check that data is only used at authorized places. This is done by storing the data with its label and any time the data needs to be accessed its label is checked with the process label and only if permitted, read/write operations on the data are permitted. This system could be implemented by the function sending the data, by checking that the function receiving has enough access, the storage location would also check the same. The function receiving could use the label of the function they got the data from for confidence label on that data. Thus, having security perspective by the one sending and also, the one receiving the data. A function call is given the label of the caller and actions allowed to the caller are only permitted to that execution. Similarly any data store being modified also stores the label with which it was modified and stores allowed to a particular label are read or written by that label. If any operations need to be done outside of the permitted value of the caller, declassification [26] is used.

E. Implicit Storage Channel

A serverless function always runs on behalf of a specific user and can be assigned a corresponding security label. The function's label determines its view of the data which it reads or writes in databases or filestores, the function can only observe the existence of data whose label does not exceed the function's label. In a situation where multiple functions with incomparable labels write to the same store location(database or filestore). We avoid information leaks in this situation by employing faceted store semantics, where each record can contain several values with different security labels [27]. Implicit storage channels is when the attacker infers secrets by observing the labeled values exist within particular store locations without observing the actual values [28]. An attacker could check that a location contains sensitive data by writing to a particular location and reading from there. One could block writing to that location but that would also leak that it contains sensitive data. One alternative would be to have extra data copies for different labels. This is shown by data store semantics where each record can contain multiple values. Though this has a high runtime cost [29].

F. Audit Logs

If an IFC system is made at the cloud level, including the network, OS and continuing to the application/function level, all this data flow can be used as a logging system. Enforcement of IFC can provide the opportunity for recording flow decisions to build a provenance like audit graph. This can be analysed to understand where, how, why and by whom the data was manipulated within the system. This audit data, captured during IFC enforcement, can help to demonstrate compliance with regulations by providing tangible traces, showing how the data was handled [30]. Under a conservative assumption that all secrets obtained during function execution propagate to all its outputs, we can track the global flow of information in the system by monitoring inputs and outputs of all functions in the system [21]. Audit Logs are made by tracking all information flow using tainting. An important detail is to get all the information before logging which would be necessary for analysing, thus, the point where the logs are stored needs to be as late as possible to get the most logs and we also need to consider the high performance penalty cost for it and minimize it. Major challenges with this system is being able to track all information flows and the logs being enough to recreate the situation or analyse the situation completely [31].

VII. IFC Implementations for Serverless Systems

A. Hardware Level - General Implementation

RIFLE [32] translates normal binary code to run on hardware with IFC tracking. Dynamic information flow tracking can be used at this level for checking the use of spurious values being used as instructions or pointers [33]. In serverless systems this method any attacks on the underlying hardware where the calls are sent to using the virtual machine can be checked.

B. Kernel Level - Cloud Implementation

Information flows in a system are only generated through system calls and shared memory between processes. If shared memory is restricted then information flows could only be generated using system calls. The entities defined in this model are processes, files, pipes and sockets. Privileges are only associated with processes(active entities). All labelled entities are allocated their labels when they are created. For a process

creating some entity the sub-rules associated with the flow are that the created entity inherits the labels. Certain processes have privileges, allowing them to change their labels that is, those processes are able to change their security context using declassification. The labels of passive entities (files, pipes and sockets) can't be changed. Processes are further associated with privileges over their tags. System calls creating flows are intercepted and IFC constraints are applied, enforcing IFC according to the labelling, other system calls are left unintercepted. The cloud tenant decides the labels and tags for processes and calls [34]. This case is valid in serverless systems as we run only one function inside a microVM and if there is any information flow between functions or third party services it has to go outside the function/microVM using system calls, as nothing else is running on the microVM thus, no shared memory among usage. This system can be used to restrict any malicious code injected from sending data outside the microVM and thus, restricting the outflow. The access to the microVM is protected and restricted by the cloud provider.

C. VM Level - General Implementation

Argos [35] modifies the QEMU(type-2 hypervisor which can be combined with KVM to make a type-1 hypervisor) virtualisation framework to extend the target code so that it defines isolation regions and checks information flow meta-data. It uses dynamic taint analysis to detect exploits and protects unmodified operating system processes. It checks the network data throughout execution to identify their invalid use as jump targets, function addresses, instructions. It has a very high overhead but could be used to find signatures which can be used with almost no overhead to find exploits during run-time [35]. This system is used to protect the VM itself so that the VM runs and terminates smoothly. This system could be used to confirm that the VM is restricted to the allowed and approved capabilities. Running this system on actual machines doesn't seem feasible because of the overhead but a system like this is very useful to actually analyse the new attacks that are being created and get their signatures to block them on actual running machines.

D. OS Level - General Implementation

If IFC is enforced at OS level, the applications running above the OS, run under the policy constraints expressed by the IFC labels tags. They do not need to be trusted not to leak data through the monitored labels in the processes [36]. This system has DIFC implemented at the granularity of processes, and integrates DIFC controls with standard communication abstractions such as pipes, sockets, and file descriptors via a application level reference monitor. This interface helps programmers secure existing applications. This system enforces the DIFC policy during runtime. The application consists of two types of processes. Untrusted processes are generally used for most of the work. They are constrained by, and maybe unaware of the DIFC controls. Trusted processes are aware of DIFC and setup the privacy and integrity controls that constrain untrusted processes. Trusted processes also have the privilege to selectively violate information flow control for example, by declassifying private data, or by endorsing data as high integrity. The system represents each resource a process uses to communicate as an endpoint, including pipes, sockets, files, and network connections. A process can specify what subset of its privileges should be exercised when communicating through each endpoint. Uncontrolled channels are modeled as endpoints that exit the DIFC system. This can be used as a security policy for end-to-end integrity protection. It can pull third-party plugins into its address space, but with end-to-end integrity protection, users can enforce that selected plugins never interact and potentially corrupt sensitive data, either on input or output [36]. This system can also be used by the cloud provider to restrict the usage of the functions as the microVM, the OS and underlying hardware is managed by the cloud provider. The cloud provider can have different levels setup and the functions can decide the amount of freedom they require and the cloud provider could isolate categories of functions in their own DMZ (Demilitarized Zone).

E. Network Level - Serverless Implementation

This system has agents residing in function containers to monitor storage and network behaviors. These agents dynamically generate taint labels that describe each function's file accesses and network requests. These labels are reported to a centralized controller. The controller then aggregates this information to discover the flow paths of the application. The information flow monitored by the controller can be restricted based on the security policy. Function calls are monitored with taint labels that and called/triggered using REST based APIs in this system. The system works with deploying a transparent forward proxy in each container that begins proxying network requests when the container starts. The network proxy performs network level tainting. The proxy inspects the REST call to determine appropriate labels with which to taint the current flow, which are mostly mentioned in the Rest call itself, and are compared with the policy file. Each invocation request is a unit of work in FaaS, functions are short-lived and taint labels are assigned per request. The "taint explosion" problem occurs because of this. The network taints can be summarized when a function makes multiple calls to the same domain, thus compressing the taint labels. Function level operations are checked across workflows to capture inter-function security violations. There can be Data leak through the network as Static network policies are bypassed by passing data to downstream functions with network access. This can be mitigated with Network level taint tracking. Another type of attack could be the Cross invocation side channel where residual data in warm containers is leaked across invocations and this can be mitigated using File access taint tracking and function garbage collection [13]. A similar design is mentioned in [37] where labels are used on data and communication between libraries is done using messages and based on the level of the process any data in the message above that label level is removed. Continuing on the previous system, the agent contains a system call tracing mechanism for monitoring function file I/O, allowing the system to detect cross-invocation flows resulting from container reuse. After the function finishes execution, all data on disk that was modified by the function is erased from the container. As current attacks require an explicit data flow from one function execution to another this procedure is sufficient to deny cross-invocation capabilities to the attacker. Commercial platforms provide only a small writable partition using an in-memory filesystem, the approach is significantly more efficient than

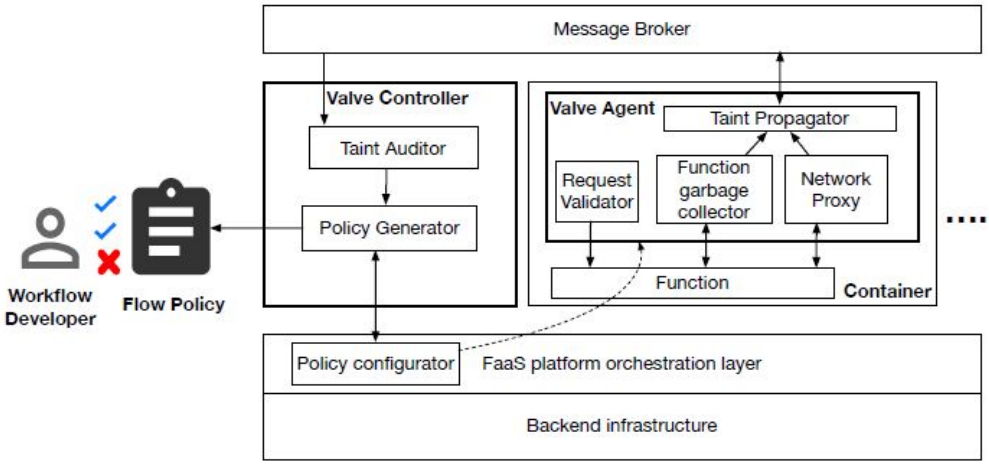


Fig. 1. Network Layer Design Architecture Presented in [13].

destroying and re-provisioning the entire container. The agents also check taint labels on file accesses and file access behaviors as they may violate security constraints of the application. Garbage collection is performed after each function invocation, the set of modified files are not tainted by previous function executions [13] (Fig. 1).

F. Service Level - Cloud Implementation

The system's Model generator component is responsible for building a model that semantically simulates the lifecycle or runtime execution of the candidate application. Then, the IFC engine performs information flow analysis on the unmodified application's bytecode with the help of the generated model. The vulnerability detector pinpoints insecure flow paths that violate data integrity and confidentiality. The result publisher component refines and reports the analysis results to the cloud provider and tenant. Based on the results, it decides if a security certificate should be granted to the candidate application, which is sent to both the cloud provider and tenant [38] (Fig. 2). The system uses IFC based on System Dependence Graph (SDG) and program slicing techniques for security inspection. The SDG has the advantage to model the information flow through a program by capturing both data and control dependencies [39]. Even with this model, the system is prone to stealth type attacks. This system can be added to serverless systems as a separate utility which could be used when any abnormality is detected to check the request, because this system has a very high overhead.

G. Application Level Implementation on Hadoop using Inlined Reference Monitor (IRM) - Cloud Implementation

IRM [40] implementation carefully leverages object encapsulation, control-flow safety, and type-safety properties of the binary language in which the function code is expressed, to guarantee that the surrounding untrusted code into which the IRM is in-lined cannot corrupt or circumvent the IRM's security programming at runtime. IRM whose programming is in-lined into untrusted binary jobs as they arrive at the cloud edge. After in-lining, the modified jobs self-enforce the security policy. The in-lined enforcement code maintains and consults an information flow graph (IFG) implemented as a distributed data resource within the cloud. The IFG tracks information flows between the various principals, and the IRM prohibits job operations that introduce explicit flows that violate any defined policy. This makes it easy to implement and adapt to real world clouds, since the cloud and the enforcement can be maintained completely orthogonally. It achieves this by enforcing an IRM that is in-lined into untrusted binary jobs at the cloud's edge. The resulting jobs self-monitor their accesses and collectively maintain a distributed information flow graph within the cloud, which tracks the history of flows and prohibits policy-violating operations. Well-established IRM design methodology is applied to secure the IRM against attacks from the code into which it is in-lined, protecting it even from threats that know all the IRM's implementation details. This system is limited by enforcement of mandatory access controls of explicit information flows between principals [41]. A system like this is hard to implement in serverless systems as the functions last under a second and aren't running all the time, so IRM would have a heavy startup overhead and wouldn't be very useful as it isn't running all the time, else it'll have to converted to a system which stores its state in a database and retrieves it everytime it starts up. I think that the cost and performance overhead would outweigh it's benefits. A system like this could be implemented at the cloud infrastructure level where the cloud enforces some principals and based on environments which need more security could have more/stricter enforcement which would increase the processing and runtime of the function and thus, the cost. This system will be useful at a cloud level for applications deemed dangerous by the cloud and yet requiring a lot of privileges. This system could be used to limit these applications from a moral and legal perspective. The idea of this system is to enforce a system that does not believe the cloud infrastructure. We could consider an implementation of a web-based application, where the front page is always running on a low system with an IRM system enabled and every other process is done using serverless systems, and thus, the scaling and running of all the serverless systems will be managed by the cloud infrastructure itself and these functions, would connect and interact with the

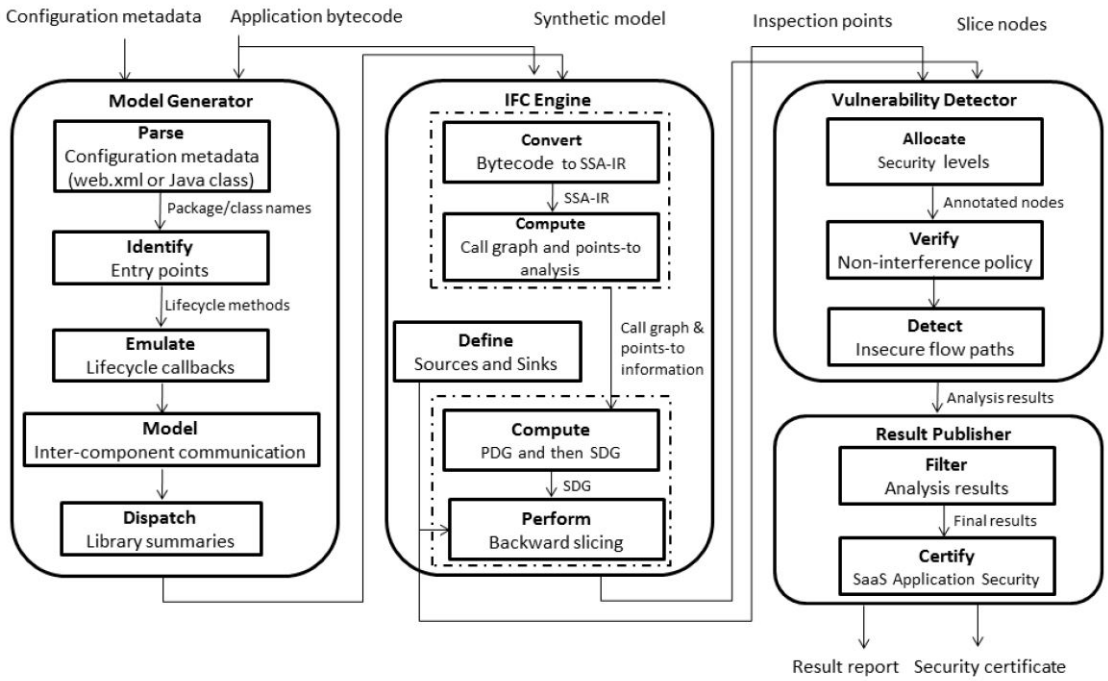


Fig. 2. Service Layer Design Architecture Presented in [38].

IRM system, thus, creating their own IFC system and enforcing security in this untrusted cloud infrastructure. This kind of system can be very useful for banks or similar institutions.

H. Chinese Wall - General Implementation

Chinese Wall Security Policy [42] makes use of subjects and objects to prevent information flows which cause conflict-of-interests between tenants. Data is divided into conflict categories (conflict-of-interest class) and subdivided into sub-divisions based on their profile. When data from a sub-division is accessed, all data from that sub-division can be accessed but data from the broad conflict class can't be accessed anymore but data from other broad conflict classes can be. Access to data is constrained by what data the subject has already accessed. All subjects are allowed to access at most one data-set which belongs to a same conflict-of-interest class. A subject can freely access any object in the sanitized security group, which is a conflict category.(data which doesn't need restriction). [43] For scaling purposes we could use a decentralised Chinese wall mechanism mentioned by Minky. [44] The decentralised mechanism uses Law Governed Interaction mechanism where authorization is required before accessing any resource. Chinese-Wall Process Confinement (CWPC) could be used for practical application-level distributed coalitions that provide fine-grained access controls for resources and that emphasize minimizing the impact on the usability [45]. The centralized system can be applied to serverless systems by classifying the functions using conflict categories and using that strategy to allocate VM's. Thus, there is lesser conflict of functions on the same VM. The CWPC way can be used to check for data request access on the functions from the service provider and the validation method which can be used based on the Minky Law Governed Interaction way where functions have to authorize the usage of data by the other functions if there is an conflict. This is an idea of to increase trust between functions and their working together with conflicts. This system requires a lot of trust by the functions to the cloud provider and the cloud provider needs to maintain strict security policies so that no information gets leaked as the cloud provider will have a lot of sensitive information with this method.

VIII. Advantages

DIFC [46] will be useful on serverless systems for managing and securing information flows both within and between virtual machines and, the overall flow within the cloud. Continuous check of data at every usage point will prevent data leakage and unauthorized use. It would also allow the applications to define their own independent security terminology dynamically [1]. IFC tracks all data flows in order to detect policy violations, it can be used to provide detailed logs for audit purposes [1]. The IRM system makes it possible for systems like banks to switch to public cloud, still having and enforcing their own high level of security. The warm container method increases confidence and decreases security vulnerability with an IFC implementation.

IX. Trade-off

- Data sharing between virtual machines could be done through the intranet having IFC enabled on the network level of the cloud system or have the data sourced through a secure system where it gets authorized. Using a secure system for authorization would reduce the burden on the developer but will only leave a generic check mechanism on IFC and possibly a higher overhead.

- There is a higher development and design time for DIFC integration which can help reduce the security vulnerabilities for the systems integrating themselves with the security policy.
- Implementing a DIFC model at different levels in the cloud system where it is integrated with the other levels and set by the cloud provider, one of the main advantages could be the network analysis which restricts the flow of sensitive data outside the cloud intranet. We can also restrict the flow of data between functions with labelling but the problem here would be that on a real world scale which have billions of functions and a very vast and diverse infrastructure, a lot of labels would be required for some part of data from a label to pass to another and restrict to others/some specific ones.
- Flow of data received by a function can be blocked by blocking transitivity, but it might break the functionality if it is required, so transitivity is also a parameter that would have to be considered with labels. With a vast network, the checking of this data at every point would be a heavy overhead apart from managing the data and where it could flow. This would in turn increase the development time and one would have to continue checking till it reaches the end point, checking all allowed use cases, adding extra labels which are allowed to use its data and adding restrictions based on when that data flow is supposed to stop.
- Updating the infrastructure or functions would affect each other and has a very high chance of breaking functionalities, thus, upgrading would have to be in phases where both phases are working at a time till everyone moves to the new model. So the effort for updating is increased as there can't be fast updates and any update would force every function that it depends on and those functions that depend on it to be updated and tested.
- Onboarding of new personal would take a lot of time to this system.
- Taint tracking systems would not work if the developer tried to evade them. Using channels outside of the policy are known as covert channels [47].
- Setting up a system like this for a cloud provider would take a lot of effort and money, which wouldn't necessarily result in an increase in revenue but would lead to more confidence by the tenants in the cloud system. Tenants setting up their security from a cloud provider they don't trust would require very heavy security and is almost impossible without set models and to only setup for one particular application. Thus, the feasibility of this system is forced on people making a generic model which is adapted and updated by the tenants and cloud providers like other services.
- The cloud provider will have a lot more sensitive and analyzed data because of this system, thus, the cloud provider keeping this data secure and not exploiting this data would require legal implications so as to keep the cloud provider in check and regular audits by a central authority which regularly audits the cloud system and the data used for the cloud auditing and confidence that it hasn't been tampered with can be gotten by this(DIFC) system itself.

X. Implementation

A nontermination sensitive DIFC can be setup on the cloud which is integrated into the network level, OS level, only allowing authorized system calls checked by the cloud provider and the specific tenant receiving the request based on their request policies. Using the VM level implementation mentioned above securing itself from the bare-metal machine attacks. A general norm of label-set which are configured in the cloud system and could be extended by the application, along with an intranet DMZ set which blocks the flow of data outside the DMZ unless it contains authorized labels. DIFC having library extensions(eg boto3 by amazon for AWS for python language) which be imported and extended to the applications. A system like this could have checks from the starting point i.e the REST call till the end of point where the data will flow and could also have legal limitations which check the data based on location among others. All third-party plugins into each function are used with end-to-end integrity protection and the functions can define policies so that selected plugins never interact and, potentially corrupt sensitive data or only certain plugins interact with sensitive data. A Chinese wall setup could be used to present and restrict any functions being run on the same machine which have conflicts with each other. There could be a service level implementation where any suspected activity could be checked before sending it to the actual function, as this would have a high overhead but the payment of this system would have to be figured out. Any function which wants extra security could have sticky policies which could be used to log all data decryption, thus always having a log of everybody getting the data at this source and the data is only present here and only gets decrypted through the sticky policy. The overall flow of data is monitored using the IFC system and logged and thus, these detailed logs can be used for verification and checking of the cloud infrastructure, tenants and other functions. The logging level shown will only be for their data and the other data will be obliqued. The system will have multiple data copies for different labels to stop implicit storage channels and with function owner will be notified if any activity which does not follow the policy defined for the storage channels. Warm Startup is present with all sensitive data being removed which was identified by tainting. This system would improve security confidence between functions and also with the cloud provider. Blockchain methodology could be added to the Audit log to show that it hasn't been modified and present confidence in the Audit Log.

XI. Future Research Direction

- A Dynamic IFC Model for serverless systems is made by [21] in which the main points are described above in the Network Level Section. The model has inherent assumptions where almost all of the TCB is considered safe and data integrity isn't considered. Reuse of containers and warm starts or multiple invocations to the same function aren't considered. The system could be extended to include all these points.

- With serverless systems a lot of the services required are outsourced, for example the authentication service for AWS i.e AWS Cognito is used in a lot of serverless functions rather than using their own. What kind of extra risk does outsourcing things bring and how does IFC mitigate it is another field where more information with some data for proof is needed, the ideas are expressed in this paper.
- The monitoring of function requests moves the stateless architecture to a stateful architecture. How could we implement IFC while keeping a stateless architecture.
- A computation layer based on label, isolated in a DMZ where only the public output is allowed to leave the overall virtual DMZ network, so all private computation is done inside.
- An cloud implementation which presents the model with an norm based label structure and doesn't have a heavy overhead on performance and time compared to the present system. A prototype proof of concept for the real world.
- An implementation of a model described above could be a starting point with all the system design features mentioned considered. The implementation shouldn't consider TCB to be safe.

XII. Conclusion

A survey of the present IFC implementations is presented and system designs which are relevant to Serverless Systems that could be added to Serverless Systems Architecture and, an idea of an IFC model that could be effectively applied in a decentralised model like serverless systems. The overall idea of this paper assumes that all of the TCB is vulnerable and gives implementations/ideas which could be used to increase confidence between functions with other functions and cloud provider and also mitigate security vulnerabilities making the system safer.

Acknowledgment

I thank Danfeng Zhang for his helpful suggestions on early drafts of the paper.

List of Abbreviations

ACL: Access Control List
DIFC: Decentralized Information Flow Control
DMZ: Demilitazied Zone
IFC: Information Flow Control
RBAC: Role-Based Access Control
TA: Trusted Authority
TCB: Trusted Computing Base
TNSI: Termination-Sensitive Non-interference
VM: Virtual Machine

References

[1] Jean Bacon, David Eyers, Thomas FJ-M Pasquier, Jatinder Singh, Ioannis Papagiannis, and Peter Pietzuch. Information flow control for secure cloud computing. *IEEE Transactions on Network and Service Management*, 11(1):76–89, 2014.

[2] Alexander Posashenki. How serverless is chanding security: The good, bad, ugly, and how to fix it. https://distillery.com/blog/serverless-is-changing-security/, 2019.

[3] Alexandru Agache, Marc Brooker, Alexandra Iordache, Anthony Liguori, Rolf Neugebauer, Phil Piwonka, and Diana-Maria Popa. Firecracker: Lightweight virtualization for serverless applications. In *17th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 20)*, pages 419–434, 2020.

[4] Tim Wagner. Understanding container reuse in aws lambda. https://aws.amazon.com/blogs/compute/container-reuse-in-lambda/, 2014.

[5] Liang Wang, Mengyuan Li, Yinqian Zhang, Thomas Ristenpart, and Michael Swift. Peeking behind the curtains of serverless platforms. In *2018 {USENIX} Annual Technical Conference ({USENIX}{ATC} 18)*, pages 133–146, 2018.

[6] Kostya Kortchinsky. Cloudburst: A vmware guest to host escape story. *Black Hat USA*, 19, 2009.

[7] Nelson Elhage. Virtunoid: A kvm guest- host privilege escalation exploit. *Black Hat USA*, 2011, 2011.

[8] Dirk Leinenbacih and Thomas Santen. Verifying the microsoft hyper-v hypervisor with vcc. In *International Symposium on Formal Methods*, pages 806–809. Springer, 2009.

[9] Rich Jones. Gone in 60 milliseconds intrusion and exfiltration in server-less architectures. https://media.ccc.de/v/33c3-7865-gone_in_60_milliseconds/, 2016.

[10] Ioana Baldini, Paul Castro, Kerry Chang, Perry Cheng, Stephen Fink, Vatche Ishakian, Nick Mitchell, Vinod Muthusamy, Rodric Rabbah, Aleksander Slominski, et al. Serverless computing: Current trends and open problems. In *Research Advances in Cloud Computing*, pages 1–20. Springer, 2017.

[11] Avraham Shulman, Ory Segal, and Shaked Yosef Zin. Methods for securing serverless functions, January 3 2019. US Patent App. 16/024,863.

[12] Develop fast, stay secure. https://snyk.io/.

[13] Pubali Datta, Prabuddha Kumar, Tristan Morris, Michael Grace, Amir Rahmati, and Adam Bates. Valve: Securing function workflows on serverless computing platforms. *International World Wide Web Conference Committee (IW3C2)*, 2020.

[14] Johannes Manner, Stefan Kolb, and Guido Wirtz. Troubleshooting serverless functions: a combined monitoring and debugging approach. *SICS Software-Intensive Cyber-Physical Systems*, 34(2-3):99–104, 2019.

[15] Jeremy Daly. Event injection: Protecting your serverless applications. https://www.jeremydaly.com/event-injection-protecting-your-serverless-applications/, 2018.

[16] Michael Dalton, Christos Kozyrakis, and Nickolai Zeldovich. Nemesis: Preventing authentication and access control vulnerabilities in web applications. 2009.

[17] Thomas Pasquier, Jean Bacon, Jatinder Singh, and David Eyers. Data-centric access control for cloud computing. In *Proceedings of the 21st ACM on Symposium on Access Control Models and Technologies*, pages 81–88, 2016.

[18] Thomas FJ-M Pasquier, Jatinder Singh, Jean Bacon, and Olivier Hermant. An information flow control model for the cloud.

[19] Dorothy E Denning. A lattice model of secure information flow. *Communications of the ACM*, 19(5):236–243, 1976.

[20] Joseph A Goguen and José Meseguer. Security policies and security models. In *1982 IEEE Symposium on Security and Privacy*, pages 11–11. IEEE, 1982.

[21] Kalev Alpernas, Cormac Flanagan, Sadjad Fouladi, Leonid Ryzhyk, Mooly Sagiv, Thomas Schmitz, and Keith Winstein. Secure serverless computing using dynamic information flow control. *arXiv preprint arXiv:1802.08984*, 2018.

[22] Hayawardh Vijayakumar, Guruprasad Jakka, Sandra Rueda, Joshua Schiffman, and Trent Jaeger. Integrity walls: Finding attack surfaces from mandatory access control policies. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, pages 75–76, 2012.

[23] Andrei Sabelfeld and David Sands. A per model of secure information flow in sequential programs. *Higher-order and symbolic computation*, 14(1):59–91, 2001.

[24] Nevin Heintze and Jon G Riecke. The slam calculus: programming with secrecy and integrity. In *Proceedings of the 25th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 365–377, 1998.

[25] Siani Pearson and Marco Casassa-Mont. Sticky policies: An approach for managing privacy across multiple parties. *Computer*, 44(9):60–68, 2011.

[26] Andrei Sabelfeld and David Sands. Dimensions and principles of declassification. In *18th IEEE Computer Security Foundations Workshop (CSFW'05)*, pages 255–269. IEEE, 2005.

[27] Thomas H Austin, Tommy Schmitz, and Cormac Flanagan. Multiple facets for dynamic information flow with exceptions. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 39(3):1–56, 2017.

[28] Thomas H Austin and Cormac Flanagan. Permissive dynamic information flow analysis. In *Proceedings of the 5th ACM SIGPLAN Workshop on Programming Languages and Analysis for Security*, pages 1–12, 2010.

[29] Thomas H Austin and Cormac Flanagan. Multiple facets for dynamic information flow. In *Proceedings of the 39th annual ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 165–178, 2012.

[30] Thomas FJ-M Pasquier and David Eyers. Information flow audit for transparency and compliance in the handling of personal data. In *2016 IEEE International Conference on Cloud Engineering Workshop (IC2EW)*, pages 112–117. IEEE, 2016.

[31] Afshar Ganjali and David Lie. Auditing cloud management using information flow tracking. In *Proceedings of the seventh ACM workshop on Scalable trusted computing*, pages 79–84, 2012.

[32] Neil Vachharajani, Matthew J Bridges, Jonathan Chang, Ram Rangan, Guilherme Ottoni, Jason A Blome, George A Reis, Manish Vachharajani, and David I August. Rifle: An architectural framework for user-centric information-flow security. In *37th International Symposium on Microarchitecture (MICRO-37'04)*, pages 243–254. IEEE, 2004.

[33] G Edward Suh, Jae W Lee, David Zhang, and Srinivas Devadas. Secure program execution via dynamic information flow tracking. *ACM Sigplan Notices*, 39(11):85–96, 2004.

[34] Thomas FJM Pasquier, Jean Bacon, and David Eyers. Flowk: Information flow control for the cloud. In *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, pages 70–77. IEEE, 2014.

[35] Georgios Portokalidis, Asia Slowinska, and Herbert Bos. Argos: an emulator for fingerprinting zero-day attacks for advertised honeypots with automatic signature generation. *ACM SIGOPS Operating Systems Review*, 40(4):15–27, 2006.

[36] Maxwell Krohn, Alexander Yip, Micah Brodsky, Natan Cliffer, M Frans Kaashoek, Eddie Kohler, and Robert Morris. Information flow control for standard os abstractions. *ACM SIGOPS Operating Systems Review*, 41(6):321–334, 2007.

[37] Jatinder Singh, Thomas FJ-M Pasquier, Jean Bacon, and David Eyers. Integrating messaging middleware and information flow control. In *2015 IEEE International Conference on Cloud Engineering*, pages 54–59. IEEE, 2015.

[38] Marwa Elsayed and Mohammad Zulkernine. Ifcaas: information flow control as a service for cloud security. In *2016 11th International Conference on Availability, Reliability and Security (ARES)*, pages 211–216. IEEE, 2016.

[39] Christian Hammer and Gregor Snelting. Flow-sensitive, context-sensitive, and object-sensitive information flow control based on program dependence graphs. *International Journal of Information Security*, 8(6):399–422, 2009.

[40] Fred B Schneider. Enforceable security policies. *ACM Transactions on Information and System Security (TISSEC)*, 3(1):30–50, 2000.

[41] Safwan Mahmud Khan, Kevin W Hamlen, and Murat Kantarcioglu. Silver lining: Enforcing secure information flow at the cloud edge. In *2014 IEEE International Conference on Cloud Engineering*, pages 37–46. IEEE, 2014.

[42] David FC Brewer and Micheal J Nash. The chinese wall security policy. In *null*, page 206. IEEE, 1989.

[43] Ruoyu Wu, Gail-Joon Ahn, Hongxin Hu, and Mukesh Singhal. Information flow control in cloud computing. In *6th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2010)*, pages 1–7. IEEE, 2010.

[44] Naftaly H Minsky. A decentralized treatment of a highly distributed chinese-wall policy. In *Proceedings. Fifth IEEE International Workshop on Policies for Distributed Systems and Networks, 2004. POLICY 2004.*, pages 181–184. IEEE, 2004.

[45] Yasuharu Katsuno, Yuji Watanabe, Sanehiro Furuichi, and Michiharu Kudo. Chinese-wall process confinement for practical distributed coalitions. In *Proceedings of the 12th ACM symposium on Access control models and technologies*, pages 225–234, 2007.

[46] Andrew C Myers and Barbara Liskov. Protecting privacy using the decentralized label model. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 9(4):410–442, 2000.

[47] Vineeth Kashyap, Ben Wiedermann, and Ben Hardekopf. Timing- and termination-sensitive secure information flow: Exploring a new approach. In *2011 IEEE Symposium on Security and Privacy*, pages 413–428. IEEE, 2011.

Monitoring Indoor Activity of Daily Living using Thermal Imaging: A Case Study

Hassan M. Ahmed, Bessam Abdulrazak
AMbient Intelligence Lab (AMI-Lab)
Faculté des sciences, Université de Sherbrooke

***Abstract*—Monitoring indoor activities of daily living (ADLs) of a person is subjected to dependency on sensor type, power supply stability, and connectivity stability without mentioning artifacts introduced by the person himself. Multiple challenges have to be overcome in this field, such as; detecting the precise spatial location of the person, and estimating vital signs like an individual's average temperature. Privacy is another domain of the problem to be thought of with care. Identifying the person's posture without a camera is another challenge. Posture identification is a key in assisting detection of a person's fall. Thermal imaging could be a proper solution for most of the mentioned challenges. It provides monitoring both the person's average temperature and spatial location while maintaining privacy. In this research, an IoT system for monitoring an indoor ADL using thermal sensor array (TSA) is proposed. Three classes of ADLs are introduced, which are daily activity, sleeping activity and no-activity respectively. Estimating person average temperature using TSAs is introduced as well in this paper. Results have shown that the three activity classes can be identified as well as the person's average temperature during day and night. The person's spatial location can be determined while his/her privacy is maintained as well.**

***Keywords*—Activity monitoring; activities of daily living (ADLs); thermal imaging; indoor monitoring; thermal sensor array (TSA)**

I. Introduction

Monitoring indoor activities of daily living (ADLs) can be achieved with different methods [1]–[5]. As a first step, building occupancy and estimating the number of occupying individuals should be performed. A full review for detection of building occupancy and estimating number of individuals inside a building is presented by Chen et al. [1]. In this research, a review for different systems utilizing sensor fusion for building occupancy detection is presented. A comparison between the sensors utilized is conducted as well. Another complete review for different approaches used in detecting occupancy of the buildings, counting number of individuals and tracking those individuals is introduced by Saha et al. [2]. The authors in that review, focused on two aspects, the mathematical point of view and the corresponding metrics for evaluating these approaches. They focused on the prediction and performance metrics in order to establish a benchmark system for indoor occupancy estimation. A third systematic review concerned about the ambient assisted living technologies and focusing on their impact on individuals' health is presented by Choukou et al. [6].

Inertial sensors are utilized in Voung et al. [3] to monitor the indoor wandering patterns to detect behavior disorder for people with dementia. Their preliminary experimental results have shown that their proposed solution has outperformed the existing classification algorithms based on time-series analysis with respect to accuracy and time performance. Another research in the scope of dementia is conducted by Das et al. [4], where a one-class machine learning classification approach for detecting real-time indoor elderly individual daily activity errors is proposed. Machine learning techniques with the aid of computer vision approaches are used to monitor the daily activities of elderly people in another study Chen et al. [5]. Deep convolution neural network approach is used for recognition of daily activities such as; eating, bathroom entries, sleeping and housekeeping by the work presented by Gochoo et al. [7]. The proposed deep convolution neural network is found to outperform the existing models by F1 score of 0.951. Utilizing motion sensors is one of them [8]–[10]. As in Aloulou et al. [8], an adaptive approach for plug-and-play mechanisms of motion sensors used for ambient assistive living of elderly people is proposed. The real-life deployment for the system proposed in the previous study using motion sensors is presented in Aloulou et al. [9], where the authors deployed the system in three nursing room and monitored the daily activity for elderly people for a 14 months period. The study included eight dementia patients. A pilot study for IoT deployment of a continuous real-time monitoring of elderly people using unobtrusive technology and utilizing door sensors and motion sensors as a core sensor is conducted by the same team Aloulou et al. [10]. The aim of that pilot study is to identify the health-related problems for elderly people. A complete unsupervised approach is used to detect and model the behavioral change for elderly people by the use of passive-sensing technology; such as PIR and motion sensors is presented and deployed by Hu et al. [11]. Their experimental results have shown the ability of the system to detect the following individuals' activities; namely, sleeping, outing and visiting activities in addition to individuals' health status. A recent project focusing towards assisted living for elderly people to support aging in place is presented by Choukou et al. [12]. A case study for investigating the claim of smart flooring system to detect elderly people falling is presented by Chintanu et al. [13].

Motion sensors can be used to detect if there is any motion taking place in front of the sensor. On the one hand, these sensors are effective for detecting the movement of one person, and their effectiveness increases when sensor fusion is deployed. However, they impose several limitations that can be

summarized as follows: a) they cannot detect the exact spatial location of the monitored person or even an estimate for its spatial location, b) they cannot differentiate between the steady sitting state and the no-motion state, c) they cannot be used for fall detection as they cannot obtain depth and 2D data about the person's location, and d) they do not inform about the number of individuals inside the room. Moreover, this type of sensor has another limitation is that it does not give any vital data about the person during the monitoring period.

Another option of detecting ADLs is using another physical quantity i.e., the temperature of the person [14]–[17]. Every human being can be considered as a heat source that can be detected using Thermal Sensor Array (TSA). Thus, the problem can be reduced to being only the detection of the heat distribution of the person inside the room. In other words, it is possible to track the estimated spatial location of the person inside the room by tracking his/her thermal distribution. (Fig. 1 shows the thermal 2D distribution of two heat sources as measured by a thermal sensor array).

Given the various advantages of thermal imaging, the contribution of this paper is to assess the effectiveness of a thermal sensor array for tracking a subject's indoor activities.

The paper is organized as follows: section II discusses the advantages of thermal imaging over the mentioned limitations. Section III methodology used in conducting the experiment. Results of the experiment and related discussion are introduced in Section IV. Finally, the conclusion is presented in Section V.

II. Thermal Imaging Technology

A thermal sensor array can be used to pick up a complete capture for the thermal distribution of the room every minute and store its readings for subsequent processing [18], [19] [20]. This method has many advantages.

- First, it maintains the privacy of participating persons.
- Second, it tracks the precise spatial location of the person, giving rise to a better understanding of behavior patterns extracted from activity per spatial location.
- Third, it monitors the average temperature of the person per unit timestamp allowing for better assessment of the person's health status (e.g., if he/she is suffering from fever or relevant diseases).
- Fourth, it can track whether the person is in an upstanding posture or has lain down, using advanced algorithms that can differentiate between the thermal distribution in both cases. Upstanding posture gives more concise (confined) thermal distribution for the person, alike the horizontal or lying down position where relevant thermal distribution is much wider and scattered. This feature can also be helpful in fall detection [21].

Several researches have concluded the possibility of indoor occupancy estimation using TSA [18], [19]. They all placed TSA in a specific spatial location in the room. The person IR emission is triangulated to estimate his location as in [20]. An experimental evaluation for two low resolution thermal sensor arrays for occupancy detection is conducted by Rinta-Homi et al. [22]. Another systematic study investigated the performance of three low resolution thermal sensor arrays in detecting the indoor occupancy with the aid of machine learning algorithms [23]. In this study, the authors conquered the challenge of detecting two individuals with low proximity to each other by the use of iterative blob filtering technique to split the blob which is larger than that of a single human being. The privacy issue is respected in [24] by using a low resolution thermal sensors to detect the occupancy and to track the individuals indoors. A study is performed to detect the presence of people indoors by means of identifying their directions with respect to a room doorway using a low resolution thermal sensor array is conducted by Perra et al. [25]. The use of thermal array sensors for detecting individuals' fall is investigated practically by the study conducted in [21].

All these presented work focused on occupancy detection and estimation only without giving attention to the person's ADL indoors. In our work the focus is on monitoring a person's ADL to deduce his behavior. Following, the methodology used in our proof-of-concept IoT experiment is discussed.

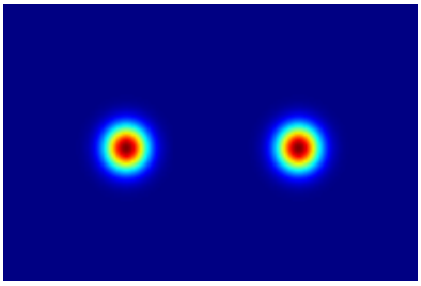


Fig. 1. Thermal Distribution of Two Heat Sources as Picked up by an Array Sensor after Interpolation.

III. Methodology

A male person is monitored over time using a thermal sensor array to track his activity. A single person as a proof-of-concept and for feasibility is monitored. An IoT system consisting of a thermal sensor array and a processing unit to analyze the acquired data is implemented. The person's activity is monitored by tracking specific thermal image pixels related to the activity spatial locations inside the room. The temperature of the corresponding pixels in the temporal domain to construct different activity vectors/arrays for the person is tracked. Then these vectors are sent to a cloud server annotated with their relevant timestamp, where they can be stored and analyzed.

The system is deployed to monitor a single person living inside a room. The person's activity is monitored on a 24-hour basis and is classified (for this work) as three classes: the sleeping activity, the daily activity and the no-activity classes. The corresponding spatial locations at which these three activities are most likely to happen are marked on the room schematic. The bed represents the sleeping activity of the person. The working table and dining table represents the daily activity of the subject. The room schematic is presented in Fig. 2 along with the sensor located inside the room and its location with respect to the person. The activities' spatial locations are presented in Fig. 3.

The IoT system consists of MLX90641 (16x12/110°x 75°/~7m axial range) [26] which is a thermal sensor array for acquiring thermal activity and a Raspberry Pi 3B+ is used as a node. The acquired thermal frame is transmitted as a 1D-array (element-by-element) along with a corresponding timestamp via CoAP transmission protocol from the Pi to a server for subsequent data processing and analysis. Another copy from the transmitted data is stored locally on the Pi for retrieval and substitution in case of transmission failure situations. The IoT system block diagram is presented in Fig. 4. The complete process of the system is presented in Fig. 5. The experimental hardware setup is shown in Fig. 6. The next section explains the temporal thermal activity tracking algorithm.

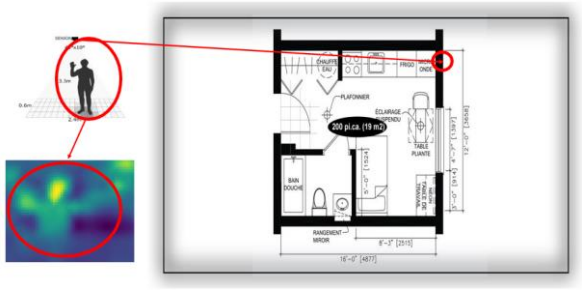


Fig. 2. Room Schematic along with its Real Photo, describing the Spatial Location of the Sensor W.R.T the Person Body.

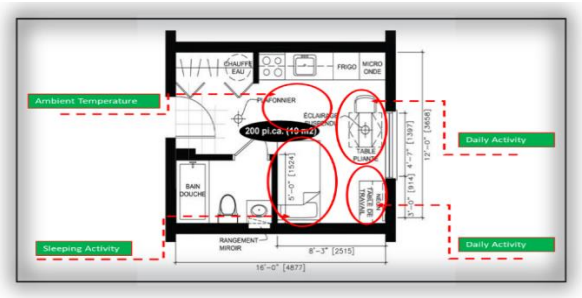


Fig. 3. Regions of Interest along with its Corresponding Activity.

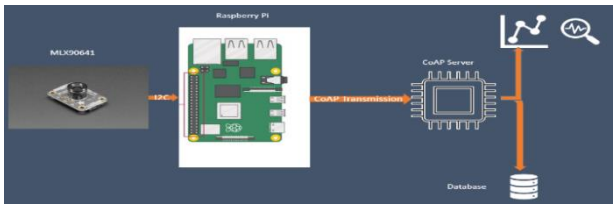


Fig. 4. Block Diagram of the Complete System.

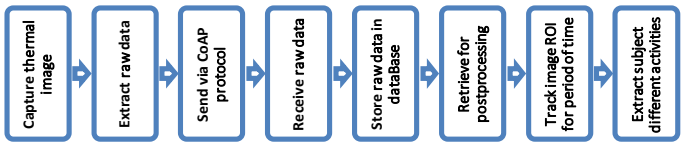


Fig. 5. Complete System Process.

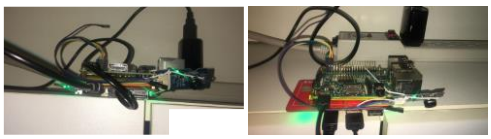


Fig. 6. Complete Hardware Setup during Runtime.

A. The Temporal Thermal Activity Tracking Algorithm

This algorithm is used to track the spatial thermal activity in the acquired image from the sensor array. It starts with tagging the corresponding pixels that are relevant to the spatial locations where the three activity classes are most likely to happen. The tagging is performed by relating specific image pixels to a location where each class of the three classes happen, i.e. for the sleeping activity class, the person is asked to lie down on the bed and a reference image is captured to identify the location of the bed and so on for the other two classes. Then the related pixels to these locations are tracked over time. This process is shown in Fig. 7.

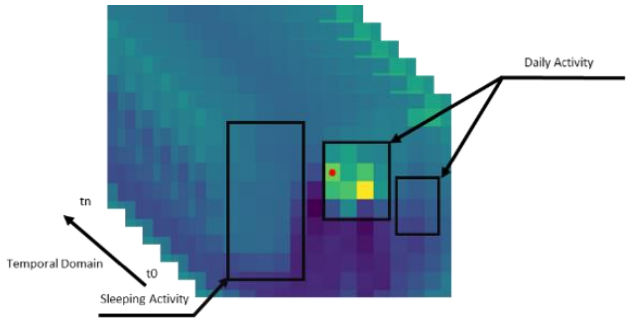


Fig. 7. Visualization of Temporal Thermal Activity Tracking Algorithm.

Each thermal activity corresponding to one of the three classes can be formulated mathematically as a 1D-array by the following formula (1):

$$A^K = [a_t^K, a_t^K, a_t^K, \dots\dots, a_T^K], \forall\, t = \{0, 1, 2, \dots, T\}\ and\ K \in \{1, 2, 3\} \quad (1)$$

Where A is the thermal activity array, t is the temporal resolution, T is the final timestamp of the monitoring period, and K is the activity class (1: Sleeping, 2: Daily and 3: No-activity).

The computational algorithm is implemented on the server side due to the limited computational resources used on the monitoring site side. The aim of this algorithm is to identify the timestamps at which each activity has started and ended as accurate as possible.

The used algorithm starts firstly by collecting the relative pixel values into 1D arrays, where each value in the activity 1D array represents the temperature value at specific timestamp in the real spatial location in the room. Each constructed 1D array (A^K) is then differentiated to its first derivative array, this will help identify the timestamps at which there were changes along the array. If the value of the array is falling from a high temperature value to a lower temperature value, then there has been an activity at this spatial place and is now ended. This is reflected in the first derivative array as a sharp impulse in the negative direction, i.e., having a negative value. It is at this value where we identify the ending timestamp of one of the three activities done by the individual. On the contrary, if the 1D array is experiencing a sudden increase in its values from a relatively low value to a higher value, then its first derivative output should have positive impulse values indicating the start of the activity at this spatial location.

This process is ruled by a try and error threshold of >3C degrees between the no activity state and the activity state. The complete No-activity inside the room is identified when there is no activity status detected on all of the three arrays representing the thermal activity of the person on the three spatial locations tagged previously.

The computational algorithm is illustrated in the following Algorithm1.

ALGORITHM1: IDENTIFY ACTIVITY START AND END TIME-STAMPS

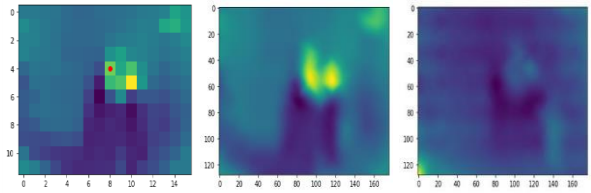
```
   Input: activity1_arr[time_stamp][Temperature],
    activity2_arr[time_stamp][Temperature]
   Output: activity1_duration_arr[time_stamp][status], status ∈{Start, End}
    activity2_duration_arr[time_stamp][status], status ∈{Start, End}
1  Initialization of variables:
   Set act1[][] ← activity1_arr[time_stamp][Temperature],
   Set act2[][] ← activity2_arr[time_stamp][Temperature]
   Set derivative1_arr [][] ← empty
   Set derivative2_arr [][] ← empty
   Set threshold ← 3C
2  if (length(act1) == length(act2))
3      for each entry in length(activity1_arr):
4          if index(entry) <= (length(activity1_arr) – 1)
5              derivative1_arr[time_stamp1][]←
               act1[time_stamp1][entry+1] – act1[time_stamp][entry]
               derivative2_arr[time_stamp2][]←
               act2[time_stamp2][entry+1] – act2[time_stamp][entry]
6          End
7      End
8      for each entry in length(derivative1_arr):
9          if (derivative1_arr[entry] >= threshold )
10             activity1_duration_arr[time_stamp][Start]←derivative1_arr[time_stamp][-]
11         end
12
13         if (derivative1_arr1[entry] <= -( threshold) )
14
15             activity1_duration_arr[time_stamp][End]←derivative1_arr[time_stamp][-]
16         end
17
18     end
19
20     for each entry in length(derivative2_arr):
21
22         if (derivative2_arr1[entry] >= threshold )
23
24             activity2_duration_arr[time_stamp][Start]←derivative2_arr[time_stamp][-]
25         end
26
27         if (derivative2_arr1[entry] <=-( threshold ))
28
29             Activity2_duration_arr[time_stamp][End]←derivative2_arr[time_stamp][-]
30         end
31
32     end
33
34 else:
35
36     Terminate with error message (“The duration of time-stamps and
       activity does not match”)
37 end
38
```

IV. Results and Discussion

The non-interpolated image (16x12) as captured from the sensor, its corresponding interpolated image (128x176) and no-activity image are presented in Fig. 8(a), (b) and (c) respectively. The person’s location inside the image is labeled by the red dot. Two different ways of sleeping activities are presented in Fig. 9.

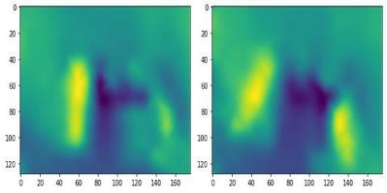
The person’s behavior for 10 consecutive days in April 2021 is monitored. Every monitored day is presented in a separate curve starting from 00:00 o’clock till 11:59 of the same day on the X-axis and the temperature on the Y-axis, except for the last day where the monitoring ended at 16:30. The orange curve represents sleeping activity and the blue curve represents daily activity for the monitored person. The first three days of monitoring are shown in Fig. 10, namely 7th, 8th and 9th April. The three consecutive days which are 10th, 11th and 12th April are shown in Fig. 11. Fig. 12 shows the person’s behavior during 13th, 14th and 15th April. Finally, the last two days are shown in Fig. 13.

Sleeping activity periods, daily activity periods, no-activity periods, and missing data periods for each day are listed in Table I and are shown as a bar chart in Fig. 14. Behavior statistics such as mean value and percentage value for each activity type along the whole monitoring period are listed in Table II.



(a) (b) (c)

Fig. 8. (a) Non-Interpolated Image showing the Person doing his Daily Activity, (b) The Corresponding Interpolated Image, and (c) Interpolated Image showing no Activity Inside the Room.



(a) (b)

Fig. 9. Two different Sleeping Activities of the Person.

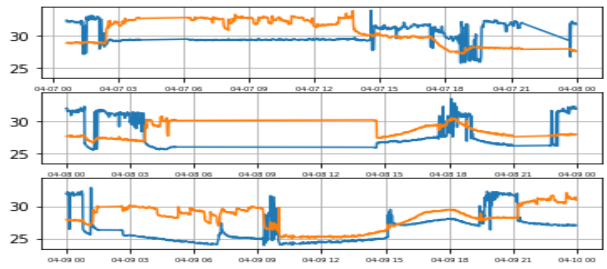


Fig. 10. Monitoring Person’s behavior during 7th, 8th and 9th April.

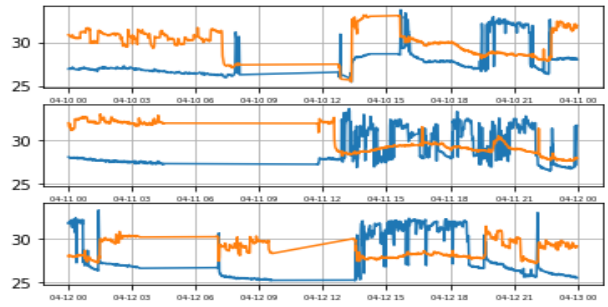


Fig. 11. Monitoring Person's behavior during 10th, 11th and 12th April.

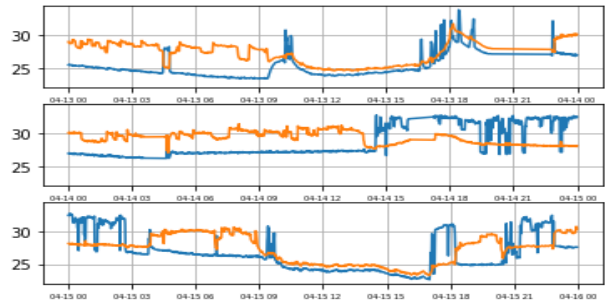


Fig. 12. Monitoring Person's behavior during 13th, 14th and 15th April.

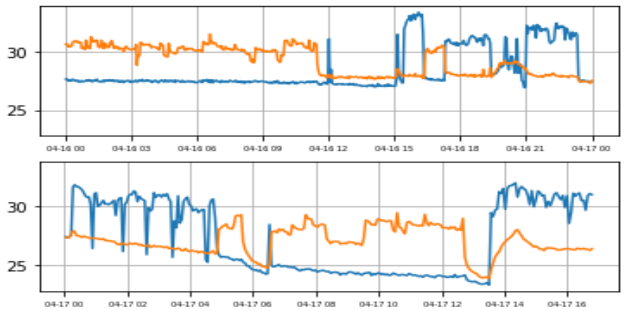


Fig. 13. Monitoring Person's behavior during 16th and 17th April.

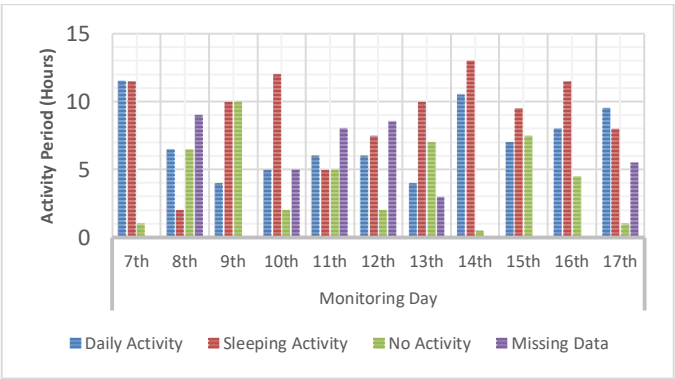


Fig. 14. Different Activity Periods for the Person's behavior during the Whole Monitoring Period.

On each day of the monitoring period, the sleeping activity and the daily activity are observed to be contrary to each other. Both activities have a temperature difference of about 3°C degrees each day. The person's average temperature is observed to be 32°C degrees on average during the whole monitoring period.

TABLE I. BEHAVIOR MONITORING PERIODS

		Activity Period (Hours)			
		Daily	**Sleeping**	**No-Activity**	**Missing Data**
Monitoring Day	7th	11.5	11.5	1	0
	8th	6.5	2	6.5	9
	9th	4	10	10	0
	10th	5	12	2	5
	11th	6	5	5	8
	12th	6	7.5	2	8.5
	13th	4	10	7	3
	14th	10.5	13	0.5	0
	15th	7	9.5	7.5	0
	16th	8	11.5	4.5	0
	17th	9.5	8	1	5.5

TABLE II. BEHAVIOR STATISTICS FOR THE WHOLE MONITORING PERIOD

Activity Type	**Mean (Hours)**	**Percentage**
Daily Activity	7.090909	0.295454545
Sleeping Activity	9.090909	0.378787879
No Activity	4.272727	0.178030303
Missing Data	3.545455	0.147727273

The most dominant activity of the person during the monitoring period is the sleeping activity as concluded from the behavior statistics in Table II with an average of 9 hours per day. Daily activity happens to be the second place with 7 hours per day, and the no activity comes in third place with about 4.2 hours per day.

Small periods of no-activity happening inside the room are inferred as a bathroom entry due to its small duration and its location between two long durations of either sleeping activity or daily activity or both of them. The person is most likely considered visiting the bathroom just after wake up. The person's concluded bathroom visits are listed in Table III.

It is concluded from Table III that the person's bathroom visit takes on average between 30 minutes to 60 minutes.

During the long no-activity durations exceeding one hour the person is considered out of the room, i.e., these durations are considered outing periods. On 9th, 13th, 15th and 16th long durations of no-activity at the same normal working hours are observed, hence it is concluded that the person was out for work at these periods.

TABLE III. THE PERSON'S BATHROOM VISITS

	Bathroom Visit (Start / Duration [minutes])			
Monitoring Day	7th	14:00 / 60		
	8th	01:00 / 30		
	9th	01:30 / 30		
	10th	07:00 / 30	13:00 / 30	21:30 / 60
	11th	17:30 / 30	22:00 / 30	22:30 / 90
	12th	01:30 / 30	18:30 / 30	21:00 / 60
	13th	09:30 / 30		
	14th	14:15 / 30		
	15th	03:00 / 30		
	17th	12:30 / 30		

It is also noticed that the person often goes to sleep after midnight and before 03:00 except for 8th, 15th and 17th April, where the person went to bed after 03:00. It is concluded here that the person is a night owl.

It is noticed that the person spends about 29% of its day as daily activity, 37% as sleeping activity, and 17% as no-activity or bathroom visits and outings. Missing data is about 14%.

In terms of monitoring the person's temperature, a significant decrease in the person's body temperature is observed during 12th April around 20:00 and during 17th April from 00:00 till around 05:00.

V. Conclusion

In this paper a monitoring system based on thermal sensor array that can capture a person's activities of daily living (ADLs) is proposed and implemented. The monitored ADLs are classified as sleeping, daily, and no-activity at all. The experiment proves that the system enables detection of a person's spatial location indoor precisely. In addition, the experiment enables prediction for the bathroom visits and the outing and estimates the person's temperature during the whole monitoring period along with maintaining the person's privacy as well. As a future development of this experiment, there are different directions to be investigated. A first direction is the hardware experimental setup; where for covering wider field of view (FOV) an approach of TSA gird should be utilized. Another direction is, in the direction of automatic recognition of the individuals' presence inside the room, where machine learning approaches should be utilized. This is considered to automate the presence recognition and to reveal more daily activities' types. Also, machine learning techniques can help differentiate between multiple heat sources not only other than multiple human beings' presence in the room but also differentiating between different heat sources such as; the heater-on and the heater-off states in the room.

References

[1] Z. Chen, C. Jiang, and L. Xie, "Building occupancy estimation and detection: A review," Energy Build., vol. 169, pp. 260–270, 2018.

[2] H. Saha, A. R. Florita, G. P. Henze, and S. Sarkar, "Occupancy sensing in buildings: A review of data analytics approaches," Energy Build., vol. 188, pp. 278–285, 2019.

[3] N. K. Vuong, S. Chan, C. T. Lau, S. Y. W. Chan, P. L. K. Yap, and A. S. H. Chen, "Preliminary results of using inertial sensors to detect dementia- related wandering patterns," in 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 3703–3706, doi: 10.1109/EMBC.2015.7319197.

[4] B. Das, D. J. Cook, N. C. Krishnan, and M. Schmitter-Edgecombe, "One-Class Classification-Based Real-Time Activity Error Detection in Smart Homes," IEEE J. Sel. Top. Signal Process., vol. 10, no. 5, pp. 914–923, 2016, doi: 10.1109/JSTSP.2016.2535972.

[5] D. Chen, A. J. Bharucha, and H. D. Wactlar, "Intelligent Video Monitoring to Improve Safety of Older Persons," in 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007, pp. 3814–3817, doi: 10.1109/IEMBS.2007.4353163.

[6] M. A. Choukou, A. Polyvyana, Y. Sakamoto, and A. Osterreicher, "Ambient assisted living technologies to support older adults' health and wellness: a systematic mapping review," Eur. Rev. Med. Pharmacol. Sci., vol. 25, no. 12, pp. 4289–4307, 2021.

[7] M. Gochoo, T.-H. Tan, S.-H. Liu, F.-R. Jean, F. S. Alnajjar, and S.-C. Huang, "Unobtrusive activity recognition of elderly people living alone using anonymous binary sensors and DCNN," IEEE J. Biomed. Heal. informatics, vol. 23, no. 2, pp. 693–702, 2018.

[8] H. Aloulou, M. Mokhtari, T. Tiberghien, J. Biswas, and P. Yap, "An adaptable and flexible framework for assistive living of cognitively impaired people," IEEE J. Biomed. Heal. informatics, vol. 18, no. 1, pp. 353–360, 2013.

[9] H. Aloulou et al., "Deployment of assistive living technology in a nursing home environment: methods and lessons learned," BMC Med. Inform. Decis. Mak., vol. 13, no. 1, pp. 1–17, 2013.

[10] H. Aloulou, M. Mokhtari, and B. Abdulrazak, "Pilot Site Deployment of an IoT Solution for Older Adults' Early Behavior Change Detection," Sensors, vol. 20, no. 7, 2020, doi: 10.3390/s20071888.

[11] R. Hu et al., "An Unsupervised Behavioral Modeling and Alerting System Based on Passive Sensing for Elderly Care," Futur. Internet, vol. 13, no. 1, p. 6, 2021.

[12] M.-A. Choukou, J. Ripat, S. Mallory-Hill, and R. Urbanowski, "Designing the University of Manitoba Technology for Assisted Living Project (TALP): A Collaborative Approach to Supporting Aging in Place," in Congress of the International Ergonomics Association, 2021, pp. 223–228.

[13] Y. Chintanu, J. Waycott, and C. Newton, "With increasing numbers of frail elderly, is smart flooring a useful strategy for falls detection and reduction?," in 53rd International Conference of the Architectural Science Association, 2019, pp. 557–566, [Online].

[14] A. Naser, A. Lotfi, J. Zhong, and J. He, "Human activity of daily living recognition in presence of an animal pet using thermal sensor array," in Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments, 2020, pp. 1–6.

[15] M. Cokbas, P. Ishwar, and J. Konrad, "Low-resolution overhead thermal tripwire for occupancy estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 88–89.

[16] S. Munir, S. Mohammadmoradi, O. Gnawali, and C. P. Shelton, "Measuring people-flow through doorways using easy-to-install IR array sensors." Google Patents, 2021.

[17] A. Naser, A. Lotfi, J. Zhong, and J. He, "Heat-map based occupancy estimation using adaptive boosting," in 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2020, pp. 1–7.

[18] Y. Yuan, X. Li, Z. Liu, and X. Guan, "Occupancy estimation in buildings based on infrared array sensors detection," IEEE Sens. J., vol. 20, no. 2, pp. 1043–1053, 2019.

[19] A. Gomez, F. Conti, and L. Benini, "Thermal image-based CNN's for ultra-low power people recognition," in Proceedings of the 15th ACM International Conference on Computing Frontiers, 2018, pp. 326–331.

[20] J. Kemper and D. Hauschildt, "Passive infrared localization with a probability hypothesis density filter," in 2010 7th Workshop on Positioning, Navigation and Communication, 2010, pp. 68–76.

[21] Z. Liu, M. Yang, Y. Yuan, and K. Y. Chan, "Fall detection and personnel tracking system using infrared array sensors," IEEE Sens. J., vol. 20, no. 16, pp. 9558–9566, 2020.

[22] M. Rinta-Homi, N. H. Motlagh, A. Zuniga, H. Flores, and P. Nurmi, "How Low Can You Go? Performance Trade-offs in Low-Resolution Thermal Sensors for Occupancy Detection: A Systematic Evaluation," Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol., vol. 5, no. 3, pp. 1–22, 2021.

[23] V. Chidurala and X. Li, "Occupancy Estimation Using Thermal Imaging Sensors and Machine Learning Algorithms," IEEE Sens. J., vol. 21, no. 6, pp. 8627–8638, 2021.

[24] R. Rabiee and J. Karlsson, "Multi-Bernoulli Tracking Approach for Occupancy Monitoring of Smart Buildings Using Low-Resolution Infrared Sensor Array," Remote Sens., vol. 13, no. 16, p. 3127, 2021.

[25] C. Perra, A. Kumar, M. Losito, P. Pirino, M. Moradpour, and G. Gatto, "Monitoring Indoor People Presence in Buildings Using Low-Cost Infrared Sensor Array in Doorways," Sensors, vol. 21, no. 12, p. 4062, 2021.

[26] N. Equivalent and T. Difference, "MLX90641 16x12 IR array," pp. 1–53, 2020.

Improving the Quality of e-Commerce Service by Implementing Combination Models with Step-by-Step, Bottom-Up Approach

Hemn Barzan Abdalla, Ge Chengwei, Baha Ihnaini
Department of Computer Science
College of Science and Technology
Wenzhou-Kean University
China

***Abstract*—e-Commerce, as a hot industry, plays an important role in people's lives. People visit e-commerce websites, check what they want, then click buy, and finally complete the transaction. The developments taking place at the level of electronic services at the global level and the intensification of competition and the increase in the experiences of electronic shoppers, the awareness and understanding of companies of the distinctive characteristics of the population in the region and their purchasing habits has become the most important for companies of e-commerce and services, where it is imperative Companies should keep pace with these developments and provide electronic services via the internet of high quality and efficiency, by focusing on the most important requirements for customer satisfaction, especially in light of the information and technological revolution. However, customers will have an awful experience if they visit crudely made e-commerce websites. Kunst A. (2019, Dec 20) claimed that around a total of 37.4% of customers complained that they had an awful shopping experience. The reason is that the service quality of e-commerce websites is not up to standard. This research aims to improve the quality of e-commerce service by using the Comprehensive and Referential Combination Model by implementing a Step-by-Step, Bottom-Up approach. Finally, we will recommend improving the quality of e-commerce service in construct and revision ways within parts of this model.**

***Keywords*—E-commerce; website; framework; criteria; model; approach; service; quality**

I. Introduction

E-commerce is also called electronic commerce or internet commerce. In general, it represents people using the internet to buy and sell commodities or provide clearly priced services in funds and data flows. [6, 23] A computer network is a fundamental tool to make e-commerce achieve its value. More and more people can use the internet to browse the website, which provides the possibility for the promotion of e-commerce websites. From this point of view, e-commerce has great development potential and broad prospects.[15, 16] An E-commerce website enables people to order the goods they want without leaving home, and they can receive the goods delivered by logistics at home. Thus, more people begin to choose e-commerce to choose and purchase their desired products. [17, 18] However, some informal e-commerce websites attract consumers to browse and view through false advertisements or other attractive ways (such as promotion).

As a result, consumers' shopping experience will be very poor, and their precious time will be wasted. Even some formal e-commerce websites will make consumers have a bad shopping experience. The former is due to fraud, extortion, and other bad means. At the same time, the latter is due to rough, lack of consideration of the user interface, inefficient navigation bar, lack of basic description of goods and payment methods without security guarantee, and other adverse factors. These will cause consumers to give up buying products, so there is a high probability that they will not come for a second time. It is very important to establish a basic framework and evaluation scheme for an e-commerce website. [19, 20, 21] The evaluation scheme consists of a series of criteria. Predecessors have not given up this kind of research and put forward their evaluation scheme.

The team of Van Der Merwe (2003) started to begin this type of research in 2003. They proposed to build a new framework and method to evaluate e-commerce websites. In other words, they want to re-establish a credible business principle. [13] They designed three phases to build the framework and method named Level 1, Level 2, and Level 3. They focus on customers, especially the customer buying cycle [13], consisting of recognition, information gathering, evaluation, and purchase completion. They put forward general direction criteria: user interface, navigation bar, content, and degree of website's credibility, technical development. [13] They set 4 criteria groups in each general criteria group. And each criteria group consists of 5 fundamental rules. They use these 100 rules to evaluate famous e-commerce websites (e.g., Amazon, Barnes & Noble, Exclusive Book, etc.) with the help of correspondence analysis developed by themselves. [13] This has greatly contributed to developing e-commerce websites for systematic evaluation. In addition, the evaluation methods provided by the later research also change due to the changes of the times and different concerns.

In 2016, the group of Watrobski developed a framework for evaluating e-commerce websites called PEQUAL [14]. The author improved the traditional equal method by combining the appropriate model and multi criteria decision analysis (MCDA) which consists of usability, site design, information quality,

trust, and empathy. [14] They established the framework of the PEQUAL [14] method. First and foremost, they choose the classical EQUAL method and choose one of the website evaluation methods other researchers develop. And then, they use relevant criteria evaluation and other parameters to grade e-commerce websites. It's based on a survey that users contribute. And then, they deal with collecting data by using PROMETHEE [14] method to conform group ranking and individual ranking of e-commerce websites in order to reduce the uncertainty of users filling in the questionnaire. Compared to the previous research, they use different criteria by using different analysis methods. Apart from academic evaluation of e-commerce. [22] Some credible sources also provided some benchmarks and some sets of criteria that are easy to understand to evaluate e-commerce websites for the public who don't know how to use professional analysis methods to evaluate e-commerce websites.

Kogan D. [7] suggested that we can pay attention to content, functionality, authority, and marketing advantages which are elements from e-commerce websites to grade individual e-commerce websites. Also, some websites also gave us several steps to evaluate e-commerce websites in a scientific way. In addition, Burke D.[1], as a famous user experience strategy master, provided 5 criteria for judge whether this e-commerce website is successful or not. Successful e-commerce websites will know how to avoid users register themselves in a complex way. They also consider demonstrating powerful ads and promotions, including a special offer. They also focus on building a strong search engine to ensure that customers will be beneficial to find desirable goods. In a nutshell, the e-commerce industry serves customers to make them have a wonderful shopping experience.

II. Problem Identification

How to build an effective model to improve the service quality of an e-commerce website? Can researchers get any inspiration from consumers visiting e-commerce websites and completing a series of transactions? Furthermore, how can researchers use an example to prove the necessity of building this model? How can researchers use the data to demonstrate the effectiveness of our model? Which specific e-commerce website do researchers need to study? What will researchers focus on a specific e-commerce website? What type of tools will researchers use to gather data? The following are some minor problems arising from the main research problems, but these problems basically need to be solved and clearly defined. What makes customers feel like an e-commerce website is good or bad? Is it based on the perceptual visual impact or content reality that researchers can be left behind to complete the process of browsing the website or completing the buying circle [13]? The entrepreneur needs to create an e-commerce website, but how can the entrepreneur create a practical website? What is the definition of utility? Or how can an entrepreneur scientifically create a qualified e-commerce website that can perform its own functions? What is the scope of an e-commerce website? Or the exact duty? Do researchers need to include technology in the evaluation of e-commerce websites? What about including language, culture, other intangible factors into evaluating the e-commerce websites? In addition, how does this research benefit human beings, specifically customers?

The above questions, the main questions need theoretical support, which refers to establish a framework and criteria, and then we compare this model with some former research. The following main questions, which are about the effectiveness of our models, can be answered in the part of research methods. And some of the questions can be answered in a short time. We decide to study the Taobao website, which is a famous e-commerce website in China. And we study how they can improve the quality of this service because we can probably find the link between their strategy and our model. Some of the secondary questions need to be judged by experience, and some can be done by consulting data. For example, we can consult some relevant information by identifying keywords to answer the first question because it’s a common background problem. Finally, we can get the answer by presenting some relevant data which is referred from searching websites.

These studies effectively studied how to establish e-commerce websites through a series of criteria, and then at the same time, through clear steps, detailed data of relevant variables to make their established criteria very powerful persuasive. There are other studies that cover not only the establishment of guidelines but also the primary demographic of business sites, namely customers. They take some of the important common behaviors and psychology of customers and make them into core principles, and then combine them with building websites to make them more approachable. Other researchers, however, have looked to the source of the problem and devised designs designed to help companies coordinate their organizational structures and optimize their workforce for market readiness. These results are convincing, but they fail to connect the two most important elements of e-commerce: the company and the customer. And customers can connect through e-commerce sites. Therefore, our research aims to achieve a joint model to solve problems from beginning to end through a series of scientific and organized methods so as to optimize the company's sales situation and customers' shopping environment.

III. Object Determination

Based on this research background, previous studies are to study how to establish a reasonable and complete evaluation mechanism and then use their own data processing methods to evaluate a selected group of e-commerce websites and then process the data. Then the relevant charts are established, and the evaluated websites are compared with each other to get the results. Finally, they give constructive suggestions. Although this will enable some people to obtain references through these studies, which is about getting the evaluation results of popular websites, We should focus on improving the service quality of e-commerce rather than just staying in the evaluation stage. However, an e-commerce website is the facade of the e-commerce industry, so we are committed to solving how to improve the service level of an e-commerce website.

IV. METHODOLOGY

A. Overview

First and foremost, our main goal is to improve the quality of the e-commerce website's service. According to the Oxford dictionary definition, service is a system supplying a public need such as transport, communications, or utilities such as electricity and water. The quality of e-commerce service mainly depends on whether consumers think the experience is good or bad after they visit the website, view the content of the website, select goods on the website, and complete the transaction. Therefore, consumers' evaluation after visiting a website is an important index to measure an e-commerce website. How can an e-commerce website's provider improve the quality of the service? We have come up with a solution for this problem. The research is based on that in order to make consumers have a better shopping experience; we need to think about how to improve the serviceability of e-commerce websites and solve the problem from the source. Evaluation can only detect the problem but cannot solve the problem. Therefore, we need to solve this problem with the model composed of framework and criteria for e-commerce websites. We need to modify or design a required e-commerce website based on our established model. This is due to the fact that we need to improve our e-commerce serviceability. So we can focus on how to help owners improve or design their e-commerce websites. An E-commerce website is an owner, a company's facade, and a business card. We have three main phases to construct this solution: building-up, validation, and application for the phase of building-up. We pay attention to and study the consumer's shopping behavior.

The consumer's shopping behavior has five main steps: visit the website, query the goods, view the content of the goods, buy the goods, and pay attention to the logistics trends. Our framework of this model is based on the behavior of customers and established criteria for each part. We can also help build a part of the model by combining an example of a successful e-commerce website. The framework of this model consists of the user interface, searchability, content readability, service safety, website accessibility, future value. We will subdivide each part, namely, criteria. Then we need to prove that our model is practical and reasonable. I will visit the databases of Google Scholar and Kean Library (e.g., ProQuest Central, ProQuest: business) to access some research papers about e-commerce by observing their frameworks and criteria and the data processing methods used.

We even assess the China database (e.g., CNICC) to recently analyze the condition of e-commerce in China. So we can be more informed about this field. We will compare with previous studies to show that evaluation alone is not enough. We need to put this model into action: improve the service level of an e-commerce website. Then, we may distribute the questionnaire and set up questions about consumers. For example, what they want most, what websites they tend to like, and so on. Then we will make suggestions to the owners on how to use our model to modify or design their e-commerce website to improve their service level.

B. Method of Constructing Combination ModelProposed Solution

1) General working plan: Fig. 1 consists of two main phases: Building-Up and Application. For Building-Up, it has problem identification that states in the introduction part, the construct combination model in Step-By-Step, Bottom-Up Approach with the help of validation, and refine the details about constructing the criteria in each part. For confirmation, it has to find support from customers by using a questionnaire, getting data support from previous scholarly journals, some relevant data to make statements persuasive. The application phase has two ways: redesign and revise by using the combination model we established. Finally, we will give recommendations for each condition.

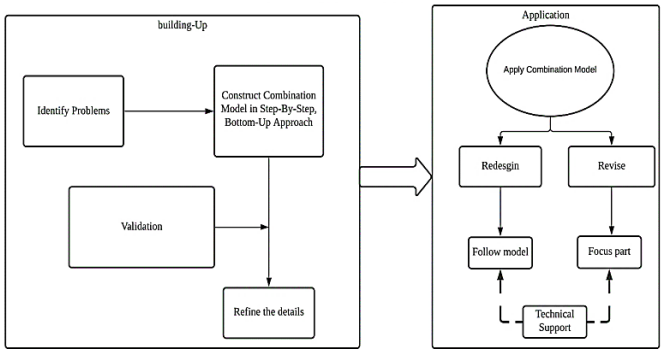


Fig. 1. General Working Plan.

2) Specify method in hierarchy approach: Now we want to construct an efficient method to construct method. The main method we use is called step-by-step, bottom-up, as shown in Fig. 2. We use this method to propose solutions to our research topic. This method is a hybrid method, which is composed of seven sub-methods. Each sub-method carries on the results generated by the previous method. Finally, a complete solution is developed.

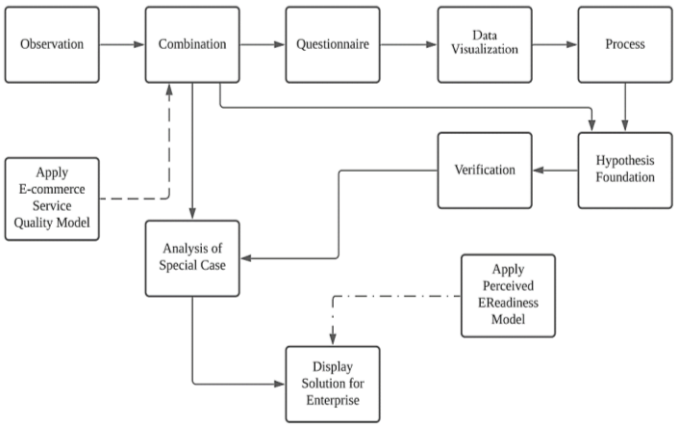


Fig. 2. The Structure of Step-By-Step, Bottom-Up Approach.

In the below explanation about the structure of the step-by-step, bottom-up approach:

a) Observation - The Behavior of Customers: The customer's shopping behavior has five main steps: visit the website, query the goods, view the content of the goods, buy the goods, and pay attention to the logistics trends. Let's model the behavior of customers who is about to visit a specific website. First and foremost, people initially enter keywords into Google search engine or Baidu search engine to find the website they want to visit. The domain name of some websites is easy to remember, and some people will directly enter the corresponding URL to find it. Usually, popular websites are located on the front page of search results, and they start to click on the website to visit it. Next, they will wait a moment to see the user interface of the e-commerce website. Customers usually visit e-commerce websites to buy what they want. So they will find what they want for the first time, they may seek to use classification screening or the search engine provided by the website to search. Then they will get the details of the product and click buy product to enter the purchase interface. Finally, through secure payment to complete the transaction and after-sales logistics services, finally get the desired goods. The Van's team in 2003 also provided the criteria model of constructing e-commerce websites based on customer behavior shown in Fig. 3 (they called it "buying circle"). This proves that our starting point is correct and consistent with other academic papers establishing e-commerce websites' criteria.

b) Combination - Combine two models with customers' feelings and structure of e-commerce website: The framework and criteria for establishing an e-commerce website are specific because an e-commerce website is a particular thing; customers can see and touch and have visual, intuitive feelings. However, the quality of service is abstract, invisible, and can only be perceived. It is not enough for us to build a good e-commerce website to improve the service quality of e-commerce. Therefore, we need to refer to other studies on the criteria of e-commerce service quality. We are going to combine our model with theirs to produce a new model.

We reference the framework of e-commerce service quality which was worked by Ishak et al. (2021). They used Delphi Method by using questionnaires and giving them to experts to hear their responses, which will be excellent support for constructing criteria in the e-commerce service. Here is the result from Ishak's team, as shown in Fig. 4. They provide what the e-commerce industry can do for customers and what kind of experience customers need: reliability, responsiveness, assurance, empathy, and tangible ways. However, they established a model based on the view of customers. So we want to refine this model and create another model based on the structure of e-commerce websites and combine them from different perspectives. We want to make more details in each part associated with customer behavior. We also refined some subcriteria and category them into more complex parts, which will list in the result section.

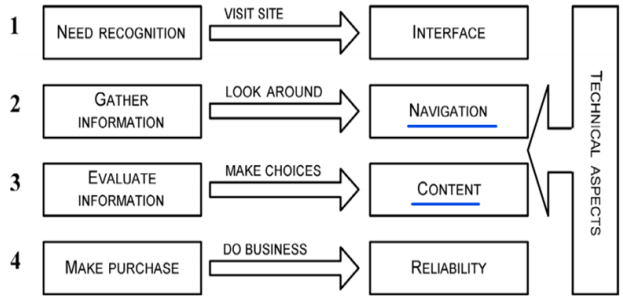


Fig. 3. Customer behavior [13].

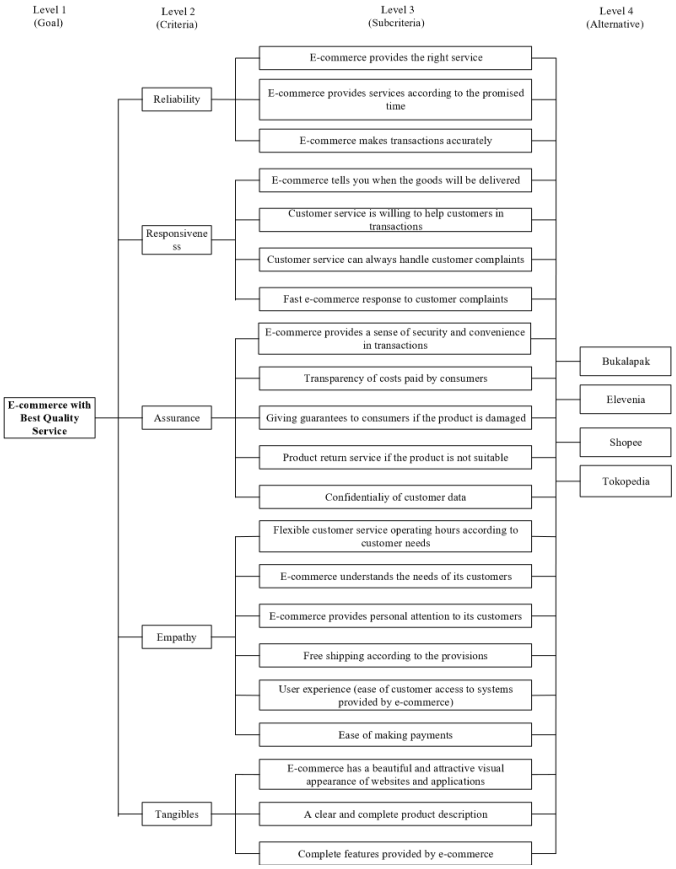


Fig. 4. The Framework of e-commerce Service Quality [5].

c) Questionnaire – prepare for building criteria of e-commerce website model: The questionnaire can get the public's view of e-commerce and help modify the model. And through the response of the people, we can see whether our research is on the right track. We use a questionnaire to verify whether our model has a mass basis and is widely accepted by the public. The questionnaire was designed for adults over 18 years old who are potential customers. Forty-three people were invited to fill in the questionnaire, of which three were not completed, which was invalid. So there are 40 valid questionnaires. Among them, there are 20 males and 20 females, and college students account for the most, accounting for 75%. The remaining 23-30 years old, 31-40 years old, and 51-60 years old accounted for 12.5%, 2.5%, 5%, and 5%. Here are Table I shows the results from the questionnaire.

TABLE I. RESULT FROM QUESTIONNAIRE

Reaction / Problem	Total Disagree	Disagree	non-agree or disagree	agree	total agree	Total
1	0	1	8	10	21	40
2	0	4	9	9	18	40
3	0	3	9	14	14	40
4	0	0	13	12	15	40
5	0	1	9	15	15	40
6	0	2	6	14	18	40
7	0	1	8	8	23	40
8	0	0	9	6	25	40

d) Structure of Questionnaire

1) You want to find the website you want to visit through keywords quickly.

2) If the website you visit doesn't respond quickly, you will feel impatient.

3) You want to see an attractive user interface (the first page you see on the site).

4) You tend to visit a creative website that gives you a visual experience.

5) If you can't find what you want through the website's search engine, you will feel distressed.

6) You want the site to provide classification options to help you search for the items you want.

7) You tend to visit websites with secure payment methods (Alipay, WeChat).

8) You tend to visit a website that provides after-sales service.

e) Data Visualization: The data in Fig. 5 obtained shows that 77.5% of the respondents hope to get the website they want quickly through keywords. 67.5% said they would be impatient with websites that can't respond quickly, while 70% wanted to see attractive user interfaces. 67.5% of people want to have a visual experience website. Nearly 75% of the people feel distressed. If they can't find the products they want by using the website's search engine, 80% of the people also hope that the e-commerce website can provide classification options to help them view the products. Nearly 80% (77.5%) love Alipay and WeChat's payment sites, and the percentage of people also hope that e-commerce can provide customer service.

f) Hypothesis Foundation – build a hypothesis model of criteria of e-commerce website: Based on the results obtained by the previous face method and our experience of browsing popular e-commerce websites (such as Taobao and Jingdong) and shopping experience, we provide a hypothesis model to combine with the previous research abstract model (combination process).

The framework of an e-commerce website, as shown in Fig. 6, consists of the user interface, readability, searchability, service safety, website accessibility, future value. This is based on the behavior of customers. The user interface is the consumer's first impression of an e-commerce website, which affects their actions. For readability, it represents the degree of customers understanding the product's basic information on an e-commerce website. For searchability, it represents the search capability of an e-commerce website. For service safety, it reflects the security level of the e-commerce website. For website accessibility, it reflects the technology level of this e-commerce website. Finally, for future value, this represents the degree to which the site has plans and goals for the organization. Table III demonstrates the visual of this framework. Here is the framework of the e-commerce website in a concise version.

The criteria of the e-commerce website model, as shown in Fig. 7, demonstrate the subcriteria of these six main parts in construct an effective e-commerce website. It's another version of the framework of an e-commerce website in a more complex way.

Then we combine our established model with the framework of e-commerce service quality, as shown in Fig. 4, by connecting with two perspectives: customers' individual emotions and the structure of e-commerce website to construct the combination model. This can provide a good blueprint for e-commerce enterprises after they make a foundation for themselves. They connected with each other by using the problem domain approach, which means that they will map (dash line) with each other if both blocks discuss the same field in the framework.

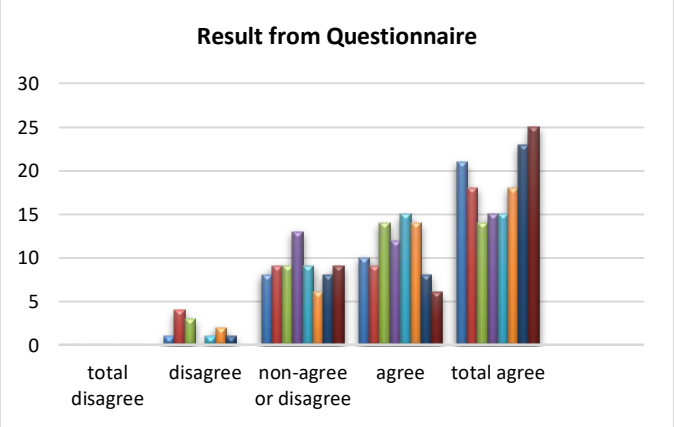


Fig. 5. Result from Questionnaire in Histogram.

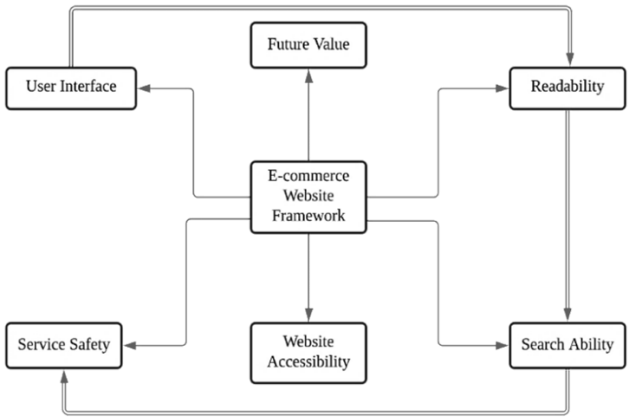


Fig. 6. Framework of E-commerce Website.

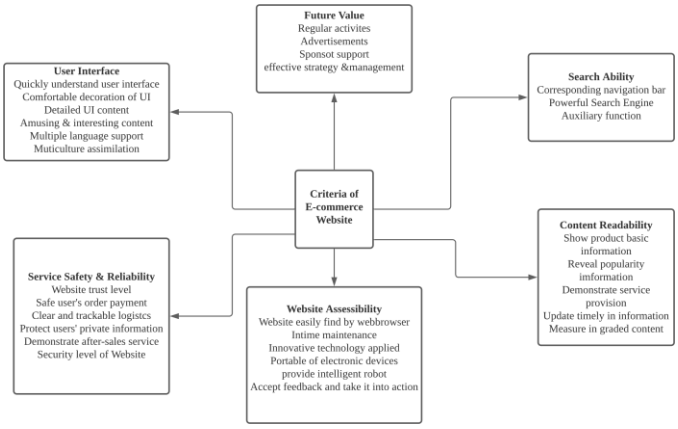


Fig. 7. Criteria of E-commerce Website.

g) Verification of Combination of Model of E-commerce: Based on the above, we start from the consumer behavior, improve a research model, establish their model, and then combine, and finally get a more comprehensive and valuable model. We have also adopted a series of scientific and practical methods. The framework of criteria of the e-commerce website model consists of the user interface, searchability, content readability, service safety and reliability, website accessibility, and future value. We find that we are supported by Sharma and Aggarwal (2019) 's work because they built the model of successful e-commerce based on the hypothesis as shown in Fig. 8, which consist of five main parts to determine the success of e-commerce service with the help of Exploratory Factor Analysis (EFA) method. They also verify it by using Partial Least Square –Structural Equation Modeling(PLS-SEM) [11] by providing loadings, Cronbach's α. And we consider the value of Cronbach's α for each variable in Table II. This value can be regarded as good if they fall between 0.8 and 0.9. From Table II, we can find that each variable can be considered a credible measurement of e-commerce success. All these variables discussed similar topics and included some functions that we used in our hypothesis model of e-commerce website criteria. So our model is greatly supported by this credible data.

h) Analysis of One Special Case - Study interviews from Alibaba: Leavy, B. (2019, Match 18) interviewed Zengming, who is a famous Alibaba strategist. He asked about Alibaba's milestones, a series of turning points, the impact on China's industry, the application of intelligent commerce, the model of mutual cooperation in the network, and the final strategic positioning. Zengming answered these questions in detail. From this interview, we mainly focus on Alibaba's business scope and strategic positioning. The business scope of Alibaba is core business, cloud computing, logistics services, micro-financial services [9] mainly.

They also have other initiative actions such as entertainment on the internet. From this, we can think that Taobao, tmall, financial services, and rookie are all based on these kinds of business scope. From this professional interview, we can find that Alibaba initially set up a goal for small and micro enterprises to have their own position and influence. It is a B2B business platform for all other franchised enterprises. Alibaba has been doing this business for a period of time, accumulated some capital, and gained particular strength. Then, in order to prevent other enterprises from seizing market share, they made some adjustments to their goals, such as creating a Taobao website which is a business platform for major e-businesses to join in and allow them to carry out legal business activities on their own platform and making tmall [9] separate and independent, in order to speed up the penetration of China's market.

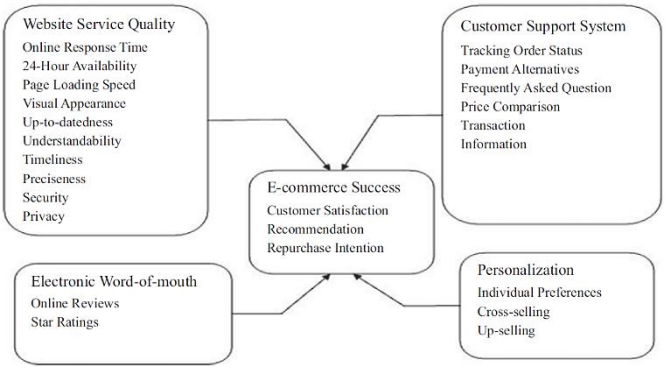


Fig. 8. Model of Successful E-commerce [11].

TABLE II. FIVE BIG VARIABLES TO DETERMINE THE SUCCESS OF E-COMMERCE [11]

Variable	hems	Loadings	Cronbach's a	CR
Website service quality (WSQ)	WSQ1	0.66	0.805	0.80
	WSQ2	0.756	0.799	
	WSQ3	0.493	0.802	
	WSQ4	0.496	0.808	
	WSQ5	0.598	0.802	
	WSQ6	0.604	0.802	
	WSQ7	0.433	0.804	
	WSQ8	0.465	0.802	
	WSQ9	0.418	0.808	
	WSQ10	0.349	0.806	
Customer support system (CSS)	CSSI	0.502	0.803	0.72
	CSS3	0.334	0.807	
	CSS3	0.506	0.801	
	CSS1	0.709	0.806	
	CSS5	0.559	0.812	
	CSS6	0.497	0.803	
Personalization (PER)	PER1	0.798	0.822	0.76
	PER2	0.787	0.820	
	PER3	0.859	0.815	
Electronic word-of-month (EWOM)	EWON11	0.820	0.810	0.85
	EWOM2	0.798	0.808	
ecommerce system success (ISS)	ESS1	0.624	0.806	0.78
	ESS2	0.791	0.809	
	ESS3	0.709	0.803	

As time goes on, Alibaba hopes to achieve something in other fields, such as finance and technology. They created two companies: Alipay and cloud computing. Among them, Alipay has a profound impact on China's payment system. Among them, Alipay had a profound effect on China's payment system. People use Alipay to pay for goods, make commercial loans, and transfer accounts. Alipay even contracted for various platforms' two-dimensional code scanning services, such as providing health codes to provide passage and implementing epidemic prevention policies. Or use two-dimensional code to go shopping, daily behavior. Taobao is an e-commerce website whose essential function is to provide customers with online shopping services. Among them, he innovated away: at that time, it was very advanced. Taobao provided leasing services for businesses, that is, let businesses open shops online. Companies can rent a physical store in the urban area to sell goods without high rent. They only need to use the platform provided by Alibaba to complete their business online. From this point of view, they are mutually beneficial and win-win on the premise that there is no relationship between exploitation and exploitation. At the same time, Zengming also said in the dialogue that they have an intelligent business expansion. Through the internet and big data, customer companies are combined to operate together. To find the corresponding laws and survive in this cruel market.

Most importantly, Alibaba is very clear about its position. They know what they want to do and the significance of doing it. They put forward the theory of point, line, and surface [9], and they take on the role of the line as the provider of infrastructure. And they are bigger and stronger, which shows that they have done an excellent job in the part of the line. So they can succeed and have the energy to do other business sectors. They also mentioned the importance of intelligence data combined with the internet to provide better service for customers. They focus on making customers have better e-commerce service and provide a smooth communication approach between customers and business. Therefore, ant financial services are born to help provide loans to promote the sustainable development of the economy. As shown in Table III, Alibaba's strategies demonstrate their theory in detail from four main aspects: feature, point, line, and plane.

i) Construct solutions for e-commerce enterprise: First and foremost, e-commerce enterprises should determine what industry they belong to. The distribution of the e-commerce industry in China, as shown in Fig. 9, demonstrated Chinese e-commerce industry distribution. We can find that clothing, shoes, and hats, textile and chemical fiber, agriculture, forestry and animal husbandry, digital home appliances, mechanical equipment, chemical plastics, food, sugar and wine, building materials, hardware tools, medical and pharmaceutical products account for a high proportion to a low balance. These are shown in Fig. 9 clockwise, which started from clothing and shores. And others industry occupies about 31.90 %. It's not difficult to find that many industries are related to e-commerce. What kind of products do enterprises need to provide depends on the company's technology type and development direction? Companies need to pay attention to which industries have good development potential and the broad market. To pay attention to these industries and prepare for the transformation of these industries.

TABLE III. THE THREE STRATEGIC POSITIONS IN A BUSINESS ECOSYSTEM [9]

Strategic Position			
Feature	**Point**	**Line**	**Plane**
Value proposition or service	Selling a function or capability	Creating a product	Connecting related parties
Competitive advantage	Expertise	Value, cost, and efficiency	Matching efficiency
Organizational capabilities	Simple; no complex operations	Streaming and optimizing workflows	Designing systems and institutions to mediate relationships
Core strategy	Advance into the next rising plane and find one's niche in a fast-growing line	Use the resources of robust planes to incorporate strong points	Enable the growth of points and lines
Web-celeb analogy	Factories, clothing designers	Ruhan	Taobao, Weibo

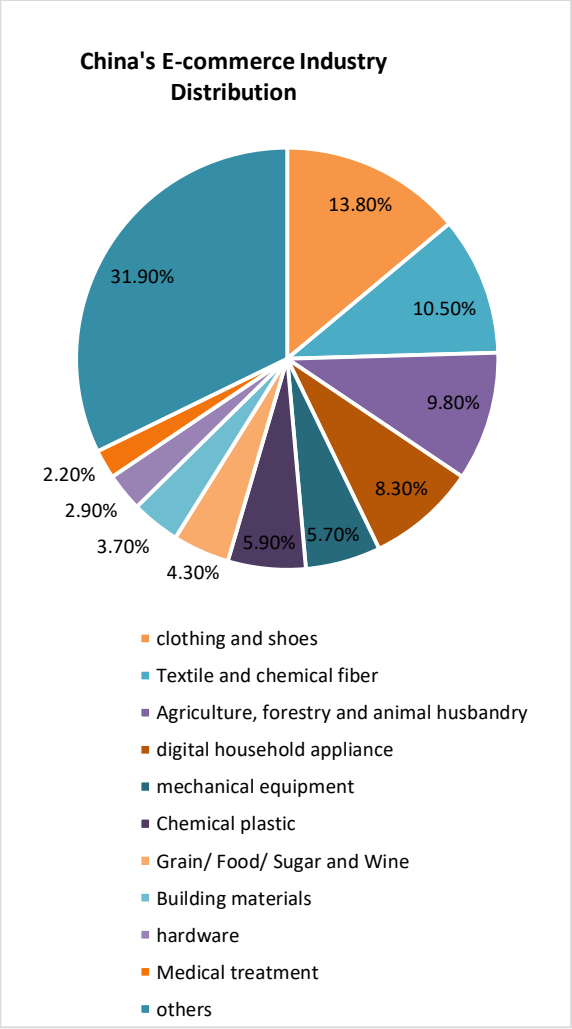


Fig. 9. Chinese E-commerce Industry Distribution [3].

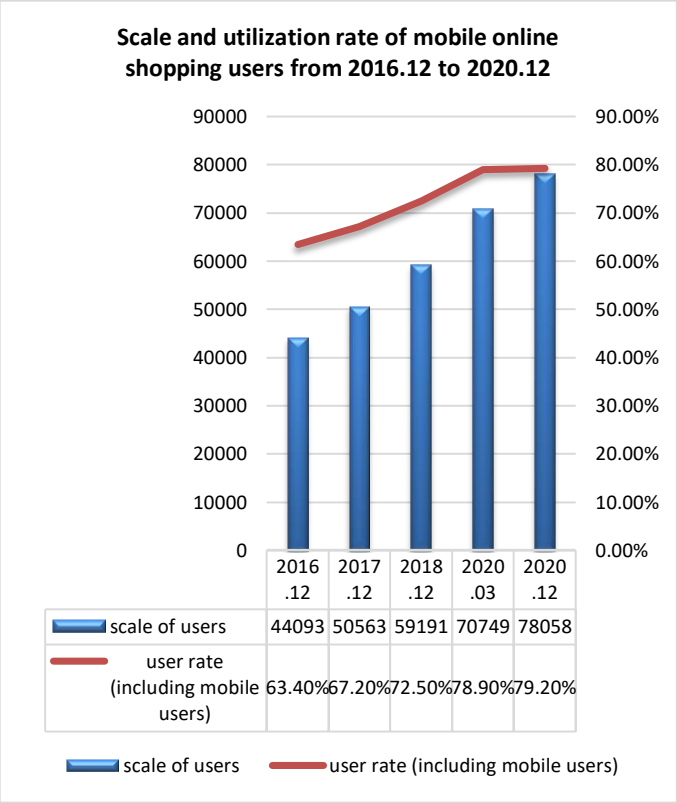


Fig. 10. Scale and Utilization Rate of Mobile Online Shopping users from 2016.12 to 2020.12 [2].

Furthermore, China's e-commerce sector is still showing a good development trend in China. From the condition of users using electronic devices as shown in Fig. 10, we find that the Internet-scale is increasing year by year. By 2020, the ranking will reach 780.58 million people, and about 80% of users will use online shopping. Internet in China is still in progress, promotion. The market scale is gradually expanding. As shown in internet listed companies as shown in Fig. 11, 12.6% of the listed enterprises are e-commerce enterprises. It shows that many enterprises still choose the field of e-commerce. From this, we can see that e-commerce in China has a good development prospect and is still creating a new high.

Next, Enterprises need to know what kind of customers they serve, so they are roughly divided into four categories according to customer types. They are B2C (business to customer), B2B (Business-to-Business), C2C (Consumer To Consumer), B2G (business-to-government), respectively. And B2C and B2B are the main forms of e-commerce enterprises in China. Companies need to choose partners according to their products or services to determine the type of e-commerce company. The group of Tan also used the Perceived eReadiness Model, as shown in Fig. 12, which will help them analyze the situation of the e-commerce industry in China. [12] Based on this model, they have made corresponding improvements and put forward the factors that affect the development of e-commerce in China. They also used the nine most essential variables (factors) they considered to measure their degrees of reliability by using cronenbach alpha, as shown in Table IV, which measures the reliability. The nine variables are awareness (A), Business Resources (BR), Commitment (C), Governance (G), Government eReadiness (GVeR), Human Resources (HR), Market forces eReadniess (MFeR), Supporting industries eReadiness (SIeR), and Technological resources (TR). [12] The variable of cronenbach alpha will be higher, which shows that it tends to be more favorable for e-commerce development. However, human resources may not be considered as a negative part in China of 2020. In addition, Fig. 10 shows that about 78058 million people have electronic devices to surf on internet in 2020. This may be much helpful to improve the development of human resources. Tan (2007) also proved that this model is appliable in China, as shown in Fig. 8. This model provides an excellent structure even for business starters who want to build a company with other people. This is because this model demonstrates some essential elements in building a perceived or hardworking organization. This model gave the standard to e-commerce enterprises about the quality of employees, executors, strategies, external forces, practical enterprise's managing structure. In concrete, this model puts forward the following requirements for the company: the level of employees and the reserve of professional knowledge need to meet the standards to be competent for their work. Companies need a real-time view of current events, external industry information, and their field and e-commerce field to predict future development. Prepare for transformation and revision. At the same time, this model also puts forward requirements for employees' work experience and the application of electronic aids. Especially for strategic planning, employees at the forefront of the field are essential. At the same time, it also puts forward a series of essential requirements for decision-makers, requiring them to have a clear strategic plan, which can be related to previous interviews, because Alibaba is the leader of B2B in China. This model can be combined with Alibaba's strategy for reference. At the same time, this model also requires to be aware of external forces. For example, government policies, administrative decrees, preferential policies, market changes (changes in demand), and cooperation between financial and trust institutions. Enterprises can use this model to initially form the company structure, thinking direction, and employee type. And then, they can move to the production of the website as the company's facade, portal, the first step of customer and enterprise communication. Then we use this model to combine this model, which is about applying this model into the hypothesis model combined e-commerce service quality's model, strategy in one particular case into a more comprehensive and referential combination model as shown in Fig. 12.

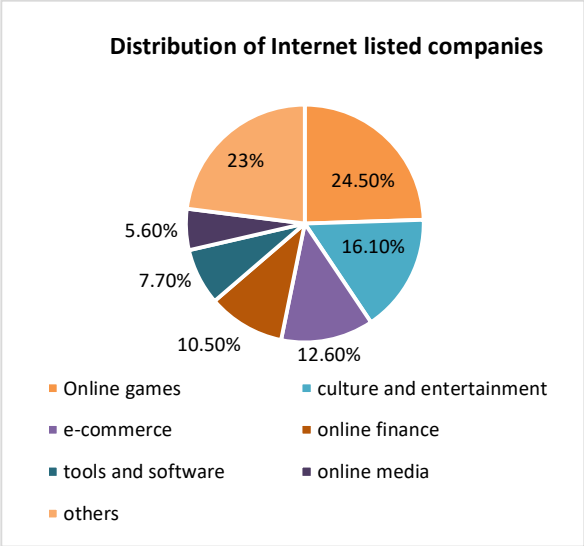


Fig. 11. Distribution of Internet Listed Companies (China Internet Network Information Center) [2].

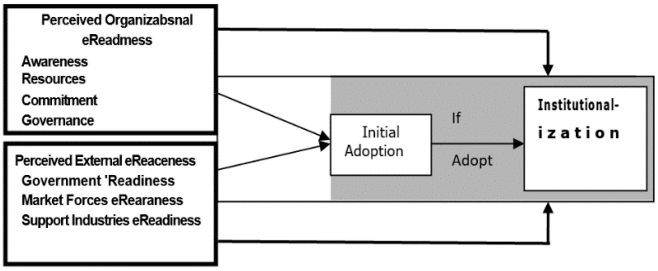


Fig. 12. Perceived eReadiness Model Framework [12].

TABLE I. MEASURE OF RELIABILITY IN EACH TERM OF PERCEIVED EREADINESS MODEL [12]

Instrument reliability	
	Cronebach alpha
A	0.91
BR	0.70
C	0.91
G	0.92
GveR	0.78
HR	0.59
MfeR	0.86
SieR	0.83
TR	0.82

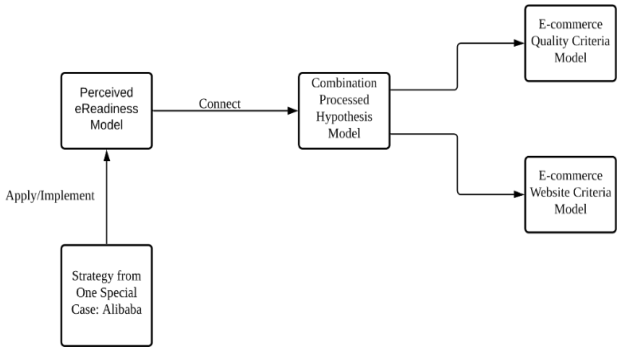


Fig. 13. Comprehensive and Referential Combination Model.

V. RESULT

The Comprehensive and Referential Combination Model as shown in Fig. 13, in the last part, which consists of the Perceived eReadiness Model as shown in Fig. 12, applied strategy as shown in Table III, Combination Processed Hypothesis as shown in Fig. 4 and Fig. 7, which are mapped in the domain). This model has been proved to be reasonable, supported by data, and comes from scholarly journals. Next are the detailed versions of the e-commerce website criteria model.

1) Part of Combination of Hypothesis Model - Details of criteria of e-commerce website.

a) User Interface

- Customers understand the user interface quickly When browsing the website. a. Website has tutorial guidance. b. The user interface is simple and easy to understand.
- Customers feel comfortable in the decoration of the user interface a. font size b. shape c. font color d. the suitability of background color, typesetting.
- The content of the user interface is detailed. a. Related to the theme of the website b. The contents are arranged in an orderly and regular way. c. There exist pictures to help explain the contents and other auxiliary materials d. There is enough information to enable customers to have a complete understanding of the product.
- The contents are interesting, a. cartoon elements or video provided or any other visual dynamic aids b. List recommended items shown in the user interface. c. Provide Popular current products.
- Support multiple languages and accept the local culture.

b) Search Ability

- There is a corresponding navigation bar on the e-commerce website. a. it's conspicuous so that users can easily find it. b. There is a classification or a subdivision of the classification. c. There is a special effect to display the column subdivided under the classification.
- Provides a powerful search engine. a. Users can find the goods they want by entering keywords. b. accurate 2. valid links (user can access these links).
- Auxiliary function a. When inputting text into the search box, there is help content display. b. The speed of search result.

c) Content Readability

- Provide the product's basic information: pictures, size, type, price, inventory surplus, shipper and address, price of each piece (unit).
- Provide the popularity information: monthly sales volume, the number of people who view and buy, comment information.
- Provide the service: seven days no reason to return the goods, customer service consultation,

- Timeless of the product information a. Update regularly. b. No grammatical errors, typos.
- Graded content: ensured that minors could not search for adult contents.

d) Service Safety and Reliability

- The trust level of the website, a. the last line of the user interface, has key information: company information, address, telephone, zip code, and copyright. b. There is a certificate of business granted by the state and a series of trust licenses.
- The user's order payment approach is safe or even multiple, including Apple Pay, WeChat, Alipay.
- The logistics are precise. a. the transparency of the delivery mode of goods b. Easy to grasp the direction (e.g., using GPS and graph).
- Promise to protect customers' personal information from leakage. a. protects the customers' filling orders. b. Protect the process of the transaction.
- Perform product after-sales service. a. Support comments from buyers. b. 24 hours or most of the time, customer service consultation is available. c. Provide suggestion channels: telephone contact information provided, email feedback. d. Provides real-name registration.
- The security level of the website (HTTPS).

e) Website accessibility

- Users can easily find the site from various search engines.
- Intime maintenance log a. Update records. b. Update relevant personnel.
- New technology is used in the website. a. The speed of opening the website: network response, display interface speed. b. A working database: record the relevant information of various commodities, which information can change in real-time.
- The availably of computers and mobile phones a. Succeed in displaying in other electronic products display screen (compatibility).
- Design intelligent robot feedback to help answer questions.

f) Future Value: An E-commerce website has regular activities, advertisements, sponsor support, organized and effective strategy, and management.

2) Validation for Established Model

a) Find Support from Customers: According to questionnaire data, we find that nearly 75% of them agree with our perspective of establishing the framework and criteria on average. This means that our starting point is correct, and we are suitable to analyze the customer's behavior and what they want.

b) Theoretical Support from Previous Research: Apart from we mentioned Data support in the previous, here is the theoretical backing of prior research. Van Der Merwe's team (2003, Dec 1) quoted other people's opinions about the evaluation website; they think that the framework and criteria should be established from the customer's buying circle. Next, Lu J. (2003) provided a way of building a company assessment framework, website capabilities, and customer assessment. [10] They set 8 groups to establish criteria for each part. In a way, part of his research is similar to our study because they also consider the factors of customers.

What's more, Hasan et al. (2013) got assistance from three other research which is to use framework and criteria to evaluate e-commerce websites. [4] They even observed a series of actions generated by users using e-commerce websites, which they recorded in real-time. In short, they kept their behavior by having the subjects perform the same task. This is similar to our study of customer behavior. These studies provide great theoretical help for our research.

VI. Recommendation

Enterprises need to go through the first step in the Comprehensive and Referential Combination Model. Then they can go through the next step, which is about designing e-commerce websites that are parts of a combination of hypothesis models. Companies that want to build an e-commerce website and carry out online shopping business use this model to design user interface, search engine, commodity content interface, and payment interface. The content of each criterion can be used as a reference to remind the creator to consider joining or make corresponding fine-tuning as needed. After the design, they invite some volunteers or experts to put forward suggestions and opinions on this website. Then release and open users to enter the website for testing and make corresponding adjustments according to the feedback. If the company has completed the production of the website, the evaluation is given by testers after testing the website is not ideal. Then you can refer to the model for comparison and add or modify the missing part. To improve the service level. Also, the company can refer to Alibaba's operation mode and connect the company's information with customers and partners to ensure the timely transmission, delivery, and exchange of information. At the same time, according to current events and market changes, we should make corresponding adjustments to the strategic plan and clarify our role in the whole business field, that is, what we can provide, what we want to do, and the purpose of doing these things so that these can increase the level of future value.

VII. Conclusion

The significance of our results is to solve the problem of improving the service quality of e-commerce websites by using the Comprehensive and Referential Combination Model by implementing a Step-By-Step, Bottom-Up approach that constructed efficiently. In the result, we also suggest how to use this model, which is in the recommendation section.

Although this study proposes solutions and processes to improve the service quality of e-commerce websites and has specific theoretical and public support, it lacks experts to verify

the design in person. We provide what we need to pay attention to and what kind of functions. However, we are lack of detailed quantitative and detailed description of each criterion. For example, under what circumstances and how many large fonts are used. This requires a lot of work and some professionals to investigate and analyze e-commerce websites in various fields.

Moreover, we need to make an overall analysis of this joint model. For example, we need to find out the related problems for this model and then distribute the questionnaire to experts. Because this model involves many aspects and many questions need to be prepared, it is inappropriate to distribute this questionnaire to the internet. The public does not have professional knowledge in this field. Therefore, the right person to fill in the questionnaire is someone familiar with the field of e-commerce. We may move towards this in the future. These will take a lot of time. If there is time support, the instrument and the corresponding helpers will provide support. This research may continue in the future or even develop in other directions, for example, using this model to evaluate popular websites for e-commerce research. We can also keep the original order and optimize the combination model's design, which contains several models, by implementing more relevant variables.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support from Leading Talents of Provincial Colleges and Universities, Zhejiang-China(#WB20200915000043). With Wenzhou-Kean University.

REFERENCES

[1] Burke D. (2015, May). 5 Criteria for a Successful eCommerce Website Avoiding mistakes that would be embarrassing in a retail store. Retrieved from https://www.getfused.com/blog/posts/5-criteria-for-a-successful-ecommerce-website-87

[2] China Internet Network Information Center (2021, Feb. 23). The 47th statistical report on the development of Internet in China. Retrieved from http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/202102/P020210203334633480104.pdf

[3] Foresight Industrial Research Institute. (2016, Apr., 21). Market Outlook and Investment Strategy Planning Analysis Report of China E-Commerce Industry (2016-2021). Retrieved from https://bg.qianzhan.com/report/detail/459/160421-9a14f4f1.html

[4] Hasan, L., Morris, A., & Probets, S. (2013). E-commerce websites for developing countries - a usability evaluation framework. Online Information Review, 37(2), 231-251. doi:http://dx.doi.org/10.1108/OIR-10-2011-0166

[5] Ishak, A., Ginting, R., & Wanli, W. (2021). Evaluation of e-commerce services quality using Fuzzy AHP and TOPSIS. IOP Conference Series: Materials Science and Engineering, 1041(1), 012042. https://doi.org/10.1088/1757-899x/1041/1/012042

[6] Jenn Vande Zande. (n.d.). What is e-commerce? Definition, benefits, examples. Retrieved from https://www.the-future-of-commerce.com/2020/01/19/what-is-e-commerce-definition-examples/

[7] Kogan D. (2013, July 24). 8 Ways To Evaluate Your eCommerce Website. Retrieved from https://www.1digitalagency.com/8-ways-to-evaluate-ecommerce-website/

[8] Kunst A. (2019, Dec 20). U.S. online shoppers with negative shopping experiences 2017. Retrieved from https://www.statista.com/statistics/705252/shopping-experience-of-online-shoppers-in-the-us/

[9] Leavy, B. (2019, March 18). Alibaba strategist ming zeng: "smart business" in the era of business ecosystems. Strategy and Leadership. Emerald Group Publishing Ltd. https://doi.org/10.1108/SL-01-2019-0006

[10] Lu, J. (2003). A model for evaluating E-commerce based on Cost/Benefit and customer satisfaction. Information Systems Frontiers, 5(3), 265. Retrieved from https://kean.idm.oclc.org/login?url=https://www.proquest.com/scholarly-journals/model-evaluating-e-commerce-based-on-cost-benefit/docview/232037789/se-2?accountid=11809

[11] Sharma, H., & Aggarwal, A. G. (2019). Finding determinants of e-commerce success: A PLS-SEM approach. Journal of Advances in Management Research, 16(4), 453-471. doi:http://dx.doi.org/10.1108/JAMR-08-2018-0074

[12] Tan, J., Tyler, K., & Manica, A. (2007). Business-to-business adoption of eCommerce in China. Information and Management, 44(3), 332–351. https://doi.org/10.1016/j.im.2007.04.001

[13] Van Der Merwe, R., & Bekker, J. (2003, December 1). A framework and methodology for evaluating e-commerce Web sites. Internet Research. https://doi.org/10.1108/10662240310501612

[14] Watrobski, J., Ziemba, P., Jankowski, J., & Wolski, W. (2016). PEQUAL-E-commerce websites quality evaluation methodology. In Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016 (pp. 1317–1327). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.15439/2016F469

[15] Teo, T.S.H. & Liu, J. (2007). Consumer trust in e-commerce in the United States, Singapore and China. Omega, 35(1):22-38.

[16] Sun, P. C., Luo, J. J., & Liu, Y. L. (2010, November). Perceived Risk and Trust in Online Group Buying Context. In Information Management, Innovation 323 Management and Industrial Engineering (ICIII), 2010 International Conference on (Vol. 3, pp. 660-663). IEEE.

[17] Sulaiman, A., Jaafar, N.I., & Kadam, P. (2005). Factors affecting online purchasing among urban internet user sin Malaysia. Fourth International Conference on eBusiness, November 19-20. Bangkok, Thailand.

[18] Stewart, D. W. (1981). The application and misapplication of factor analysis in marketing research. Journal of Marketing Research, 18(1), 51-62.

[19] Saibaba, G., and P. Vaidya Sanivarapu. "Developing an Userfriendly Online Shopping Web-Site." Indonesian Journal of Electrical Engineering and Computer Science 12, no. 3 (2018): 1126-1131.

[20] Al-Qirim, N., 2007. The adoption of eCommerce communications and applications technologies in small businesses in New Zealand. Electronic Commerce Research and Applications, 6(4), pp.462-473.

[21] Sharma, S. and Crossler, R.E., 2014. Disclosing too much? Situational factors affecting information disclosure in social commerce environment. Electronic Commerce Research and Applications, 13(5), pp.305-319.

[22] Yang, Z., Shi, Y. and Yan, H., 2016. Scale, congestion, efficiency and effectiveness in e-commerce firms. Electronic Commerce Research and Applications, 20, pp.171-182.

[23] Hemn Barzan Abdalla, Lu Zhen, Zhang Yuantu, 2021. A New Approach of e-Commerce Web Design for Accessibility based on Game Accessibility in Chinese Market. International Journal of Advanced Computer Science and Applications, 12, pp.1-8.

Facilitating Personalisation in Epilepsy with an IoT Approach

S.A McHale, E.Pereira
Department of Computer Science
Edge Hill University
Ormskirk, UK

***Abstract*—The premises made in this paper put the future of personalisation in epilepsy into focus, a focus that shifts from a one-size fits all to a focus on the core of the epilepsy patients' individual characteristics. The emerging approach of personalised healthcare is known to be facilitated by the Internet of Things (IoT) and sensor-based IoT devices are in popular demand for healthcare providers due to the constant need for patient monitoring. In epilepsy, the most common and complex patients to deal with correspond to those with multiple strands of epilepsy. These extremely varied kind of patients should be monitored precisely according to their identified key symptoms and specific characteristics then treatment tailored accordingly. Consequently, paradigms are needed to personalise this information. By focusing upon personalised parameters that make epilepsy patients distinct this paper proposes an IoT based Epilepsy monitoring model endorsing a more accurate and refined way of remotely monitoring the 'individual' patient.**

Keywords*—IoT; healthcare systems; smart healthcare; personalisation

I. Introduction

By integrating IoT sensor-based devices deployed remotely and personalised patient data into a combined monitoring framework a vision of personalisation is realised. This study revealed some irrefutable evidence derived from patient profile analysis and experimental data that seizure detection using sensors positioned on different parts of a patents body ultimately makes an impact on the monitoring of epilepsy, endorsing that modern computer science is providing a timely chance for a more personalised approach to the monitoring and management of epilepsy.

The chances of capturing seizure data can be greatly increased if a correctly assigned sensor is placed on the correct part of the patient's body and ultimately, such a concept could 'enhance the overall monitoring scheme of a patient usually performed by caring persons, who might occasionally miss an epileptic event' [1].

This paper is organised as follows. In section II and section III the state of the art is analysed; the complexity of epilepsy together with smart healthcare monitoring approaches are highlighted. There is also a focus upon the sensors available for epilepsy following on with an emphasis on the limitations for a personalised approach. Section IV presents the driving questions in this study and describes the experiment and findings from capturing seizure data. Section V introduces the proposed IoT based Epilepsy monitoring model and reveals the PMP (Personalised Monitoring Plan) framework whereby the patient can be matched with the correct device, while Section VI presents how this was evaluated and Section VII outlines long term use. Concluding remarks are drawn in Section VIII.

II. Motivation

Epileptic seizure monitoring and management is challenging. Most current studies of epileptic seizure detection disclose drug resistant epilepsy still lacks an ultimate solution, despite the increase in anti-epileptic drugs [2].

Epilepsy is not a single disease, but a family of syndromes that share the feature of recurring seizures. In some instances, it may be related to a genetic aetiology, or it can occur in association with metabolic disorders, structural abnormalities, infection or brain injury [3].

In the United Kingdom epilepsy affects 3 million people and in the United States it is the 4th most common neurologic disorder, only migraine, stroke, and Alzheimer's disease occurs more frequently [4]. There are around 60 different types of seizure and a person may have more than one type. Seizures vary depending on where in the brain they are happening. Some people remain aware throughout, while others can lose consciousness [5].

Aside from their unpredictability, the worst part of having seizures is their utter complexity. The complex nature of epilepsy is noticeable in the variation of seizures types and symptoms between one patient and another. Distinguishing or classifying an individual epilepsy patient makes it difficult to manage and monitor. The negative impact of uncontrolled seizures spreads beyond the individual to affect their family, friends, and society. Chronic anxiety is experienced by the families and friends of people with epilepsy and many lives are adjusted to ensure the safety of their loved one. Novel approaches to epilepsy treatment are still greatly needed [6] novel therapies that better manage and monitor seizures as well as technology can help to handle the consequences of seizures.

Insufficient knowledge about epilepsy, which is a very common disorder, has a great and negative impact on people with epilepsy, their families and communities, and the healthcare systems. There is need for a better understanding of the disease to make way for new approaches to monitor it.

In the modern day of personalised medicine and rapid advancements in IoT a question that needs addressing is whether epilepsy monitoring can benefit from personalised

approach. Can the IoT have the potential to significantly improve the 'patients' daily lives whose seizures cannot be controlled by either drugs or surgery [7]?

A. Smart Healthcare Monitoring Approaches

In the history of time it is only relatively recently that computers began to assist healthcare monitoring, in 1950s' patients began to be continuously monitored by computerised machines [8] and clinical monitoring was first envisaged in the home [9]. For computer assistance to epilepsy it was not until 1972 in the field of imaging, when computerized tomography (CT) was invented by the British engineer Godfrey Hounsfield [10] and only in recent decades where specific epilepsy healthcare 'monitoring systems' have been proposed.

Much of this recent growth being due to the advent of current IoT technology whereby the rise of 'smart environment' approaches to healthcare monitoring is witnessed. There are many IoT approaches for the monitoring and management of epilepsy many of which encompass a network of connected smart devices which are equipped with sensors either embedded in clothing or smart phones, to either detect, predict or manage epilepsy. Discoveries disclose how IoT is utilised to support the ever-growing trend of personalised healthcare. These recent 'smart' approaches in healthcare demonstrate the trend toward 'sensor use' and 'remote monitoring'.

III. Related Work

Researchers are bounding toward the new generation of smart technology and IoT (Internet of Things). Novel devices such as smart watches, smart bands & smart clothing are all competing for the ultimate solution. Yet it is found there is limited research which focuses upon the concept of a more holistic, personalised approach to help manage epilepsy.

One study deemed the significance of attention on smart technologies and its potential to identify early indicators of cognitive and physical illness [11] and observed that researchers have argued and predicted that assessing individuals in their 'everyday environment' will provide the most 'valid' information about everyday functional status [12].

Indeed, there is evidence recently of this indication as several IoT platforms to manage & monitor healthcare remotely, are observed. For example, one IoT paradigm comprising of Wireless Health Sensors (WHS) permits the continuous monitoring of biometric parameters such as pulse rate, pulmonary functional quality, blood pressure and body temperature [13]. This IoT paradigm is being used to assist predictive analysis via smart healthcare systems by a medical practitioner. Using sensors connected to Arduino patient status is tracked, and by a Wi-fi connection data is collected and transmitted and can receive user requests. This data is shared with doctors through a website where the doctor can analyse the condition of the patient and provide further details online and intimate patient about future severity well in time [13].

A. Sensors for Epilepsy

EEG, an electroencephalogram is a recording of brain activity. This is the chief gold standard method used within hospitals to detect and monitor seizures. Several approaches have been reported with the aim to embed this method in other settings and platforms. Developments in some topics have been published, such as modelling the recorded signals [14] [15] or the design of portable EEG devices to deploy such models.

As an alternative and sometimes supplement to EEG there exist many sensors embedded in clothing or worn on the body to obtain bio-signals such as gyroscopes, accelerometers, pulse rate, temperature sensors, magnetometers, galvanic skin response sensors (GSR), implanted advisory system, electromyography, video detection systems, mattress sensor, and audio systems [16].

A large amount of apps have been published more recently especially in the commercial sector for the detection and management of seizures using either the Smartphone sensors or external sensors, for example Epdetec [17] and Myepipal [18] and web logging which facilitates the way a patient records daily information concerning her/his epileptic events, medication, and news, My Epilepsy Diary [19] and Epidiary [20] . Another app attracting attention and recently reviewed in the press is the Alert App by Empatica. This app sends caregivers an automated SMS and phone call when it detects unusual patterns that may be associated to a convulsive seizure [21] yet it is only designed to work with the Embrace Smartband by Empatica and can prove expensive for the user [22].

Regrettably, there are few specific sensor detection options for each specific seizure type, this is an imminent requirement for patients and their carers. Ideally when choosing a seizure detection device, the patient-specific seizure semiology's should be considered [16]. Thus, highlighting the need for a type of monitoring that distinguishes one patient from another and depicting the need for devices to pinpoint the patient-specific signs and symptoms.

B. Addressing the Gaps

Despite the focus in literature on smart healthcare monitoring approaches there is limited emphasis on the embracing of a truly personalised approach for epilepsy as previously described. Even though the 'diversity' of epilepsy is acknowledged and has been identified in other studies, i.e. by highlighting the importance of distinguishing each 'seizure type', there is still a gap to address such parameters. The 'seizure type' is just one of many parameters that can distinguish one seizure patient from another. Therefore, these very individual characteristics can be further identified to address the challenge to achieve a truly personalised approach to managing epilepsy.

More so recently it is recognised that devices should specially take into account the user's seizure types and personal preferences [23], focus should be shifting not only on the desires of the users but seizure detection devices should be able to 'adapt' to the patient's characteristics and seizures [23].

It is already becoming known that wearing sensors on the body is starting to be popular, as observed recently in a 2018 study where a great interest was highlighted in the use of wearable technology for epilepsy carers, this being independent of demographic and clinical factors and remarkably outpacing data security and technology usability

concerns thus demonstrating the vital factor of comfortability [24]. Yet as discovered during a review to select the best sensor for each individual patient there was limited data on which was the best sensor for each seizure type, this was unfortunate despite an internationally active research effort, signifying the gap in knowledge, again, for understanding the individual epilepsy patient [16].

IV. Experiemnt and Findings

This section discusses the experiment that was performed to capture seizure data, obtained from sensors, which are positioned on different parts of the patient's body. This was done to test the assumption that it is 'the individual profile' that makes the difference in which device to choose. The results from this experiment are used to inform a typical model or a PMP (Personalised Monitoring Plan) discussed in the next section.

The actual 'sensor', and their 'position' (worn by the patient) are significant for epilepsy and the focus in the experiment was on how patients exhibit behaviour, rather than any actual testing of devices. It was therefore important to choose the most accurate sensors for monitoring epilepsy; those were found to be the accelerometer and heart-rate sensors, although latest studies suggest making use of other sensors too such as peripheral temperature, photo plethysmography (blood circulation), respiratory sensors [25], and galvanic (changes in sweat gland activity) among others [26].

A. Preliminary Investigations

Numerous studies have been previously been conducted with sensors and use for epilepsy [27] [28]. Since the 'gold standard' for epilepsy monitoring is video-EEG monitoring (which takes place within hospitals) [29] the driving questions addressed here were:

1) Can the patient be just as accurately monitored at home with an inexpensive, easily obtainable accelerometer and heart-rate sensor-based device?

2) Can the individual requirements of the patient be pinpointed? If so, is it possible that these sensors can be worn at home (a personalised approach) and be just as effective as using EEG monitoring in the hospital setting?

From the analysis of the patient data it is clear that a patient profile based on particular characteristics can indicate which position the sensor is best placed on the patient's body.

Sample patient profiles where selected based upon criteria informed from discussions with clinicians. For example, Patient Profile 1 seizures begins with the right arm suddenly raising, therefore can the sensor be placed upon the right shoulder? Patient Profile 4 has a lot of shaking during their Focal Onset Seizures with shaking starting on the left arm so therefore can the sensor be useful attached to the left wrist? Whereas Patient Profile 5 begins their seizures with severe tremors on the right leg, can the sensors detect movement and heart-rate changes with sensor in this position?

During the investigation practicable devices to use in the experiment to monitor epilepsy were analysed. The 'Fitbit Ionic' was chosen as the best option since both the heart-rate and accelerometer can be extracted. The commercial activity device has been used in other studies, most notably recently whereby it used data from more than 47,000 Fitbit users in five U.S. states and data revealed that with Fitbit use the state-wide predictions of flu outbreaks were enhanced and accelerated [30]. This use demonstrates the viability and potential suitability of Fitbit as a healthcare device.

B. Experimental Description

The objectives of the experiment were to assess the movement from the accelerometer sensor and the pulse from the heart-rate sensor in the detection of epileptic seizures. Participants with confirmed epilepsy are recruited. The non-invasive wrist, leg, knee or arm-worn sensors are used to acquire heart-rate activity and movements. The study evaluated the movement from the accelerometer sensor and the pulse from the heart-rate sensor in the detection of an epileptic seizure. Over a period of 5 days the patients were asked to wear the device and continue recording seizures in their seizure diary. The study also evaluated any differences in result due to the 'position' of the sensor on the body together with the patients' acceptability & comfort.

The instructions contained daily forms for the patient to complete, hence, keeping a diary of the times of seizure, if they did not use this method an EEG recording was obtained. This way the actual time stamp of the patients recorded seizure can be checked against the server time stamp observations of the seizure, so for example if the patient records their seizure at 10.20am and the server readings reveal heart-rate peaks and rapid movement from the accelerometer also at 10.20am, then this confirms the server readings match the patients (or EEG) known seizure occurrence, see Fig. 1, Seizure Time Stamps.

C. Experimental Results

The heart-rate and accelerometer sensors used to detect characteristics of seizure events can successfully record seizure data, without need for participant cooperation beyond wearing the sensor-based device, even recharging the battery (battery life is 5 days when fully charged) was not required by the participants. Both the sensors detected the 'shaking' seizures correctly as can be seen in observation d HP4, in Fig. 2, Observations, set 1.

The effectiveness was also verified by "Non-seizure times" which are more easily recognised in sleep due to inactivity demonstrating that the sensors worked properly: see the above random time periods, in Fig. 2 whereby seizures did not occur for HP4 in Observation 'k' and 'i'.

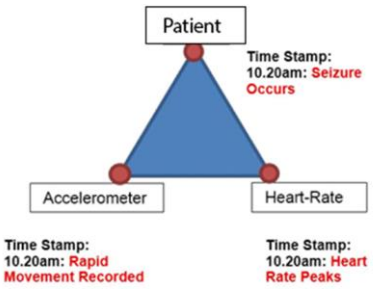


Fig. 1. Seizure Time Stamps.

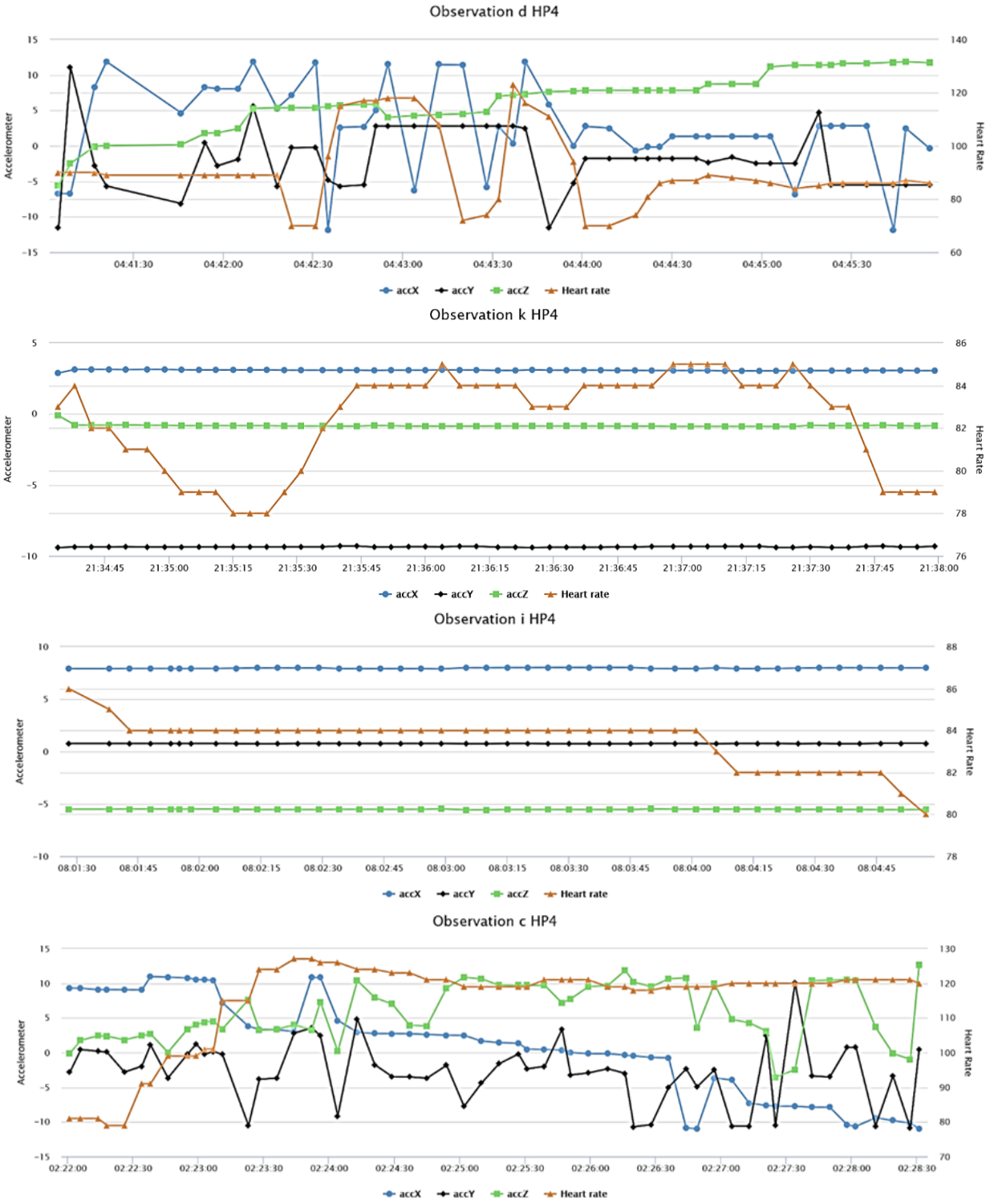


Fig. 2. Observations, Set 1.

Both accelerometer and heart-rate sensors have been used to detect seizures in numerous previous studies [25] but in this study it was found that when used together in one device they did not always work in sync "together". This is because when the sensors were worn on the non-dominant side and a seizure occurs only the heart-rate change was indicated: the accelerometer showed no change. Yet when in correct position on the body they work in union as an excellent detection method. Therefore, demonstrating that body placement or position is paramount. For example, one patient's dominant side was the right arm. This means seizures are known to occur on the right. The results from "Observation c HP4" (Fig.2), can be seen. During shaking from the right wrist at the recorded time: 00.23am during a GTCS seizure all 3 measurements on

axis X, Y and Z showed sudden movements and the heart-rate increased to its highest peak at 128. Before the seizure the heart-rate was much lower at 80, then rising rapidly to 90 and up to 128. This suggests both the sensors detected the seizure correctly.

Yet the results from "Observation a HP4", seen below in Fig. 3, Observation Set 2 indicate that during a GTCS at the recorded time '12.44pm' the 3 measurements on axis X, Y and Z did not show any sudden movement, in fact barely any movement at all, yet the heart-rate increased to its highest peak at 124, in keeping with typical heart-rate increase measurement during a GTCS for HP4. Since the accelerometer was positioned on the left wrist this reveals the sensor did not detect movement therefore demonstrating the sensor was positioned in the wrong position.

Knowing the individual characteristics of the patient profile prior to sensor-based device recommendation is key, for example the HP1 with FAS (Focal Aware Seizures) and FSIA.

(Focal Seizures with Impaired Awareness): the question here was "did the 2 sensors work in union to detect the Focal seizures?" Some heart-rate increase was detected but the accelerometer was primarily redundant, for example in 'Observation 10' (Fig.3.) a seizure occurs with sensor positioned on right wrist at the observed time: 09.44am, in this observation the heart-rate sensor detects some change over a 2 minute period i.e. The heart-rate begins at 87 increases to 90 then back to 87 then declines to 86 then steadily back to 90. At 09.45am the heart-rate does show increase to 95 and goes back down to 85. Heart-rate range is 87 -95, with some sudden movement from accelerometer at time of increased heart-rate.

Yet, during 'Observation 11', seen in Fig. 3 for HP1 with the seizure observed at 20.36pm the heart-rate range is 81-84 with little sudden movement. Likewise, in 'Observation 12' (Fig. 4, Observations set 3) seen above: the seizure occurrence at 21.06pm demonstrates the heart-rate range: 81-86 and little sudden movement.

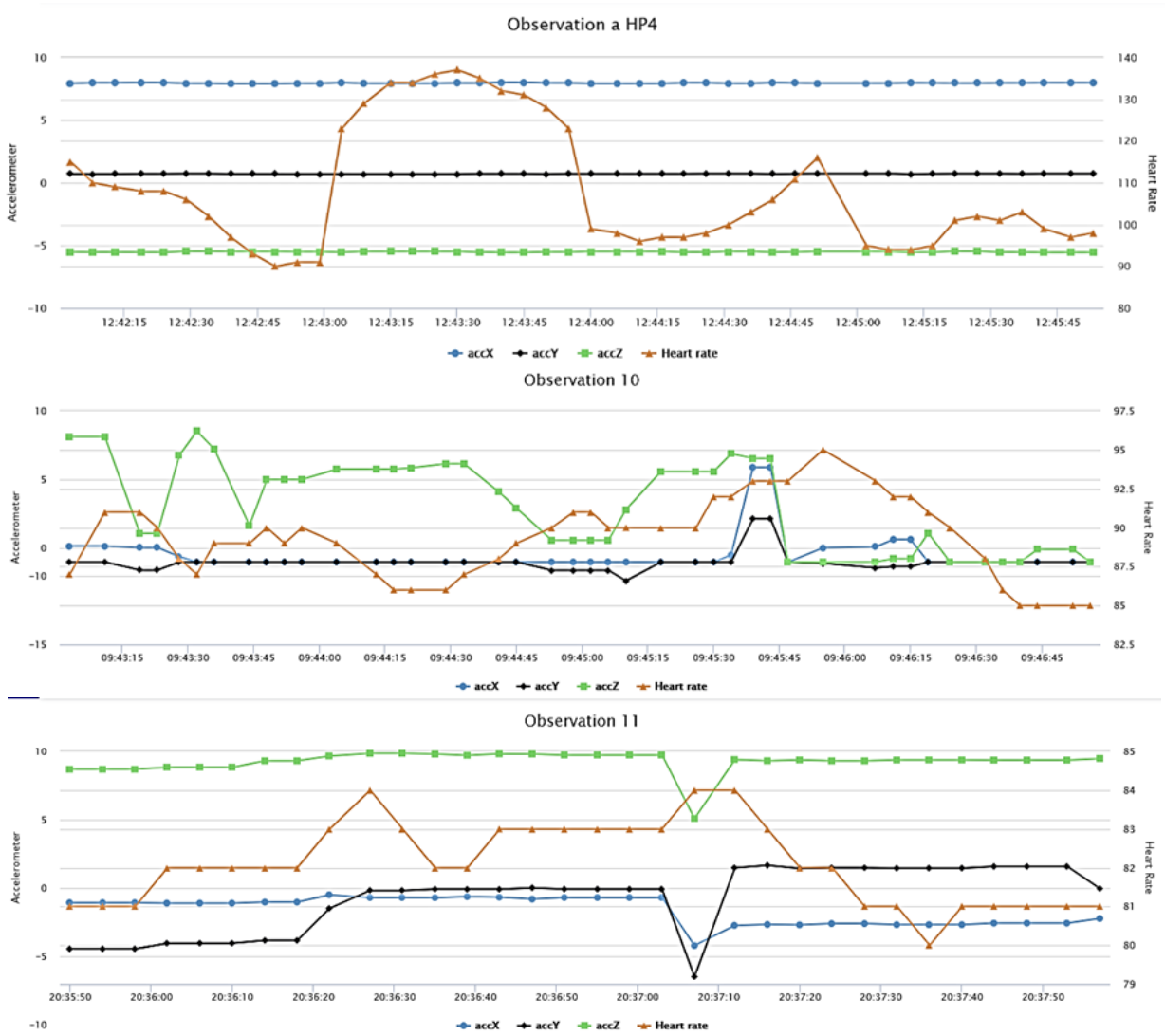


Fig. 3. Observations, Set 2.

A pattern emerges for HP1 in other observations whereby the heart-rate decreases, for example in 'Observation a HP1' (in Fig. 4.) above at the time of the seizure '09.55am' the heart-rate decreases from 90 to 84, and likewise in 'Observation f HP1' (Fig.4) with heart-rate decrease from 104 to 79 during the observed time of seizure at 11.01am and in 'Observation g HP1' (Fig.4) with heart-rate decrease from 100 to 84 during the observed time of seizure at 20.44pm.

In 'Observation g HP1' the accelerometer indicates movement from all 3 X, Y and Z axis on the accelerometer at the time of the seizure. This is further observed in 'Observation e HP1' (Fig. 5, Observations, set 4) whereby there is sudden change in the accelerometer, but this is '21.01pm' 'after' the time of seizure at '21.00' where in fact again the heart-rate shows decrease.

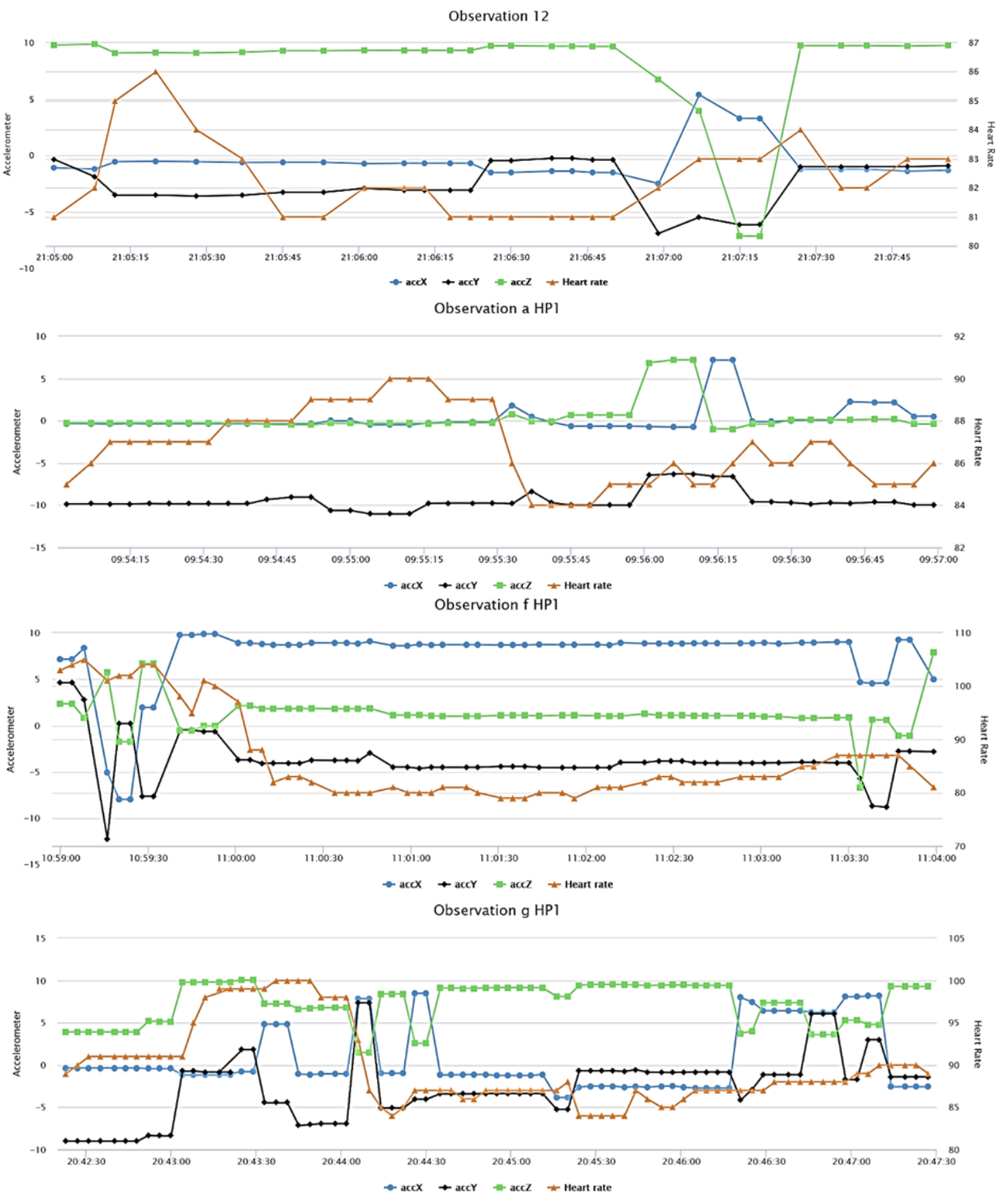


Fig. 4. Observations, Set 3.

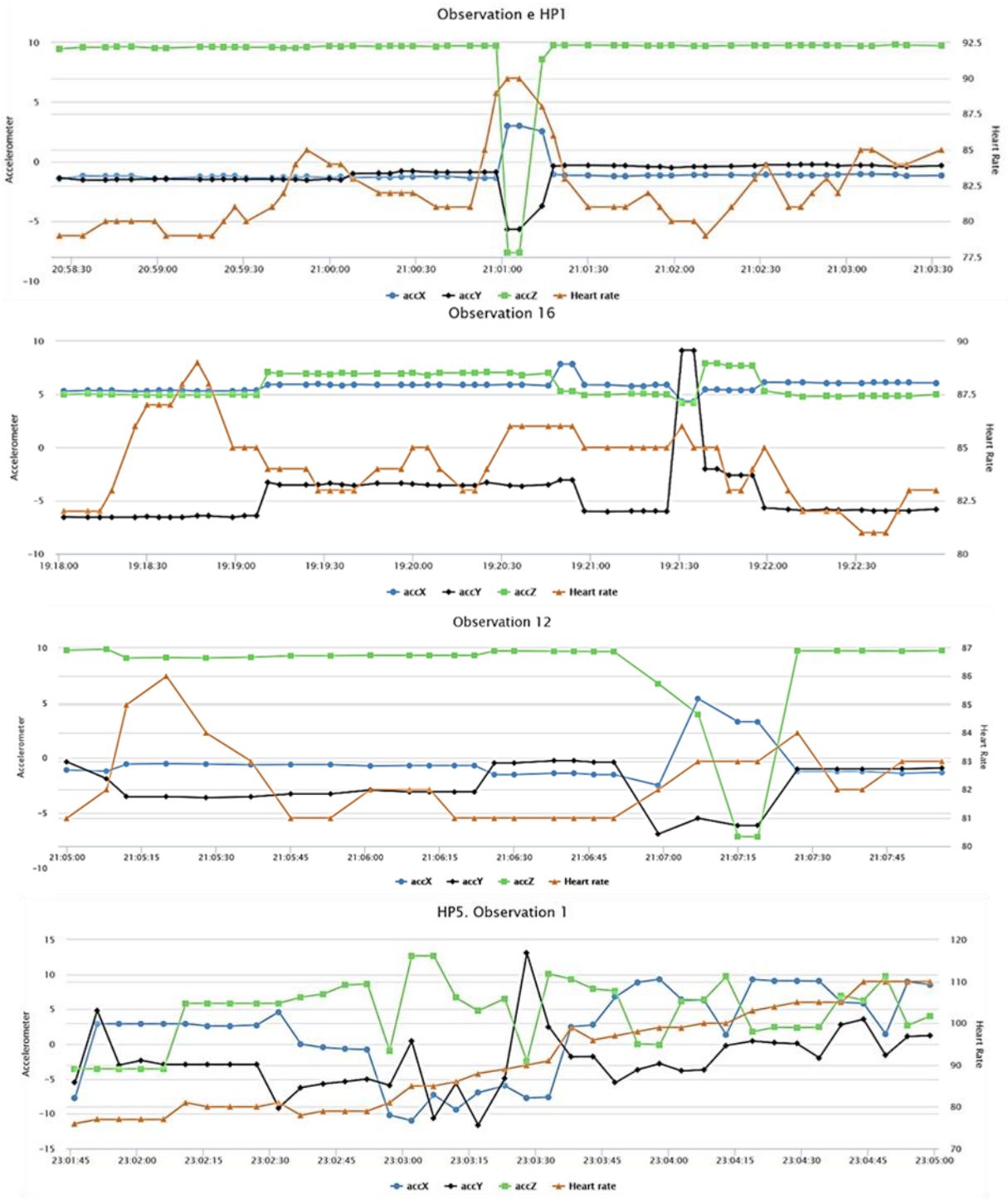


Fig. 5. Observations, Set 4.

Only 1/14 seizures observed for HP1 demonstrate significant movement for the accelerometer during seizures. Therefore, the accelerometer sensor is not useful for detecting these non-shaking seizures, although some patients' profiles (found earlier in the collection of the 'anonymous patient profiles') reveal absence seizures can evolve to convulsive generalized tonic clonic movements. Therefore, it is endorsed that Focal seizures alone, go undetected without HR changes [31].

Predominantly during non-seizure time periods for HP1 there is no significant variance in accelerometer measurements when compared to 'seizure occurrence' time periods. For example, "Observation 16" (Fig. 5) shows a 4-minute snap-shot when HP1 has no seizure: the 3 axis X, Y and Z move similarly in "Observation 12" (Fig. 5).

Ultimately there is some evidence demonstrated here that heart-rate 'change' occurs during the seizures for HP1

(increase and decrease). Yet since the seizures for HP1 occur for such a short time (20 seconds) heart-rate fluctuations can be missed or miss-interpreted as 'false alarms' perhaps due to agitation before or after the seizure.

The key and common signs and symptoms are 'LOC' (Loss of Consciousness) and 'Automatisms' for a patient such as HP1. Automatisms reveal themselves in a multitude of forms including repetitive movements, such as, lip smacking, chewing or swallowing, picking at clothes or skin or even staring [32] and these are difficult to detect with any sensor.

Hence, these repetitive movements deemed as other behavioural components of seizures which include non-motor components and post-ictal phenomena cannot be detected by the accelerometer [33] in line with the theory that seizures that are typical to the dominant body area not wearing the sensor-based device will not be detected.

Some patients exhibit automatisms such as sudden sweating events [34] and since sweating is associated with Focal Seizures a more appropriate sensor for a patient with FAS and FSIA seizure types would be Galvanic Skin Response Sensor (GSR), which refers to changes in sweat gland activity [35] as evidenced in other studies performed to detect seizures this galvanic skin response (GSR) sensor has been used in multi-modal platforms [34].

Empaticas' 'Embrace Plus' smart watch [35] can be useful for Focal seizures as it has other sensors in addition to Galvanic Skin Response (GSR), for example EDA sensor and peripheral temperature, are just one of many sensors available in this device for researchers [36]. The Electro Dermal Activity (EDA) represents the electrical changes on the surface of the skin (not just for sweat). Although witnessed in some studies finding that EDA increases during GTCS were greater than during CPS (Complex Partial Seizure, now: focal seizure with loss of consciousness) nonetheless it is a useful sensor for this type of seizure [36].

The peripheral temperature sensor also has evidence for use in detection in non-convulsive seizure's (CPS) (focal) [36].

Although the Fitbit Ionic used in this study is not a conventional device for monitoring epilepsy, it can be adapted to detect seizures as demonstrated. It is a less expensive everyday 'patient friendly' option as opposed to EEG monitoring whereby the patient wears electrodes that are not comfortable: this is because the EEG-electrodes must be attached to the scalp which hampers the patient's movement making long-term home monitoring not feasible.

In this sense this less expensive, comfortable alternative to EEG monitoring can be especially useful for patients with non-epileptic events. During early analysis in this experiment many patient profiles (found in the collected 'anonymous patient profiles') where identified as having non-epileptic events and were categorised under "Non-Classified". Although no confirmed 'epilepsy' these patients are still suffering with seizure signs and symptoms: as observed below in 'HP5. Observation 1' (Fig. 5). This patient has a non-epileptic shaking event at 23:03pm, the observation indicates heart-rate increase from 80 up to 109, the accelerometer also indicates rapid activity at the time of the shaking event.

One of the challenges in using sensor-based IoT devices to achieve a personalised approach is the barriers found in the use of them in hospital settings. Although EEG monitoring is the chief gold standard method used within hospitals to detect and monitor seizures, there is limited evidence found how sensor-based IoT devices and experiments are used in hospital settings.

There are very few experiments with sensor-based IoT devices that have been endorsed by the hospitals and a large problem is poor information when caring for people with epilepsy or doing epilepsy clinical trials [37] yet there is great potential to vastly increase the efficacy of epilepsy management using biomedical devices that can improve the quality of information. As available devices and sensors grow, if clinicians could be provided with more guidance in understanding and choosing which sensor suits which situation then a personalised approach can be achieved.

D. Calibration

Prior to this experiment upon hospital patients, this study was conducted with 2 non-epileptic volunteers who were asked to undergo the testing and perform 'simulated seizures' in a controlled environment with the sensor-based device positioned on different parts of the body at different times of the day. This was to calibrate the main hospital patient experiments. The findings are discussed below, with evidence of some of the Volunteer Observations together with the simulated individual profile characteristics and criteria used for observations.

As can be seen from 'V1.Observation 1' in Fig.5. Volunteer Observations, set 1, below, the first volunteer, with seizures occurring on the dominant right side of the body simulated a GTCS shaking from the right arm at the recorded time: 20.38pm. The sensor-based device was worn on the right wrist. The 3 measurements on axis X, Y and Z showed sudden movement and the heart-rate increased to its highest peak at 100. Before the seizure the heart-rate was steadier at 78-81, then after the seizure the heart-rate decreased to 80. This suggested the sensors detected the simulated seizure correctly.

When V1 simulated a seizure again from the right arm they placed the sensor-based device on the left wrist at the event time: 10.10am. As identified in 'V1. Observation 3' (Fig. 6) the heart-rate shows an increase (78 to 123) yet the accelerometer axis is smooth. Similarly, in 'V1. Observation 4' (Fig.6.) the volunteer placed the device on the 'non-dominant' left leg and again the heart-rate increased dramatically but the acceleration generally smooth, although some movement on all 3 axis at the time: 09.47am of the event. Since 'some movement' was detected here a further test was performed with the device on the dominant right leg and here the difference is apparent, seen in V1. Observation 5 in Fig.7. This confirms the theory identified in earlier hospital patient observations that the sensor-based device position is paramount.

For V2 the dominant side is left. It is evident when V2 placed the sensor-based device on the left leg during a simulated seizure at 22.14pm both the heart-rate and accelerometer sensors are reacting vigorously, seen in 'V2.Observation 1' in Fig. 7.

Similar to V1 when V2 places the sensor-based device on the 'non-dominant' side (right leg) during a simulated seizure only the heart-rate sensor reacts, seen below in 'V2.Observation 1a' (Fig. 7) again fueling the theory that 'position' of the sensor-based device is paramount.

E. Known Characteristics

Detection of seizures using an everyday sensor-based device and data transfer to online database was successful. This presented evidence that remote monitoring of specific epilepsy patients' profiles with known characteristics can be improved. The comfortable sensor-based device with heart-rate and accelerometer provided accurate data and is a more dependable method than a patient's paper diary.

Difference was observed due to 'position' on the body of the sensor-based device, demonstrating that because of the known patient specific characteristics a personalised approach is achieved. Furthermore, it was discovered that the 'type' of sensor used is principal in its correspondence with a patients' particular 'seizure type' together with the particular associated signs or symptoms.

The sensors and techniques used in this experiment enable some assurance in long term remote monitoring. The use of such sensor-based device used in this experiment can reduce the frequency of visits to hospitals and improve daily management of epilepsy thus, these sensing techniques have shown that results can be achieved in the measurement of specific epileptic seizures based on observations.

As established through these experiments' timely detection along with known patient characteristics is one of the keys to monitoring epilepsy.

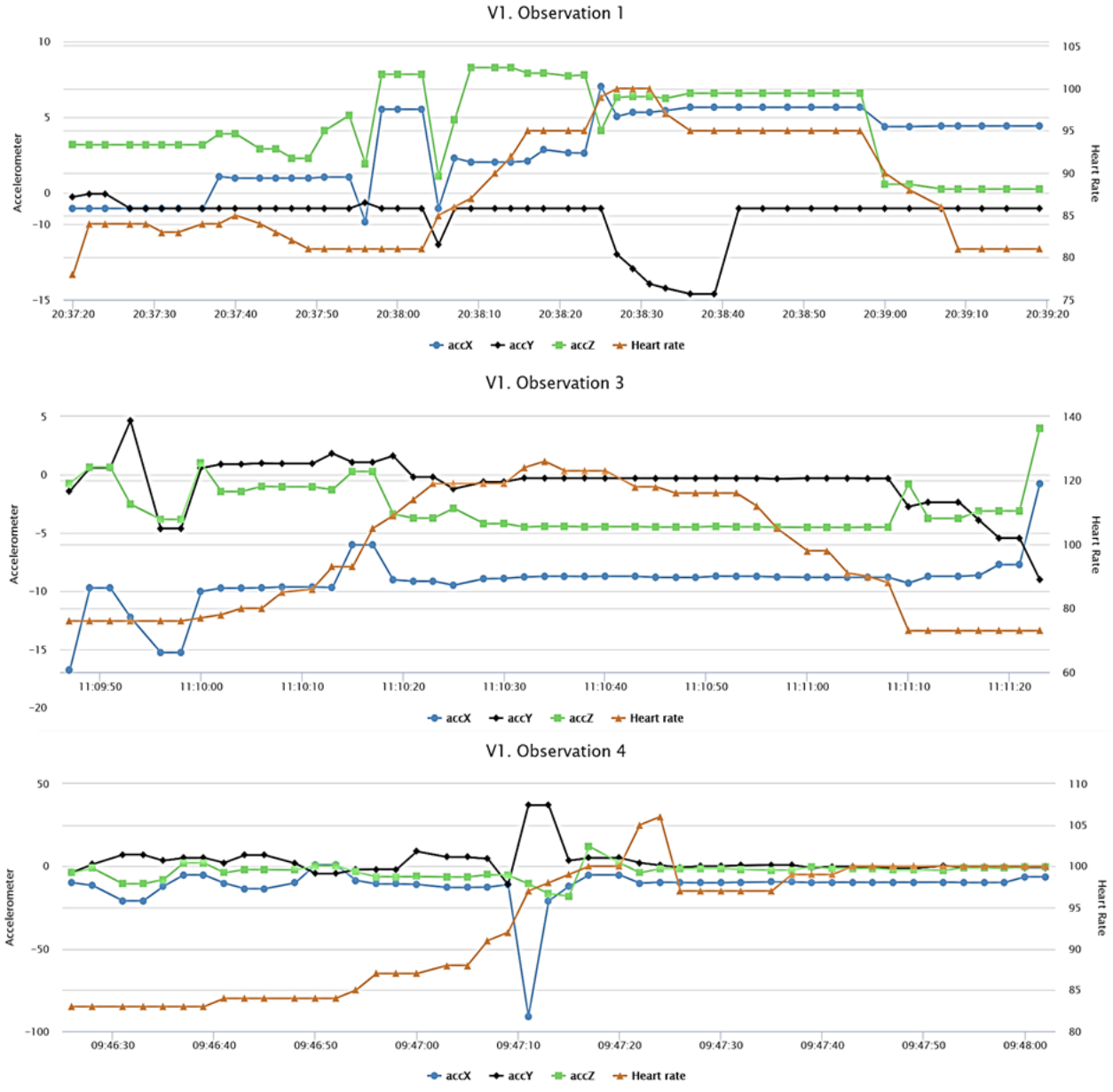


Fig. 6. Volunteer Observations, Set 1.

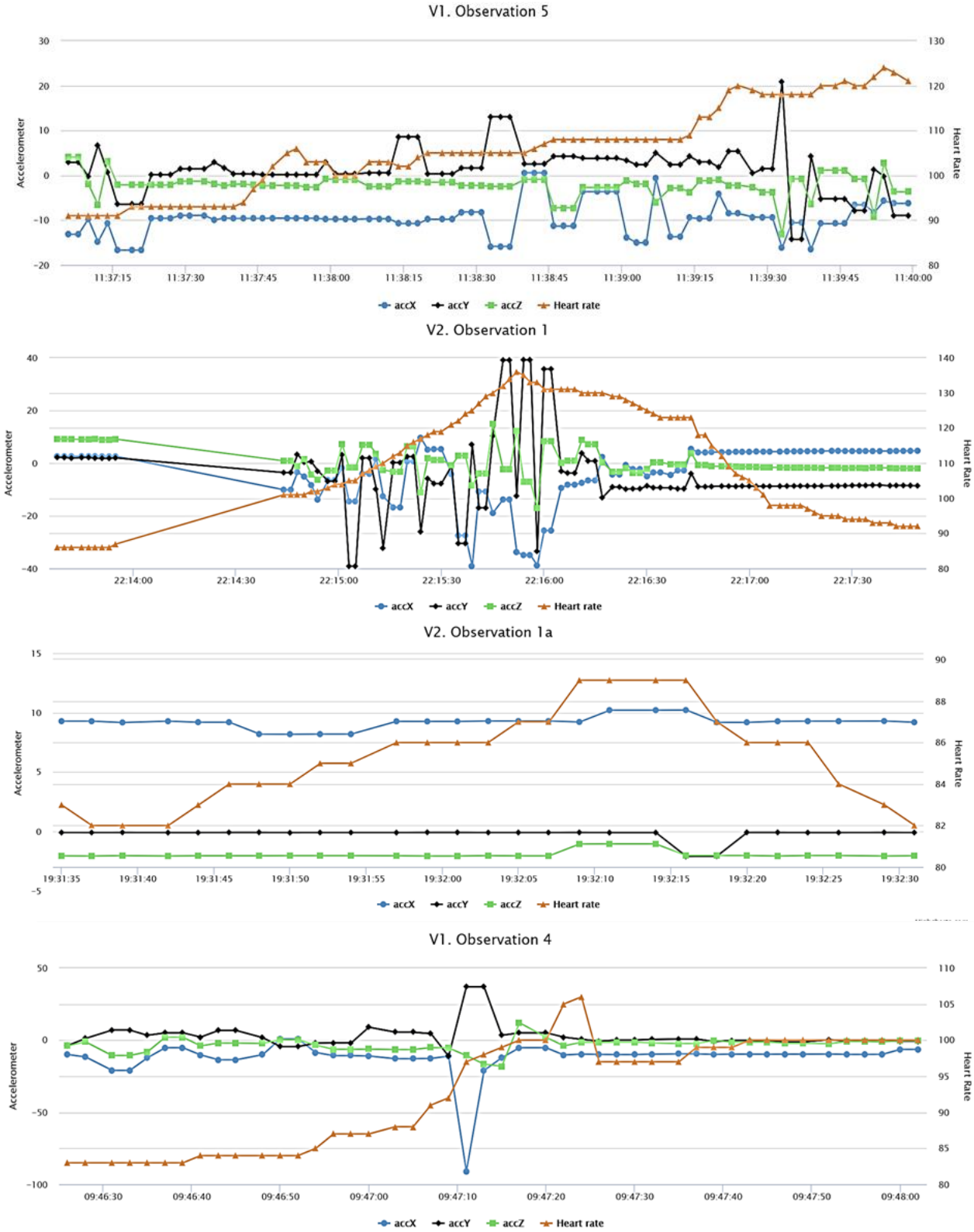


Fig. 7. Volunteer Observations, Set 2.

V. IoT Based Epilepsy Monitoring Model

The purpose of the IoT based Epilepsy monitoring model [38] is to support a 'Personalised Monitoring Plan' framework in collecting data from a variety of potential epilepsy device sensors and also provide optimal analysis tools to utilise the sensor data thus supporting clinicians to monitor epilepsy patients.

A. PMP Framework

This section proposes a Personalised Monitoring Plan (PMP) framework. In the previous section experiments were performed to capture seizure data, obtained from sensors, which are positioned on different parts of the patient's body. The results from this experiment are used to inform a PMP (Personalised Monitoring Plan), seen below in Fig. 8 which recommends which sensor-based device to use based on those very individual, personal characteristics of a given patient.

The proposed 'Personalised Monitoring Plan' (PMP) framework is a model for which doctors and healthcare professionals (HCPs) can use to assist in identifying which device they should recommend to the individual patient for remote monitoring.

The PMP framework integrates two types of 'personalisation:

- The patient as the individual (derived from an ontology language).
- Use Patients in a category (using the K-means Clustering method).

Both these personalisation elements are described below in in the next sections. The third tool of the PMP framework supports the decisions surrounding recommending the correct IoT sensor-based devices. The main purpose is to help HCPs decide which IoT Sensors to recommend for monitoring and which position on the patient's body.

The PMP framework ultimately allows users to provide a description of the 'seizure condition' of a single patient or a patient type, and to automatically obtain a PMP adjusted to the patient requirements.

The proposed framework consists of two features: the first being 'Personalisation' (based on this study) and the second is the anticipated 'Remote Monitoring', shown in pink and blue respectively in Fig. 8, PMP Framework.

B. Personalisation Elements

The personalisation contributions are part of the preceding findings in this study. The ontology language was created to support the need of the healthcare process to transmit, re-use and share individual patient profile data related to their seizures.

The ontology was achieved by the initial examination of 100 anonymous epilepsy patient medical records. The data was analysed to discover if values for each of the attributes are different for each patient, together with the investigation of epilepsy 'terminology' and existing seizure type classifications/categories were analysed so that an 'individual' seizure type patient profile could be formed. A close collaboration with clinicians helped to build a data model fit for real-world adoption inside hospital settings and thus an ontology was developed to model the concept of the epilepsy patient profile, namely ESO 'Epilepsy Seizure Ontology'. This was a driving force for the PMP Framework and a critical aspect for this concept. In order to make ESO useable for HCPs (Health Care Professionals) the ontology was transformed into a language that is understandable by humans and machines, this was accomplished by XML and the outcome was PPDL (Patient Profile Description Language).

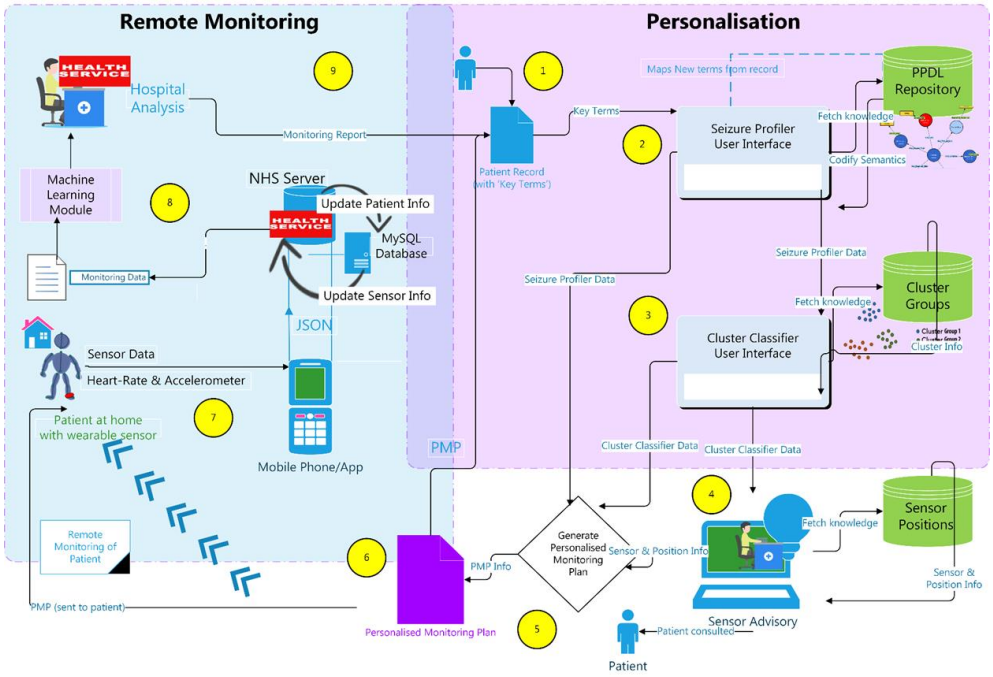


Fig. 8. PMP Framework.

The second personalisation element was achieved by using K-means Clustering analysis. Different clustering techniques were initially analysed to find the most appropriate approach for the acquired epilepsy data and an in-depth focus upon 'clustering considerations' was undertaken to confirm validity.

The outcome was a set of 6 distinct 'clustering' groups, shown in Table I. These 6 cluster groups revealed six completely different categories of patients each with their distinct seizure related information.

The results revealed the distinct groups of epilepsy patients that share similar characteristics using Clustering Analysis. This will enable the health carers to define a 'type' of epilepsy patient.

C. PMP Framework Loop and Maintenance

With this PMP framework, the PPDL (Patient Profile Description Language) can be directly maintained (and extended) by HCP's. The framework has the flexibility, (as the ontology grows with new seizure related concepts), to deal with the mounting diversity of seizure type patients. Therefore, the PMP framework is somewhat reliant on the integration of new knowledge in the PPDL. As the HCP of the PMP Framework approves the recommendations, the information about the patient and the advice may change and this new information can cause the PMP Framework to continue providing new suggestions to the HCP.

This loop will stop either when the framework is not able to provide new recommendations or when the HCP considers that the current condition of the patient is correctly represented by the recommendation. At any time, the PMP for the patient is fundamentally controlled by the HCP who is using the framework.

Consequently both the personalised 'seizure related data' and the 'cluster classifier data' of a patient may evolve as the patient disorder changes, for example when the information about the patient changes in the patient record of that patient or as a result of the application of the PMP Framework to find out new 'seizure type' knowledge about the current patient. The datasets are expected to evolve and are continuously stored as part of the record of that patient.

D. IoT based Epilepsy Monitoring Model

To achieve the type of monitoring described in the PMP framework, several IoT components can be deployed to retrieve sensor data from the epilepsy patient to be accessed remotely. These components include the integration of the personalisation components described in the PMP framework, those of an internet connection and protocols which form the 'network layer', a cloud platform to manage the data analysis and fundamentally the sensor-based devices forming the sensor layer. These components make the ingredients of an IoT solution, proposed in the IoT based Epilepsy monitoring model shown in Fig. 9.

The sensor layer, (discussed in section F) has the task of acquiring and sending the data from the different epilepsy devices involved in capturing seizure data, to the proposed cloud platform.

The 'IoT based Epilepsy monitoring model' proposal in Fig.9 shows areas on the body where parameters are measured, each area is indicated with a colour matching the parameter.

E. Cloud Platform

The proposed cloud platform provides all the necessary services for the clinician to manage, process and visualise the seizure data. All the processes that involve the interaction between the personalisation layer and the sensor layer are carried out through the following modules: PMP data management, machine learning module and data analysis & visualisation. All these services are hosted in the cloud and clinicians are able to access them remotely from any location.

The data analysis and visualisation module utilises the sensor data while the 'PMP data management' module pulls all the patient records from the personalisation modules and here the sensor data results are updated. Visualisation is a requirement for any such system as it is important for clinicians to be provided with user friendly GUIs so they can study the seizure data from the epilepsy sensor devices. The machine learning module is also proposed, this is a key aspect for future development and the idea is that by using algorithms the module will 'learn' when a patient is about to have seizure and warn them in advance.

TABLE I. CLUSTER GROUPS

Attribute	Cluster					
	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Seizure Type	NMA	Un-classified	FAS	Gelastic	GTCS	
Key Sign / Symptoms	LOC	LOC	Sensations	None	Myoclonus Bilateral Clonus	Bilateral Clonus
	Urinary / In-continence	None			LOC	LOC
Common Sign / Symptoms	Automatism	Cognitive Automatism	Automatism	Automatism	Sensory Cognitive	Automatism
Arm / Leg	Either	Leg	Either	Either	Leg	Either
Nocturnal / Diurnal	Nocturnal and Diurnal	Nocturnal and Diurnal	Diurnal	Diurnal	Nocturnal and Diurnal	Diurnal

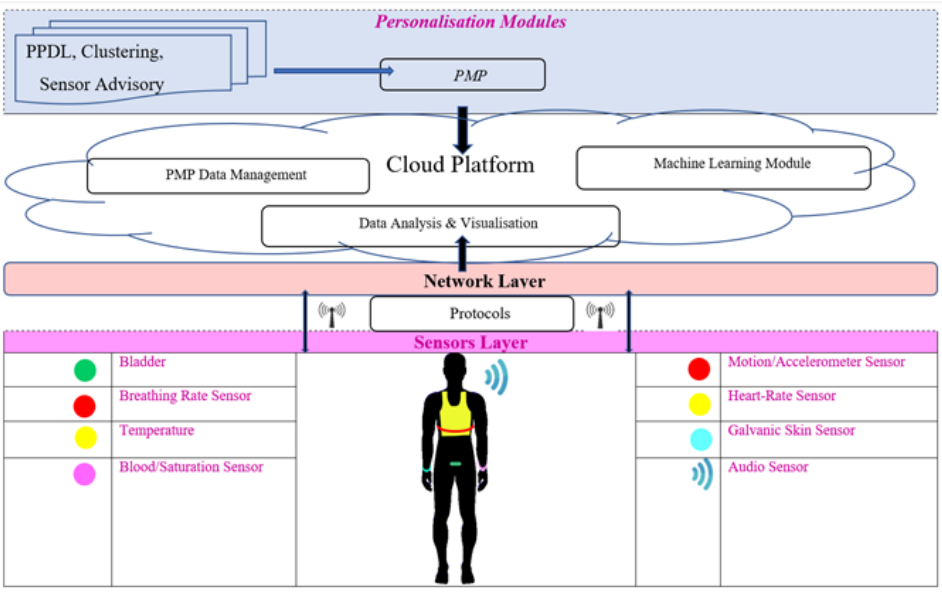


Fig. 9. IoT based Epilepsy Monitoring Model.

A pre-processing hardware and a platform are needed to communicate and transmit the sensor data which is collected using wearable sensors positioned on a patient's body. The Microsoft Azure IoT platform [39] is proposed, since this cloud computing server is trusted and safe [40].

F. Sensor Layer

The sensor layer has the task of acquiring and sending the data from the different epilepsy devices involved in capturing seizure data, to the proposed cloud platform.

The sensors previously used in the experiment, the heart-rate and accelerometer sensors were demonstrated within the PMP framework discussed in this paper. Yet there are other sensors too that can work within the context of this research. These potential sensors found in other devices are explored and proposed below.

Despite the expense, multi-modal sensor based devices are the ultimate desire to monitor an epilepsy patients seizures since multiple sensors are embedded in one device and make comfortability for the patient and all-in-one solutions for the manager of the device, furthermore epilepsy patients have revealed their preference for devices capable of monitoring several parameters [24]. Section III in this paper uncovered some of the epilepsy detection devices and monitoring systems, whilst section IV demonstrated the use of an inexpensive device with heart-rate and accelerometer sensor (justified for experimental purposes), and simultaneously identified other useful sensors for epilepsy. These other sensors which exist in devices, those that go beyond heart and movement sensing, and beyond fitness devices, are amongst a vast amount. Consequently, many studies have analysed the performance and limitations of each sensor based device, one critical evaluation by Peake [41] found many devices where not yet fully validated or tested for reliability, therefore this examination will aim only to propose validated devices for use in the IoT based Epilepsy monitoring model.

Recently in 2020, Abreu, Fred et al [42] did a significant exploration on, wearables and related devices, that can be utilised for epilepsy prediction, the findings presents devices, some with multiple sensors, characterised with respect to their applicability to research, validation status, form factor or body positioning, battery duration, method to access the data, measured signals and, their applicability to epilepsy prediction (EP) [42]. This is a vital study since the devices have already been validated, and connectivity options identified, and since they are beneficial for epilepsy they can be proposed in the IoT based Epilepsy monitoring model.

These devices and their sensors, identified as applicable for epilepsy prediction contain some of the sensors that were highlighted during the experiments i.e. accelerometer, heart-rate and GSR, but the audio and bladder sensors were not previously included. In this case the audio and bladder are added to the IoT based Epilepsy monitoring model in Fig.9 above and the connectivity options and battery life can be referred to in Table II.

The IoT based Epilepsy monitoring model proposal in Fig.9 shows areas on the body where parameters are measured, each area is indicated with a colour matching the parameter. The chosen devices have been proposed based upon the factors in the study by Abreu, Fred et al [42] for the best battery life, validity, and connectivity options selected for ease of connection to the cloud platform in the IoT based Epilepsy monitoring model. Since Embrace2 device uses its own onboard processing it is not adaptable for the model proposed in this study. Furthermore some of the device based sensors depicted in

Table II. 'Selected Sensors' are multi-modal so their use can sometimes be proposed in more than one activity area, for example the device EQ02 has both the heart-rate and temperature sensor.

TABLE II. SELECTED SENSORS

Body Area	Validated Device	Sensor	Connectivity Option	Battery Life
Breathing	HexoSkin	RESP	Cloud Storage or BLE transmission	12h
Electrodermal	Empatica E4	EDA or GVS	Cloud Storage	48h
Blood Volume	Empatica E4	PPG	Cloud Storage	48h
Movement	Bioharness3	ACC	BLE transmission	12-24h
Heart-Rate	EQ02	ECG	Bluetooth transmission	48h
Temperature	EQ02	Temp	Bluetooth transmission	48h
Audio	Alert-it	Sound Sensor	Ethernet connection	12-16h
Bladder/ Incontinence	DFree	Ultra- sound	Bluetooth transmission	24h

G. Network Layer

There are several ways the sensors can connect and send data to the cloud platform and since most of the sensor devices connect to a mobile phone they are served by Bluetooth or Bluetooth Low Energy (BLE) and use very little power. Nevertheless, each sensor-based device is provided with its own protocol and connectivity options, hence the type of IoT connectivity is determined generally by the distance that the data must travel, either short-range or long-range [43]. IoT platforms such as Azure use gateways to connect IoT devices to the cloud. The data collected from the devices moves through this gateway, gets pre-processed using in build modules (Edge) and then gets sent to the cloud. Data is protected by an additional layer of security provided by the Azure Application gateway and in addition connection security is enabled as each connected IoT device is given a unique identity key [39].

VI. EVALUATION

An evaluation was performed by taking two different epilepsy patients through the steps in the PMP framework. Two different 'use case' scenarios each with different patient profiles were tested by revealing their respective inputs and outputs. The aim was to provide the Personalised Monitoring Plan (PMP) described in this study and only the 'seizure related' information of the patient was considered. This input data is primarily composed of the patients' seizure types, signs and symptoms. Therefore, given the condition of an epilepsy patient, the PMP is used to personalise the medical knowledge available for that patient, all other unrelated medical knowledge to the patient is discarded. The input data was processed using the framework features: in summary the framework produced new datasets that were passed to the 'Generate PMP' component, these were: seizure profiler data, cluster classifier data and sensor position information. The evaluation results helped determine the effectiveness of the PMP framework and how it can be used as a tool for recommending the IoT device to an epilepsy individual patient.

VII. LONG TERM USES AND APPLICABILITY IN OTHER DOMAINS

The methods used in this study for ontology development and clustering analysis can be applied to any disease whereby recognised symptoms per patient can be individualised and be further put into sub-groups or categories. However, to fully utilise this personalised approach the application of the PMP framework can be particularly applied to patients whom have symptoms that can be monitored with different IoT sensor-based devices and personalised further by wearing the device on different body positions. In the future the following types of patients can be handled by the proposed PMP framework: (shown below together with latest progressive recommended sensor-based devices).

- Diabetes: i.e. One such recommended device could be use of 'flash glucose sensing': A device which checks blood glucose levels by scanning a sensor worn on their arm will be (available on the NHS for people with type 1 diabetes) [44].
- Sick Infants: i.e. A recommended device could be use of a miniaturised, wireless oxygen sensor wearable device the size of a Band-Aid which would allow babies to be monitored from home and able to leave the hospital [45].
- Rehabilitation: i.e. the recommended device for rehabilitation could be a Force-based sensor which can be integrated with footwear to measure the interaction of the body with the ground during walking [36]. Due to the possibility of detecting not only physiological but also movement data wearable sensors have also acquired increasing importance in the field of rehabilitation [46].

A. Machine Learning

Another significant future direction for long-term remote monitoring of epilepsy is seizure detection via 'machine learning'. By accumulating large datasets, computers can learn by recognising patterns in data.

This automated approach (without human intervention) has been proposed as a 'machine learning module' within the PMP framework and IoT based Epilepsy monitoring model to determine seizure detection or not based on the patient specific profile, the idea being that by using algorithms the module will 'learn' when a patient is about to have seizure and warn them in advance. Due to limitations in this study this module has not been built but instead is proposed at the next stage and recommended for the next level. Largely there is further work to take the PMP conceptual model and the IoT based Epilepsy monitoring model into full operation.

Recent advances in machine learning and deep learning technique inventions have shown noteworthy advantage in the

automatic learning of robust features that outperformed the human oriented features in many domains such as self-driving cars, natural language processing, and computer vision also medical diagnosis [47] [48].

Yet as identified in a review of epileptic seizure detection using machine learning classifiers a major challenge is to detect seizures correctly from a large volume of data [49], and it is highlighted that the selection of suitable classifiers and features are crucial [49].

Ultimately, along with the challenges associated with the increasing dataset sizes (hence growing epilepsy cases), and evolving date science hitches, as well as obtaining sensitive data it could be argued that the greatest challenge of all to help solve these problems is the enabling collaboration between people with differences in expertise [50].

VIII. Conclusion

The principal contribution in this study was that with the prior 'knowledge' of individual patient characteristics drawn from the PPDL repository and 'Cluster Groups' together with the supplementary 'proof of concept' knowledge obtained in the experiments each epilepsy patient can be treated distinctly and recommended an appropriate sensor-based device thus forming a patient specific unique PMP (Personalised Monitoring Plan). Hence personalisation can be achieved.

The sensors and techniques used in the experiment enables some assurance in long term remote monitoring. The use of such sensor-based device used in the experiment can reduce the frequency of visits to hospitals and improve daily management of epilepsy thus, these sensing techniques have shown that results can be achieved in the measurement of specific epileptic seizures based on observations.

As established through these experiments' timely detection along with known patient characteristics is one of the keys to monitoring epilepsy.

The integration of the components and technologies in the framework depicted in Fig.8 PMP Framework aims at providing HCP's dealing with epilepsy patients with an integrated tool that helps them in recommending the correct IoT sensor and position on the patient's body.

These decisions are made at the initial consultation and act as an 'aid' in personalising the condition of new incoming patients, and thus refine the predefined 'patient record' in order to obtain and validate a 'Personalised Monitoring Plan' which is in addition adapted to include the seizure monitoring of the patient during appointments.

The PMP Framework is designed to provide a patient-empowering support in a way that the available knowledge is continuously personalised to the condition of the seizure type patient. The IoT based Epilepsy monitoring model has been proposed and can be adopted by the PMP framework in future developments.

References

[1] M. Pediaditis, M. Tsiknakis, V. Kritsotakis, M. Goralczyk, S. Voutoufianakis, and P. Vorgia, "Exploiting advanced video analysis technologies for a smart home monitoring platform for epileptic patients: Technological and legal preconditions," presented at the 2012 International Conference on Telecommunications and Multimedia (TEMU), 2012/07, 2012. [Online]. Available: http://dx.doi.org/10.1109/temu.2012.6294719.

[2] N. Moghim and D. W. Corne, "Predicting epileptic seizures in advance," (in eng), PLoS One, vol. 9, no. 6, pp. e99334-e99334, 2014, doi: 10.1371/journal.pone.0099334.

[3] B. S. Chang and D. H. Lowenstein, "Epilepsy," New England Journal of Medicine, vol. 349, no. 13, pp. 1257-1266, 2003/09/25 2003, doi: 10.1056/nejmra022308.

[4] D. Hirtz, D. J. Thurman, K. Gwinn-Hardy, M. Mohamed, A. R. Chaudhuri, and R. Zalutsky, "How common are the "common" neurologic disorders?," Neurology, vol. 68, no. 5, pp. 326-337, 2007/01/29 2007, doi: 10.1212/01.wnl.0000252807.38124.a3.

[5] L. Chen et al., "OMDP: An ontology-based model for diagnosis and treatment of diabetes patients in remote healthcare systems," International Journal of Distributed Sensor Networks, vol. 15, no. 5, p. 155014771984711, 2019/05 2019, doi: 10.1177/1550147719847112.

[6] A. F. Van Straten and B. C. Jobst, "Future of epilepsy treatment: integration of devices," Future Neurology, vol. 9, no. 6, pp. 587-596, 2014/11 2014, doi: 10.2217/fnl.14.54.

[7] M. Tentori, L. Escobedo, and G. Balderas, "A Smart Environment for Children with Autism," IEEE Pervasive Computing, vol. 14, no. 2, pp. 42-50, 2015/04 2015, doi: 10.1109/mprv.2015.22.

[8] T. Tamura and W. Chen, Seamless healthcare monitoring. Springer, 2018.

[9] P. Bonato, "Wearable Sensors and Systems," IEEE Engineering in Medicine and Biology Magazine, vol. 29, no. 3, pp. 25-36, 2010/05 2010, doi: 10.1109/memb.2010.936554.

[10] E. Magiorkinis, A. Diamantis, K. Sidiropoulou, and C. Panteliadis, "Highights in the history of epilepsy: the last 200 years," (in eng), Epilepsy Res Treat, vol. 2014, pp. 582039-582039, 2014, doi: 10.1155/2014/582039.

[11] D. J. Cook, M. Schmitter-Edgecombe, and P. Dawadi, "Analyzing Activity Behavior and Movement in a Naturalistic Environment Using Smart Home Techniques," (in eng), IEEE J Biomed Health Inform, vol. 19, no. 6, pp. 1882-1892, 2015, doi: 10.1109/JBHI.2015.2461659.

[12] T. D. Parsons and R. L. Kane, "Computational Neuropsychology," in The Role of Technology in Clinical Neuropsychology, ed: Oxford University Press, 2017.

[13] V. J. Aski, S. S. Sonawane, and U. Soni, "IoT Enabled Ubiquitous Healthcare Data Acquisition and Monitoring System for Personal and Medical Usage Powered by Cloud Application: An Architectural Overview," in Advances in Intelligent Systems and Computing, ed: Springer Singapore, 2018, pp. 1-15.

[14] B. Direito, C. Teixeira, B. Ribeiro, M. Castelo-Branco, F. Sales, and A. Dourado, "Modeling epileptic brain states using EEG spectral analysis and topographic mapping," Journal of Neuroscience Methods, vol. 210, no. 2, pp. 220-229, 2012/09 2012, doi: 10.1016/j.jneumeth.2012.07.006.

[15] S. Xie and S. Krishnan, "Wavelet-based sparse functional linear model with applications to EEGs seizure detection and epilepsy diagnosis," Medical & Biological Engineering & Computing, vol. 51, no. 1-2, pp. 49-60, 2012/10/09 2012, doi: 10.1007/s11517-012-0967-8.

[16] A. Ulate-Campos, F. Coughlin, M. Gaínza-Lein, I. S. Fernández, P. L. Pearl, and T. Loddenkemper, "Automated seizure detection systems and their effectiveness for each type of seizure," Seizure, vol. 40, pp. 88-101, 2016/08 2016, doi: 10.1016/j.seizure.2016.06.008.

[17] "EpDetect is a mobile phone application." http://www.epdetect.com (accessed 2021).

[18] N. A. Marzuki, W. Husain, and A. M. Shahiri, "MyEpiPal: Mobile Application for Managing, Monitoring and Predicting Epilepsy Patient," in Advances in Information and Communication Technology, ed: Springer International Publishing, 2016, pp. 383-392.

[19] R. S. Fisher, E. Bartfeld, and J. A. Cramer, "Use of an online epilepsy diary to characterize repetitive seizures," Epilepsy & Behavior, vol. 47, pp. 66-71, 2015/06 2015, doi: 10.1016/j.yebeh.2015.04.022.

[20] L. Irody. "Mobile patient diaries: Epdiary." http://www.irody.com/mobile-patient-diaries/ (accessed 2007).

[21] T. Rukasha, S. I. Woolley, and T. Collins, "Wearable epilepsy seizure monitor user interface evaluation," presented at the Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers, 2020/09/10, 2020. [Online]. Available: http://dx.doi.org/10.1145/3410530.3414382.

[22] "Empatica Medical-Grade Wearable Patient Monitoring Solutions." http://www.empatica.com/en-eu/ (accessed 2020).

[23] A. Van de Vel et al., "Non-EEG seizure detection systems and potential SUDEP prevention: State of the art," Seizure, vol. 41, pp. 141-153, 2016/10 2016, doi: 10.1016/j.seizure.2016.07.012.

[24] E. Bruno et al., "Wearable technology in epilepsy: The views of patients, caregivers, and healthcare professionals," Epilepsy & Behavior, vol. 85, pp. 141-149, 2018/08 2018, doi: 10.1016/j.yebeh.2018.05.044.

[25] A. Kos and A. Umek, "Wearable Sensor Devices for Prevention and Rehabilitation in Healthcare: Swimming Exercise With Real-Time Therapist Feedback," IEEE Internet of Things Journal, vol. 6, no. 2, pp. 1331-1341, 2019/04 2019, doi: 10.1109/jiot.2018.2850664.

[26] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran, "A review on wearable photoplethysmography sensors and their potential future applications in health care," (in eng), Int J Biosens Bioelectron, vol. 4, no. 4, pp. 195-202, 2018, doi: 10.15406/ijbsbe.2018.04.00125.

[27] P. Jallon, S. Bonnet, M. Antonakios, and R. Guillemaud, "Detection system of motor epileptic seizures through motion analysis with 3D accelerometers," presented at the 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009/09, 2009. [Online]. Available: http://dx.doi.org/10.1109/iembs.2009.5334770.

[28] W. J. C. van Elmpt, T. M. E. Nijsen, P. A. M. Griep, and J. B. A. M. Arends, "A model of heart rate changes to detect seizures in severe epilepsy," Seizure, vol. 15, no. 6, pp. 366-375, 2006/09 2006, doi: 10.1016/j.seizure.2006.03.005.

[29] H. L. Varela, D. S. Taylor, and S. R. Benbadis, "Short-Term Outpatient EEG-Video Monitoring With Induction in a Veterans Administration Population," Journal of Clinical Neurophysiology, vol. 24, no. 5, pp. 390-391, 2007/10 2007, doi: 10.1097/wnp.0b013e31812f6c11.

[30] C. Viboud and M. Santillana, "Fitbit-informed influenza forecasts," The Lancet Digital Health, vol. 2, no. 2, pp. e54-e55, 2020/02 2020, doi: 10.1016/s2589-7500(19)30241-9.

[31] M. Ntekouli et al., "A mapping of epilepsy's evolution: implementation of the proposed knowledge based model," Evolving Systems, vol. 9, no. 4, pp. 299-313, 2018.

[32] "Focal Onset Seizures (Partial Seizures)." https://www.healthline.com/health/partial-focal-seizure (accessed 2020).

[33] M. Velez, R. S. Fisher, V. Bartlett, and S. Le, "Tracking generalized tonic-clonic seizures with a wrist accelerometer linked to an online database," Seizure, vol. 39, pp. 13-18, 2016.

[34] S. R. Gouravajhala and L. Khuon, "A multi-modality sensor platform approach to detect epileptic seizure activity," in 2012 38th Annual Northeast Bioengineering Conference (NEBEC), 2012: IEEE, pp. 233-234.

[35] "EmbracePlus | Empowering Breakthroughs in Neurology Research | Empatica." https://www.empatica.com/embraceplus/ (accessed 2020).

[36] M.-Z. Poh, T. Loddenkemper, N. C. Swenson, S. Goyal, J. R. Madsen, and R. W. Picard, "Continuous monitoring of electrodermal activity during epileptic seizures using a wearable sensor," in 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, 2010: IEEE, pp. 4415-4418.

[37] "Groundbreaking Announcements as National Epilepsy Awareness Month Begins." https://www.prnewswire.com/news-releases/groundbreaking-announcements-as-national-epilepsy-awareness-month-begins-300176433.html (accessed 2019).

[38] S. A. McHale and E. Pereira, "An IoT Based Epilepsy Monitoring Model," Cham, 2021: Springer International Publishing, in Intelligent Computing, pp. 192-207.

[39] M. Copeland, J. Soh, A. Puca, M. Manning, and D. Gollob, "Microsoft Azure and Cloud Computing," in Microsoft Azure, ed: Apress, 2015, pp. 3-26.

[40] R. K. L. Ko, B. S. Lee, and S. Pearson, "Towards Achieving Accountability, Auditability and Trust in Cloud Computing," in Advances in Computing and Communications, ed: Springer Berlin Heidelberg, 2011, pp. 432-444.

[41] J. M. Peake, G. Kerr, and J. P. Sullivan, "A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations," Frontiers in physiology, vol. 9, p. 743, 2018.

[42] M. Abreu, A. Fred, H. Plácido da Silva, and C. Wang, From Seizure Detection to Prediction: A Review of Wearables and Related Devices Applicable to Epilepsy via Peripheral Measurements. 2020.

[43] "IoT Connectivity Options: Comparing Short-, Long-Range Tech." https://www.iotworldtoday.com/2018/08/19/iot-connectivity-options-comparing-short-long-range-technologies/ (accessed 2021).

[44] V. Tyndall et al., "Marked improvement in HbA(1c) following commencement of flash glucose monitoring in people with type 1 diabetes," (in eng), Diabetologia, vol. 62, no. 8, pp. 1349-1356, 2019, doi: 10.1007/s00125-019-4894-1.

[45] W. P. Institute. "Engineers creating miniaturized, wireless oxygen sensor for sick infants: Mobile, wearable device the size of a Band-Aid could allow babies to leave the hospital and be monitored from home." https://www.sciencedaily.com/releases/2019/11/191114154454.htm (accessed 2019).

[46] F. Porciuncula et al., "Wearable Movement Sensors for Rehabilitation: A Focused Review of Technological and Clinical Advances," (in eng), PM R, vol. 10, no. 9 Suppl 2, pp. S220-S232, 2018, doi: 10.1016/j.pmrj.2018.06.013.

[47] X. Wang, Y. Zhao, and F. Pourpanah, "Recent advances in deep learning," ed: Springer, 2020.

[48] S. Sengupta et al., "A review of deep learning with special emphasis on architectures, applications and recent trends," Knowledge-Based Systems, vol. 194, p. 105596, 2020/04/22/ 2020, doi: https://doi.org/10.1016/j.knosys.2020.105596.

[49] M. K. Siddiqui, R. Morales-Menendez, X. Huang, and N. Hussain, "A review of epileptic seizure detection using machine learning classifiers," (in eng), Brain Inform, vol. 7, no. 1, pp. 5-5, 2020, doi: 10.1186/s40708-020-00105-1.

[50] I. Kiral et al., "The Deep Learning Epilepsy Detection Challenge: design, implementation, and test of a new crowd-sourced AI challenge ecosystem," Challenges in Machine Learning Competitions for All (CiML), vol. 1, no. 1, 2019.

EEG-based Brain Computer Interface Prosthetic Hand using Raspberry Pi 4

Haider Abdullah Ali[1]
Faculty of Automatic Control and Computers
University POLITEHNICA of Bucharest, Bucharest, Romania
Department of Computer Techniques Engineering
Madenat Alelem University College, Baghdad, Iraq

Diana Popescu[2]
Department of Robotics and Production Systems
University POLITEHNICA of Bucharest, Bucharest, Romania

Anton Hadar[3]
Department of Materials Strength
University POLITEHNICA of Bucharest, Bucharest, Romania

Andrei Vasilateanu[4], Ramona Cristina Popa[5]
Faculty of Engineering in Foreign Languages
University POLITEHNICA of Bucharest
Bucharest, Romania

Nicolae Goga[6]
Faculty of Automatic Control and Computers
University POLITEHNICA of Bucharest, Bucharest, Romania
University of Groningen, Groningen, The Netherlands

Hussam Al Deen Qhatan Hussam[7]
Department of Civil Engineering
Madenat Alelem University College, Baghdad, Iraq

***Abstract*—Accidents, wars, or different diseases can affect upper limbs in such a manner so their amputation is required, with dramatic effects on people's ability to perform tasks such as grabbing, holding objects, or moving them. In this context, it is necessary to develop solutions to support upper limb amputees to perform daily routine activities. BCI (brain-computer interface) offer the ability to use the neural activity of the brain to communicate or control robots, artificial limbs, or machines without physical movement. This article proposing an electroencephalography (EEG) mind-controlled prosthetic arm. It eliminates the drawbacks like the high price, heaviness, and dependency on the intact nerves related to the myoelectric and other types of prostheses currently in use. The developed prototype is a low-cost 3D-printed prosthetic arm controllable via brain commands using EEG-based BCI technology. It includes a stepper motor controlled by Raspberry Pi 4 to perform actions like open/close movement and holding objects. The project has successfully implemented and achieve the aim to create a prototype of a mind-controlled prosthetic arm system in addition to the necessary experimental tests and calculations regarding torque, force, and the weight that the hand can carry. The paper proves the feasibility of the approach and opens the route for improving the design of the prototype to attach it to the upper-limb amputation stump.**

***Keywords*—Prosthetic; brain computer interface (BCI); electroencephalography (EEG); raspberry pi 4; EMOTIV**

I. Introduction

A lot of current prosthetic arms like myoelectric are still expensive, heavyweight, and difficult to use. This paper is an attempt to overcome such drawbacks using EEG-based BCI technology. BCI is an interesting field of research and applications [1] that can help in creating a new way of communication for persons with severe disabilities [2]. The field of BCI has witnessed a great interest especially concerning robotic devices control, with particular focus on health applications, where the utilization of BCI to control prosthesis devices is increasing the quality of life for the patients suffering from diseases causing permanent/temporary paralysis or suffering from the loss of the limb [3]. In such medical cases, the use of BCI gives great independence to patients with severe motor disabilities, providing them with the ability to control external devices such as prosthetic arms [4].

A BCI system, consisting of hardware devices and artificial intelligence software [5], uses several signal sources or techniques to record brain activity. Among these, the electroencephalography (EEG) method offers an appropriate signal along with stability and non-clinical risks [4]. EEG is a non-invasive technique for monitoring the activity of the brain. It uses electrodes placed on the scalp to measures the voltage variation of the brain neurons caused by the ionic current. An EEG device, generally in the form of a headset, records the brain waves of a person when he/she is thinking of a particular action or implementing a muscle movement. These waves are converted into commands to control an external device in real-time. The EEG method is costly effective, accurate, and gives the patient complete control [3]. It also offers the user the possibility of taking off the EEG device when feeling inconvenience [6]. EEG-based BCI plays an essential role in the area of prosthesis control [7], managing the interaction between the patient and the device without requiring an invasive surgical procedure to reconnect nerves and allow amputees to control their prosthesis [6].

Many upper-limb amputees depend on prosthetic hand/arm to restore some of the functionality needed to perform their daily activities. These prostheses still provide less than 50% of the ability of an intact limb despite the increasing enhancements and sophistication. Therefore, frequently rejected due to fatigue and frustration of using them [8].

In this context, the primary goal of this research is to operate and control a 3D-printed prosthesis hand using EEG signals detected from the brain. This would help in developing the next generation of prostheses [9] as new materials are appearing and being used [10]. 3D printing is commonly used in many medical applications like prosthesis, implants, surgical, etc. [9,11,12], orthotics field also benefitting the advantages of this technology customization to patient need and anatomy.

The use of brainwave signals for controlling prosthetic upper limbs is reported in the literature for different applications. Elstob et al. [12] presented a solution for controlling a five-degree freedom 3D printed prosthesis hand based on the use of EMOTIV EPOC+ EEG headset for detecting brain activity and Arduino Uno as a microcontroller. They also proposed two different software frameworks to control the prosthesis hand.

Beyrouthy et al. [13] have introduced a smart 3D printed prosthesis arm controlled via brain EEG signals obtained by using an EMOTIV EPOC headset. Smart sensors and actuators have been equipped to the arm to give the arm amputee smart feedback regarding the surrounding environment. Raspberry Pi III and Arduino Mega were used to control the prosthesis arm.

Bright et al [14] have developed a low-cost EEG brain-controlled prosthesis arm using a Neurosky Mindwave headset. The prosthetic arm has two main fingers' movements: flexion and extension. Arduino Uno has been used as a microcontroller to control several servo motors to carry out commands.

Chinbat et al. [15] have used Arduino Uno as a microcontroller and EMOTIV EPOC as an EEG headset to control the prosthesis arm via brain commands. The prosthesis arm is a low-cost 3D printed arm that contains smart sensors and actuators to give the arm amputee feedback to the surrounding environment. The prosthetic arm has six movements which are the movement of each finger individually and close all the fingers.

Mariacarla Staffa et al. [3] have proposed a novel approach to control open and close actions of a prosthesis hand. They used an Arduino Uno as a microcontroller, and EMOTIV EPOC+ headset to record brainwaves. They also proposed a "Weightless Neural Network-based classifier" for supervised classification. The experimental results showed that it is proper to use "WiSARD-Classifier" as a referring model for EEG signal classification.

More recently, Parth Limbani et al. [16] have developed a prosthesis arm controlled by brainwaves. They used a NeuroSky mind wave mobile sensor to obtain the EEG signals. An open-vibe software platform has been used for signal processing and classification. Python script has been used to send data from open-vibe to the Arduino microcontroller to control the prosthesis arm. They used a 3D printed arm made of PLA material with an overall weight of about 300 gm without motors.

Literature survey showed that the integration between BCI and prosthesis is a new and promising direction of study. It was also observed that improvements are needed to increase the control efficiency while reducing the complexity of the system. In this context, this research proposes a solution that eliminates the external wireless modules and uses Raspberry Pi 4 only without an Arduino board. Several degrees of force for the finger movements were considered, and force and torque were calculated to determine the maximum gripping force and the maximum weight that the prosthetic hand can take.

II. Materials and Methods

In this research, EEG was selected for prosthesis arm control as it provides a dynamic continuous neural activity with high temporal resolution [17]. The schematic representation of the "BCI-controlled prosthesis arm" is presented in Fig. 1.

A. EEG-based BCI Headset

The BCI tool used in the system is the "EMOTIV™ Insight headset" (Emotiv, USA). The "5-channel EMOTIV™ Insight headset" with semi-dry polymer sensors has an internal sampling rate of 128 samples per second per channel was used for experiments. This tool was specially built to allow the use of BCI and research, filtering the signals and wirelessly sending the data to the computer, which offers portability [17]. This mobile EEG headset offers whole-brain sensing and advanced electronics to produce clean, strong signals. The headset can be connected to computers, tablets, and phones. It uses "Bluetooth Low Energy" or "2.4 GHz wireless" (with dongle) to connect. Also, its "LiPo battery 450 mAh" is designed to last for eight hours of functioning. The electrodes of the headset must be placed on the scalp (see Fig. 2) to read the brain activity from the head. The raw gathered information is passed to the Processing Unit over a Bluetooth connection.

B. Stepper Motor

The motor used in this system is the KH42JM2 stepper motor (see Fig. 2). It is a Uni-polar stepper motor with high torque, low vibration, and low noise [18]. It is connected to the L298n stepper motor driver since 12V is required to operate the motor.

C. Processing Unit

The Processing Unit is a computer. The main function of the Processing Unit is to processes the EEG signals. Signal processing techniques are needed to eliminate the noise and artefacts from the raw EEG signals. This is done by applying digital filters. Feature extraction is where the system is responsible for the transformation of the pre-processed brain signals into feature values matching the underlying neurological mechanism. These characteristics are used by BCI to command output devices of the prosthesis arm [17].

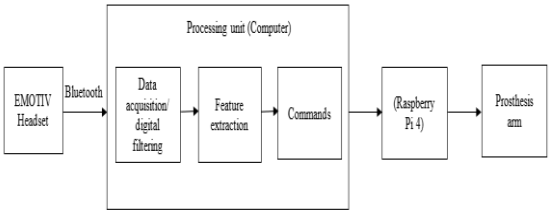


Fig. 1. Schematic Representation of the Prosthetic Arm System.

D. Raspberry Pi 4

"Raspberry Pi 4 Model B" is a fast, and powerful microcomputer that provides high-level performance that makes it an excellent choice for different control projects.

E. L298N Motor Driver

"L298N" is a high-power engine driver module. It is used to drive stepper motors as well as DC motors. It is composed of an "L298" motor driver IC and a "78M05" 5V regulator (see Fig. 2). This motor driver module can command up to 4 DC motors, or 2 DC motors with directional and speed control [19]. It is connected to the 12V stepper motor used in this system.

F. Prosthetic Hand

The prosthetic hand used in this system is a low-cost 3D printed prosthesis hand made of Polylactic Acid (PLA) - a biodegradable, strong, and durable lightweight material. The prosthesis has been developed in a project [20].

The prosthesis hand can perform actions like hand-shaking and picking up objects due to its hand close and hand opening movements (see Fig. 3). This is a prototype of the prosthesis hand that is intended to be developed in the future.

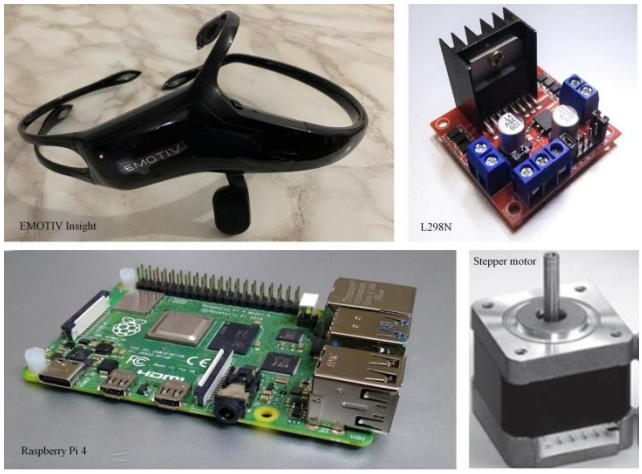


Fig. 2. Hardware Components.

Fig. 3. Prototype of the Prosthetic Hand.

III. System Implementation

This system consists of four important parts. The first one is the brainwave headset provided by EMOTIV™, the second one is the signal processing part which is carried out on the computer using EMOTIV™ cortex API, and Python for programming, the third part is the Raspberry Pi 4, and the last part is the prosthetic hand.

First, the computer is connected to the Raspberry Pi 4 through a wireless connection for commands transmission and to control the GPIO pins of the Raspberry Pi 4. Fig. 4 shows the circuit diagram of the connection between the stepper motor and Raspberry Pi 4. The stepper motor operated as a bipolar since one coil of the motor connected to the OUT1 and OUT2 of the motor driver, and the other coil connected the OUT3 and OUT4 of the motor driver.

The inputs of the L298N motor driver module IN1, IN2, IN3, and IN4 are connected to the Raspberry Pi 4 pins GPIO 18, GPIO 17, GPIO 27, and GPIO 22. A 12V power supply connected to 12V and ground pins of the L298N motor driver module. The ground pin of the Raspberry Pi 4 is connected in common to the ground pin of the L298N. An external 5V source is connected to the USB-C power port of the Raspberry Pi 4.

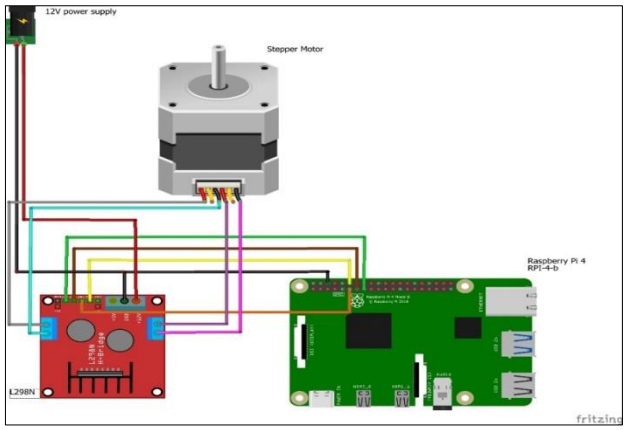


Fig. 4. Circuit Diagram of the System.

Sampled "*EMOTIV™ EEG*" signals are sent using a wireless connection to the computer that handles the processing part. Signal processing techniques are applied for noise and artefacts reduction on the raw "*EMOTIV™ EEG*" signals utilizing digital filters (Data acquisition and digital filtering module in Fig. 1). Feature extraction (module with the same name in Fig. 1) is where the system is responsible for transforming the pre-processed brain signals into feature values that match the underlying neurological mechanism. The features extracted from the headset are dispatched to the Raspberry Pi 4. Two commands were trained to control the open and close movements of the prosthesis hand. The problems encountered to control GPIO pins of the Raspberry Pi 4, especially with RPi.GPIO module were solved by using Secure Shell Protocol (SSH) to create and provide a connection channel between the computer and the Raspberry Pi 4. Python scripts are used on both computer and Raspberry Pi 4. Two Python scripts (one for hand open and the other for hand close) have been written for the Raspberry Pi 4 to control

the GPIO pins thus controlling the stepper motor. The Python code running on RPI (in Raspberry Pi 4) shown in Fig. 5 illustrates how to define the Raspberry Pi 4 GPIO pins where the stepper motor is connected, control the speed and the sequence of the motor where half step sequence has been used. When a user thinks to close or open the prosthetic hand (see Fig. 6), the control signal transmits via a wireless connection through SSH from the computer to the Raspberry Pi 4 to run the required Python script.

```
GPIO.setmode(GPIO.BOARD)
control_pins = [11,13,12,15]
for pin in control_pins:
GPIO.setup(pin, GPIO.OUT)

for i in range(25):
for halfstep in range(8):
for pin in range(4):
GPIO.output(control_pins[pin],halfstep_seq[halfstep][
pin])
time.sleep(0.002)
```

Fig. 5. Python Code for Hand Close Movement.

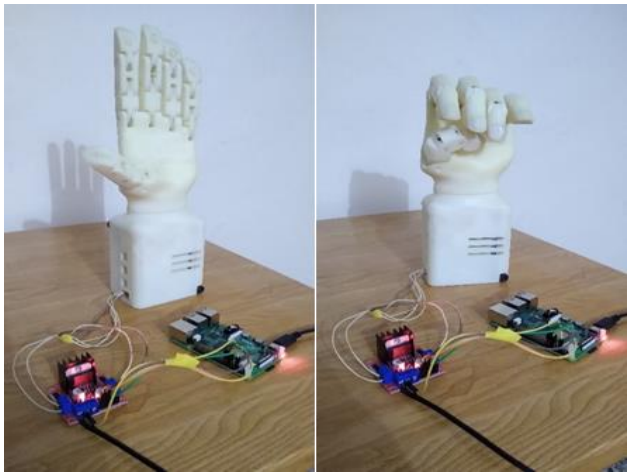


Fig. 6. Control of the Prosthetic 3D Printed Hand.

IV. Forces and Transfer of Torque

To calculate the necessary force and torque applied by the motor to move the fingers, and the maximum weight that the hand can carry, a force calculation, and laboratory test was conducted by making the prototype hold different weights. At a mass of 0.2 kg, each finger required approximately 1.962N to close, and about 9.81N in total (all fingers). However, to grip and hold objects, the force would need to be increased, about 20N (by experiment). As mentioned in the motor datasheet, the torque of the motor is 0.26Nm which is not large enough to apply appropriate gripping force; therefore, it was necessary to add a gear with a 2:1 ratio.

The KH42JM2 stepper motor used in this system is the most appropriate one due to its high torque, low vibration, and low noise. The relationship between speed and torque is inversely proportional. To gain higher torque, $\frac{Z_1}{Z_2}$ gear ratio has been used to calculate the speed of gear since using $\frac{Z_2}{Z_1}$ will give higher speed and thus lower torque.

Torque and force calculations:

To find the power:

$$p = \frac{T_1 * 2\pi * N_1}{60}$$

$$= \frac{0.26 * 2\pi * 48}{60}$$

$$= 1.307\ w$$

Where N_1 = 48 rpm

Calculating N_2:

$$N_2 = N_1 * \frac{Z1}{Z2}$$

$$= 48 * \frac{20}{40}$$

$$= 24\ rpm$$

The torque transferred by the gear is:

$$T_1 = \frac{60 * P}{2\theta * N_2}$$

$$= \frac{60 * 1.307}{2\theta * 24}$$

$$= 0.52\ N.m$$

at mass (m) = 0.2 kg

as one finger required 0.2 at gravity to close, f is:

$$F = mg$$

$$= 0.2 * 9.81$$

$$= 1.962\ N$$

Where g is the Gravitational acceleration.

Force for five fingers:

$$F = forces\ of\ one\ finger * 5$$

$$= 1.962 * 5$$

$$= 9.81\ N$$

One newton of force was added for gripping:

$$F = (1.962 + 1) * 5$$

$$= 14.81\ N$$

$$gripping\ force \cong 20\ N$$

The same procedure is applied using different mass values as shown in the Table I.

As shown in Fig. 7, when mass increases, the gripping force increases (linear).

According to the stepper motor specifications, the maximum torque meets 150*N*. If mass increases, this may damage the stepper motor or crash the prototype.

TABLE I. GRIPPING FORCE FOR DIFFERENT MASSES

Motor torque (*N.m*)	Torque with gear (*N.m*) (from calculations)	Mass (kg)	Force for one finger (*N*)	Force for five fingers (gripping force *N*)	≅
0.26	0.52	0.2	1.962	9.81	15
0.26	0.52	0.5	4.905	24.52	30
0.26	0.52	0.9	8.829	44.145	50
0.26	0.52	1.5	13.734	68.67	75
0.26	0.52	2	19.62	98.1	100
0.26	0.52	2.5	24.525	122.62	125
0.26	0.52	3	29.43	147.15	150

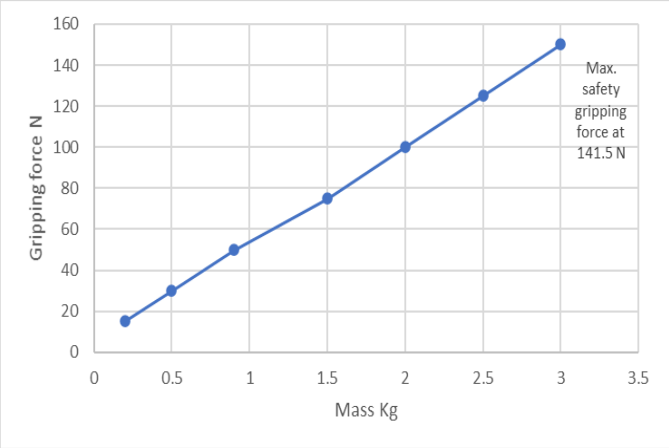


Fig. 7. Maximum Gripping Force and Weight.

The human hand can take a load of 2-3 kg easily. The force which can be applied by human hand will be:

$F = m.g$

$= 3 * 9.81 \rightarrow 27.24\ N$

Gripping forces:

$= (27.24 + 1) * 5 \rightarrow 141.5\ N$

In conclusion, the current prototype can hold 330 g (see Fig. 8) because the plastic wires used to move the fingers have lost their elasticity by time and usage. To make it handle some weight like a human hand, micro stainless-steel wires should be used. This will give the prototype the necessary elasticity and gripping force as calculated above, similar to the one of a human hand. Comparing to related works, most articles focusing on building a functional EEG-based BCI prosthetic arm without calculating or giving detail regarding torque, force, and the weight that the hand can carry. This paper has successfully achieved the aim of implement a mind-control prosthetic arm giving details and calculations regarding force and the maximum weight that the current prototype can carry to mimic the natural human hand. Additional experimental tests are intended to be achieved regarding measuring the gripping force of the hand using a hand-held dynamometer which is necessary to determine the required gripping force to grasp objects. This is the subject of further work.

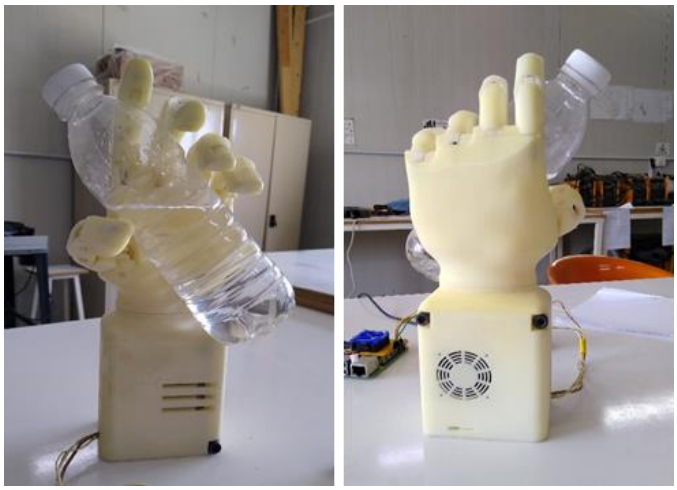


Fig. 8. Prototype Grips a Bottle of Water.

V. CONCLUSION AND FUTURE WORK

EEG-based BCI systems are still quite a new trend, especially in medical fields. In this paper, a structure and implementation of a mind-controlled prosthesis hand system have been presented. The prosthesis hand is made of strong and lightweight materials. The prototype contains one stepper motor controlled by Raspberry Pi 4 to perform open hand and close hand actions. Furthermore, significant experimental tests have been done on the prototype to determine and calculate the torque, force, and maximum weight that the hand can hold. As a result, this implementation has successfully achieved all the aims to create the prototype system of the project.

As future work, the prosthetic arm should be built with more movements involving movement of each finger individually and wrist movement using servo motors. The 3D printed hand design could be improved, so that the Raspberry Pi 4 board, servo motors, and the battery can be placed inside the arm. The entire system structure will be improved to eliminate the computer and to use Raspberry Pi 4 instead as processing and controlling unit. Moreover, the hand-held dynamometer should be used to test and calculate the griping force for the current prototype. All these tests and results will be used to design the next version of the 3D printed prosthetic hand with more improvements. Based on the prototype and considering the user's requirements, the goal is to develop a mind-controlled prosthesis upper limb using non-invasive EEG-based BCI, which is accurate, inexpensive, user-friendly, and smart. This prosthesis design will be optimized for reducing its weight and for allowing an easy, rapid, and stable fixation on the amputee stump.

REFERENCES

[1] M. H. Alomari, A. Samaha, and K. AlKamha, "Automated classification of L/R hand movement EEG signals using advanced feature extraction and machine learning," arXiv Prepr. arXiv1312.2877, 2013.

[2] R. Aldea and M. Fira, "Classifications of motor imagery tasks in brain computer interface using linear discriminant analysis," Int. J. Adv. Res. Artif. Intell., vol. 3, no. 7, pp. 5–9, 2014.

[3] M. Staffa, M. Giordano, and F. Ficuciello, "A WiSARD network approach for a BCI-based robotic prosthetic control," Int. J. Soc. Robot., vol. 12, no. 3, pp. 749–764, 2020.

[4] T. Yanagisawa et al., "Real-time control of a prosthetic hand using human electrocorticography signals," J. Neurosurg., vol. 114, no. 6, pp. 1715–1722, 2011.

[5] A. F. Glavan and C. Viorel Marian, "Cognitive edge computing through artificial intelligence," in 2020 13th International Conference on Communications (COMM), 2020, pp. 285–290, doi: 10.1109/COMM48946.2020.9142010.

[6] S. Al Taha Beyrouthy, J. K. Kork, and M. Abouelela, "EEG Mind Controlled Smart Prosthetic Arm–A Comprehensive Study," Adv. Sci. Technol. Eng. Syst. J., vol. 2, no. 3, pp. 891–899, 2017.

[7] S. Diwakar, S. Bodda, C. Nutakki, A. Vijayan, K. Achuthan, and B. Nair, "Neural Control using EEG as a BCI Technique for Low Cost Prosthetic Arms.," in IJCCI (NCTA), 2014, pp. 270–275.

[8] J. V. V. Parr, S. J. Vine, M. R. Wilson, N. R. Harrison, and G. Wood, "Visual attention, EEG alpha power and T7-Fz connectivity are implicated in prosthetic hand control and can be optimized through gaze training," J. Neuroeng. Rehabil., vol. 16, no. 1, pp. 1–20, 2019.

[9] S. Gannouni et al., "EEG-Based BCI System to Control Prosthesis's Finger Movements," 2020.

[10] S. Kholgh Eshkalak, E. Rezvani Ghomi, Y. Dai, D. Choudhury, and S. Ramakrishna, "The role of three-dimensional printing in healthcare and medicine," Mater. Des., vol. 194, p. 108940, 2020, doi: https://doi.org/10.1016/j.matdes.2020.108940.

[11] R. Marinescu, D. Popescu, and D. Laptoiu, "A Review on 3D-Printed Templates for Precontouring Fixation Plates in Orthopedic Surgery," J. Clin. Med., vol. 9, no. 9, 2020, doi: 10.3390/jcm9092908.

[12] D. Elstob and E. L. Secco, "A low cost eeg based BCI prosthetic using motor imagery," arXiv Prepr. arXiv1603.02869, 2016.

[13] T. Beyrouthy, S. K. Al Kork, J. A. Korbane, and A. Abdulmonem, "EEG mind controlled smart prosthetic arm," in 2016 IEEE international conference on emerging technologies and innovative business practices for the transformation of societies (EmergiTech), 2016, pp. 404–409.

[14] D. Bright, A. Nair, D. Salvekar, and S. Bhisikar, "EEG-based brain controlled prosthetic arm," in 2016 Conference on Advances in Signal Processing (CASP), 2016, pp. 479–483.

[15] O. Chinbat and J.-S. Lin, "Prosthetic Arm Control by Human Brain," in 2018 International Symposium on Computer, Consumer and Control (IS3C), 2018, pp. 54–57.

[16] P. Limbani, K. Chaudhari, M. Chauhan, and H. Patel, "Development of Artificial Robotic ARM based on Neural Control Interface."

[17] H. A. Ali, N. Goga, C. V. Marian, and L. A. Ali, "An Investigation of Mind-Controlled Prosthetic Arm Intelligent System," in The 16th International Scientific Conference "eLearning and Software for Education," 2020, pp. 17–26.

[18] "KH42series datasheet," Nidec Servo Corporation. http://www.nidec-servo.com/en/digital/pdf/KH42J.pdf.

[19] STMicroelectronics, "L298 datasheet," 2000. https://www.st.com/content/st_com/en/products/motor-drivers/brushed-dc-motor-drivers/l298.html.

[20] D. Fitzsimmons, S. Foy, D. Guijarro, and A. M. García, "Research, Design and construction of a prosthetic appendage: an investigation into the pros and cons of the prosthetic hand - with a solution to improve on past designs," 2017.

A Facilitator Support System that Overlooks Keywords Expressing the True Intentions of All Discussion Participants

Chika Oshima[1], Koichi Nakayama[4]
Faculty of Science and Engineering
Saga University, Saga, Japan

Tatsuya Oyama[2], Chihiro Sasaki[3]
Graduate School of Science and Engineering
Saga University, Saga, Japan

***Abstract*—This paper proposed the Keyword Movement Disclose System (KMDS), which allows a facilitator of discussion to watch a record of the moving keywords in a Discussion Board System (DBS). In the DBS, the discussion participants place each keyword in a box made for each item to be discussed. The keywords in the box were expected to show each participant's opinion and intention, because the participant's individual display was not disclosed to the other participants. Therefore, if the facilitator of the discussion can see the true opinions and intentions of all participants via the keywords in the boxes through the KMDS, the facilitator will be more appropriately advance the discussions and be able to draw conclusions based on diverse opinions. Moreover, the KMDS may contribute to the development of an artificial intelligence facilitator. In this paper, we conducted an experiment in which ten facilitators were asked to listen to a recorded discussion held by nine participants using the DBS. Five of the facilitators used the KMDS while listening the recorded discussion. It was suggested that KMDS may allow the facilitators to build a consensus from various viewpoints of the participants, although the results of the experiment did not show much difference depending on the conditions with/without KMDS.**

Keywords*—*Keyword movement disclose system; discussion board system; facilitator; putting keywords in box

I. Introduction

Although diverse perspectives shared by employees enhance corporate competitiveness, the discussions that concentrate diverse values can easily become confused [1]. The team members who differ on information diversity must engage in high quality communication to reconcile differing approaches to task completion [2]. Team members with diverse information perform better when they exchange information effectively. In contrast, their performances become poor when they rely on their own limited perspective [3]. In such a case, a facilitator [4] can appropriately advance discussions based on various viewpoints [1].

There are several studies that have automated the role of facilitators by generating a facilitator's questioning based on pattern-matching rules [5], analyzing the words written on a bulletin board and automatically facilitating the discussion [6, 7]. Although there are many discussion support systems [8, 9], few automatically facilitate during discussions. Since the facilitator role is different from those of a secretary or moderator, it is not enough to recognize the utterances of the discussion participants by voice recognition and to collect the hot opinions.

The role of facilitators includes designing the process, controlling the process, organizing and inspiring discussions, and forming an agreement [1, 4]. In this paper, we specifically focus on the fourth role. Sasaki [1] defines the fourth role as follows:

The facilitator elicits opinions from all participants as much as possible and gives a sense of conviction that the discussion was properly and adequately conducted. Identifying the right time, the facilitator encourages participants to reach conclusions and encourage consensus building.

We aim to develop an artificial intelligence (AI) facilitator that can autonomously facilitate discussions. The discussion addressed in this study calls for participants to form a collective consensus on a given set of problems and draw conclusions after considering various opinions [1, 10]. A competent facilitator reduces the peer pressure among participants and appropriately facilitates discussions based on different perspectives [1].

There are a lot of decision support systems that assist people by presenting information, knowledge, and analytical results: a system determines the best teacher using the C4.5 decision tree algorithm method [11], a system that suggests new fish that should be added to the aquarium tank based on the current environmental conditions of the aquarium [12], an automatic expert system that helps head of university department to choose lecturers and assign better course for them [13]. However, these systems do not help to make decision of the discussion.

"Wordy," which creates a word cloud based on lecture video content, allows a user to find the points they want to see in the video [14]. "Discussion Mining" generates structured data on discussion content semi-automatically and displays a graph structuralized with the pertinent information and keyword [15]. "Discussion Map" is a system, which supports consensus-building on multi-party conversations. Discussion participants themselves extract keywords during discussion and place them on the discussion map as nodes. The discussion is structured as a graph [16]. These systems allow the discussion participants to know the status of the discussion.

The Discussion Board System (DBS), ver.1.0, ver. 2.0 [1], and ver. 2.1 [10] has been developed to realize the role of a human facilitator. Before starting the discussion, items to be decided during the discussion are displayed at the top of each "box" in the DBS display. The DBS extracts nouns (called "keywords") from the participants' utterances and displays them. Each participant can put the keywords into a box according to their opinion and intention. Each participant's screen is invisible to other participants (a psychological safety zone [1] is one of the DBS's features); therefore, the true intentions of each participant will be expressed in the keywords they place in the boxes [1, 10].

The final decision made by the facilitator may change once they examine the keywords in each participant's box. How should the AI facilitator use the keywords in each participant's boxes to make a final decision? In this paper, ten facilitators listened to a discussion using the DBS, with/without watching a record of the keywords that showed when and who put each keyword into each box. Then, the final decisions the facilitators made while watching the record are compared with those made without watching the decisions.

II. Experimental Method

A. Overview

Fig. 1 shows an overview of the experiment. Ten male university students (Facilitators A–J) participated. They were paid a small reward to compensate them for their time. They were asked to listen to a recorded discussion in which nine other students (Participants O–W) discussed a fictional scenario. The participants belong to the same laboratory as Facilitators A–J. During the discussion, each participant used DBS ver. 2.1 (see Section IIB) on his own personal computer. Half of the facilitators, while listening to the discussion, were also watching the Keyword Movement Disclose System (KMDS) (see Section IIC), alerting them to when and who put each keyword in the box. After listening to the discussion, the facilitators answered questions.

B. Discussion Board System ver. 2.1

Fig. 2 indicates the display on DBS ver. 2.1 [2] that The participants used during the discussion. The boxes for each item to be discussed are displayed in a category area. Each participant puts the keywords extracted from their utterances into the box, based on their opinion and intention. The participants cannot see each other's category areas. If all participants in the discussion put the same word in the same box, the word's color changes to green. A participant can put any word in the parking area while debating whether or not to put it in the box. The comment area presents users with comments, in particular, ones that encourage those who have not spoken or moved any words on the screen for a certain period of time to join in the discussion.

C. Keyword Movement Disclose System

Half of the facilitators in this paper used KMDS to watch as keywords were moved, and they had access to data showing when and who, among the nine participants, put each keyword into a box. Fig. 3 shows the KMDS display. The top left of the screen shows the remaining time until the discussion will end, because this research assumes that company meetings feature discussion times that are set in advance. The top right area of the screen shows a timeline of the keywords' movement, and indicates when and who moved what keyword to which box. For example, one of the participants added the keyword "3 hours" to the item "performance time" when 48 minutes and 27 seconds remained for the discussion. The lower half of the window, in the category area, shows what kind of keyword each word is and how many participants moved it to each box. Although the DBS does not allow the discussion participants to see each other's category areas, the KMDS discloses the movements of the keywords of all participants to the facilitator.

KMDS usually displays these keyword movements in real time; however, in the experiment detailed in this paper, a pre-recorded discussion and a pre-prepared record of keyword movements in KMDS were used.

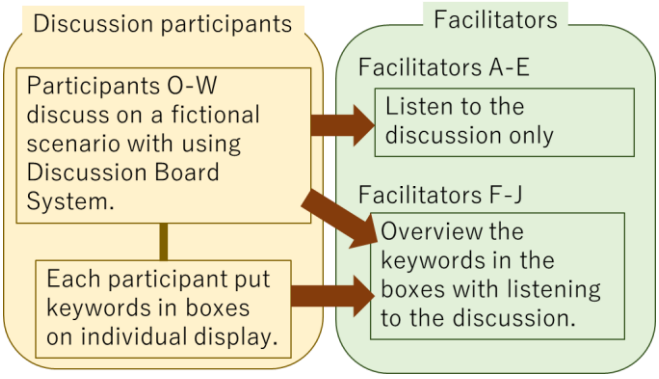


Fig. 1. Overview of the Experiment.

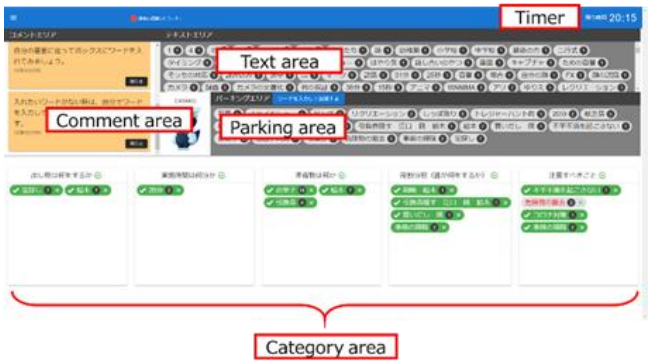


Fig. 2. Discussion Board System ver. 2.1 [1].

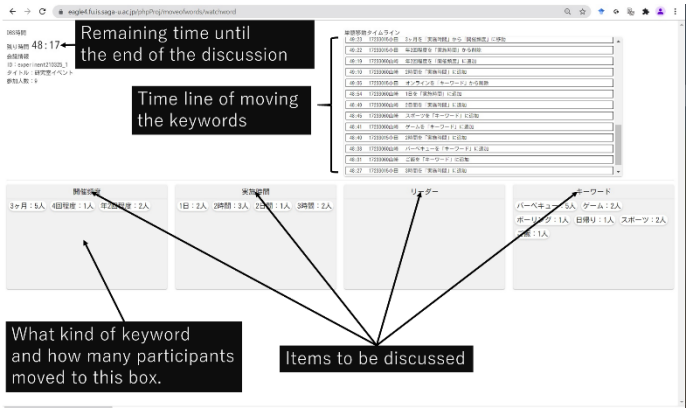


Fig. 3. Keyword Movement Disclose System.

D. Method

Each facilitator was asked to listen to a prepared one-hour discussion. The instructions for the facilitators were as follows:

From now on, you will listen to the recording of a discussion by nine people. They are discussing a "social event" to be held this year to build the relationships among the members of the laboratory. In the one-hour discussion, there are four items to decide.

Item 1: Frequency of holding the event

Item 2: Length of time per event

Item 3: Elect one leader and one sub-leader to put on the social event.

Item 4: Select two or three keywords that represent the social event from the following:

training camp / travel / welcome party / exchange party / game / sports / movie / birthday party / tournament / meal / drinking party / BBQ / festival / camping / online / other (write down here).

You are allowed to take notes on your smartphone and/or paper. The family names of the nine members are as follows:

Participants O, P, Q, R, S, T, U, V, and W (real names were indicated).

You are assumed to be a senior employee, compared to these nine discussion participants, and you have the authority to make final decisions on these four items. Listen to the discussion and come to the conclusion that you think is best for the laboratory. You are not someone who simply concludes an agreement or a secretary taking notes. Think about what the best conclusion is based on the discussion.

The KMDS was always displayed on half of the facilitators' computer screens, along with the following instructions for Facilitators F–J:

A record of the nine participants' moved words in a discussion support system appears on the KMDS screen in chronological order. Press the start button at the same time as the start button for the recording. Watch this screen while listening to the audio recording.

The facilitators were informed in advance that they would be asked what conclusion they consider to be the best for the laboratory concerning each of the four items after listening to the discussion.

After listening to the discussion, we will use a questionnaire to ask what conclusion you consider to be the best for the laboratory for each item, the reasons for your conclusions, and your degree of certainty that the conclusions are good ones.

In the recording, Participants O–W had discussed "a social event to be held this year to build the relationships among the members of the laboratory." The participants were asked to decide the same items as above four items during the discussion.

E. Questionaire

The questionnaire asked the following:

Question 1-1: Please draw the conclusion that you think is the best for the laboratory about "Item 1: Frequency of holding the event."

Question 1-2: Please explain why you drew this conclusion.

Question 1-3: Please indicate how confident you are of the suitability of the conclusion you made (5: very confident – 1: not sure at all).

Question 2-1: Please draw the conclusion that you think is the best for the laboratory about "Item 2: Length of time per event."

Question 2-2: Please explain why you drew this conclusion.

Question 2-3: Please indicate how confident you are of the suitability of the conclusion you made (5: very confident – 1: not sure at all).

Question 3-1: Please draw the conclusion that you think is the best for the laboratory about "Item 3: Elect one leader and one sub-leader to put on the social event."

Question 3-2: Please explain why you drew this conclusion.

Question 3-3: Please indicate how confident you are of the suitability of the conclusion you made (5: very confident – 1: not sure at all).

Question 4-1: Please draw the conclusion that you think is the best for the laboratory about "Item 4: Select two or three keywords that represent the social event from the following."

Question 4-2: Please explain why you drew this conclusion.

Question 4-3: Please indicate how confident you are of the suitability of the conclusion you made (5: very confident – 1: not sure at all).

Question 5: Please tell me as much as you can about what kind of situation, intention/opinion you think Participant U had in the discussion.

III. Results

A. Conclusions by the Participants and Facilitators

Fig. 4 to 7 show the rates of conclusions that the participants considered and which ones remained in each box; they also indicate whether the facilitators were working with/without KMDS on each item. Each x-axis of the figures indicates the answers. Each y-axis of the figures shows the ratio of the number of responses to the number of the participants/facilitators.

Fig. 4 shows the results of the conclusion on Item 1 (Question 1-1), "Frequency of holding the event." Although one of the facilitators without KMDS considered "once every one or two weeks" to be the best conclusion for the laboratory, the other answers and remaining keywords in the box for Item 1 were "once every three months."

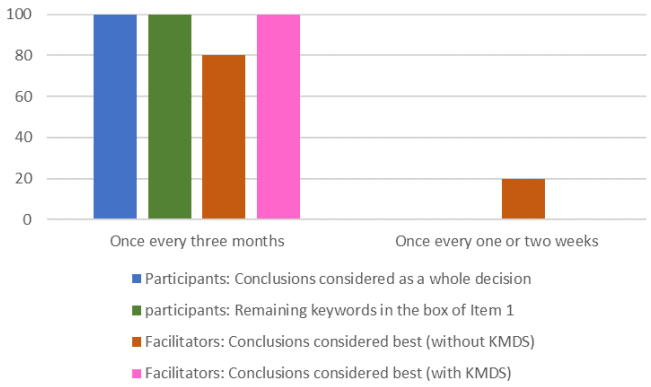


Fig. 4. Results of the Conclusion on Item 1.

Fig. 5 shows the results of the conclusion on Item 2 (Question 2-1), "Length of time per event." Although "two or three hours" and "three hours" were the most considered conclusion by the participants, "One day" was the most-often remaining among the keywords in the box for Item 2 and considered to be the best conclusion by the facilitators without KMDS.

Fig. 6 shows the results of the conclusion on Item 3 (Question 3-1), "Elect one leader and one sub-leader to put on the social event." All participants considered "Participant T is the best for the leader of this event" to be the conclusion of this discussion. In contrast, opinions on appropriate sub-leaders were divided among participants P, Q, and R. About half of the participants selected Participants Q or R, while most facilitators without KMDS concluded in favor of Participant R. Furthermore, the facilitators with KMDS, instead of giving the participants' names, based their conclusions on personal opinions, such as "participants who live near the university" (Facilitator F), "the sub-leader is expected to be a different year from the leader" (Facilitator F), and "everyone should take turns" (Facilitator H).

Fig. 7 shows the results of the conclusion on Item 4 (Question 4-1), "Select two or three keywords that represent the social event from the following." Most participants and facilitators concluded "game (online)." However, the keywords left in each participant's box varied.

B. *Reasons for the Conclusions*

This section focus on reasons for the conclusions. We classified the reasons (answers for Questions 1-2, 2-2, 3-2, and 4-2) into six types:

1) Drew the same conclusions and reasons as the majority of the participants.

2) Although they drew the same conclusions as the majority of discussion participants, the reasons given for the conclusions were different from the reasons the participants gave.

3) Drew the same conclusions and reasons as a minority of discussion participants.

4) Although they drew the same conclusions as a minority of discussion participants, the reasons for these conclusions were different from those given by the participants.

5) Although the conclusion was not based on the participants' comments during the discussion, the reason for the conclusion was same as what someone said.

6) The conclusions and reasons did not come from what anyone had said.

Table I shows the results after classifying the reasons for the conclusions, according to the above six types, and the results of the facilitators' confidence in their conclusions. There were no significant differences between the results of the conditions, without and with KMDS. However, in the reasons for Question 3 (answering Question 3-2), Participants F and H described why each of them did not write down the specific names of the participants (see Fig. 7) as follows:

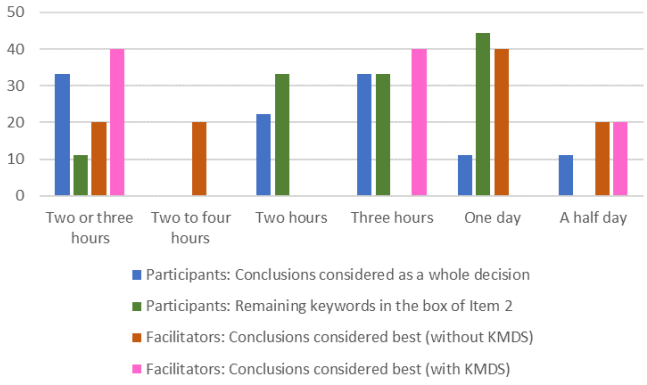


Fig. 5. Results of the Conclusion on Item 2.

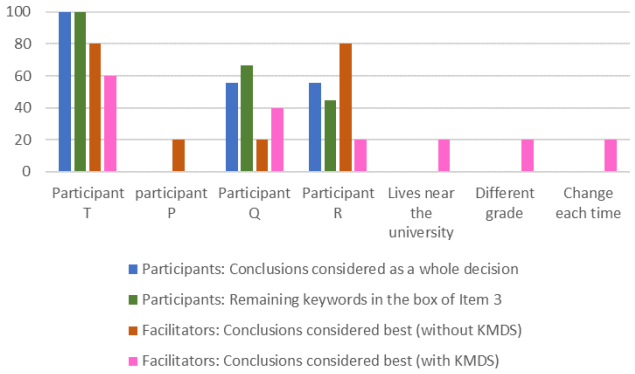


Fig. 6. Results of the Conclusion on Item 3.

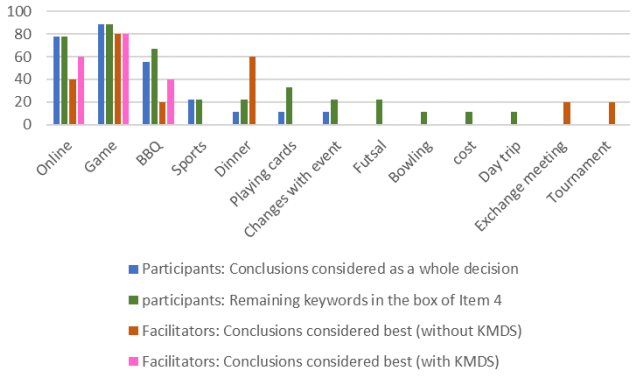


Fig. 7. Results of the Conclusion on Item 4.

Facilitator F: I did not write the specific name because I felt the discussion was a little forced. But it seems that this (Participant F's conclusion) was the direction of the participants' discussion. Moreover, I considered that it is this method (a participant who lives near the university, the years of leader and the sub-leader are different) that the decision on the event's content and communication among the participants proceeded smoothly (when they will hold the laboratory's event).

Facilitator H: The conclusions of the participants were basically a majority vote, but I felt that the consent of the participants was not so much obtained, and that some imposed their views on others. Some participants were concerned about the time for research presentations, classes, and job hunting. So, I think it is best for the participants with plenty of time to take charge in turn, according to the situation of each participant, rather than having one person takes charge.

C. Guessing Participant U's Opinions and Intentions

The results of Question 5, where the facilitators without KMDS were asked about Participant U were as follows:

Facilitator A: She (Participant U) wanted to match the faced and named of the laboratory members.

Facilitator B: Since she was just assigned to the laboratory, she thought they should have an opportunity to interact as much as possible.

Facilitator C: She wants to match the names with the faces of everyone in the laboratory.

Facilitator D: She was trying to make the meeting go smoothly by giving affirmative opinions.

Facilitator E: She was difficult to get into the discussion, because the story of the discussion changed a lot in the second half of the discussion.

The results of Question 5 where the facilitators had access to KMDS were as follows:

Facilitator F: She would like to try to match the names and faces of the laboratory members.

Facilitator G: I do not know who Participant U was.

Facilitator H: Basically, she made statements as prompted, and she did not have a positive opinion. Although she had enough opportunities to speak, she was a bystander. She was not interested in the conclusions of this discussion.

Facilitator I: She was not very enthusiastic about the event itself.

Facilitator J: Since she said that she could not match the laboratory members' names and faces, I think it was difficult for her to speak in this discussion where there were many senior members.

In fact, Participant U said that she wanted to match the names and faces of the laboratory's staff members. She spoke very little. She left the keywords in the box for Item 4, BBQ, sports, game, and online.

TABLE I. RESULTS OF REASONS AND CONFIDENCE

		Question 1		**Question 2**	
		Reason (six types)	***Confidence (1–5)***	***Reason (six types)***	***Confidence (1–5)***
Without KMDS	***A***	1	4	4	3
	B	5	5	2	4
	C	2	4	4	3
	D	2	4	4	4
	E	1	5	2	4
	Avg.	-	4.4	-	3.6
With KMDS	***F***	1	4	4	4
	G	1	5	2	5
	H	2	4	2	4
	I	1	4	1	4
	J	1	5	2	4
	Avg.	-	4.4	-	4.2
		Question 3		**Question 4**	
		Reason (six types)	***Confidence (1–5)***	***Reason (six types)***	***Confidence (1–5)***
Without KMDS	***A***	2	4	2	5
	B	2	4	2	5
	C	2	4	1	4
	D	2	4	2	4
	E	4	4	2	4
	Avg.	-	4.0	-	4.4
With KMDS	***F***	6	5	2	3
	G	2	5	2	4
	H	6	3	6	4
	I	2	3	1	4
	J	2	4	2	3
	Ave.	-	4.0	-	3.6

IV. DISCUSSION

More innovation will occur if the minority has a high level dissent and is highly involved in team decision making [17]. However, words spoken by multiple participants and/or at the end of a discussion often become a final conclusion if it is not possible to examine various opinions among the participants. A facilitator is responsible for avoiding such situations. If an AI facilitator only statistically analyzes and acquires the words that are spoken many times and adopts the opinions of the majority as a conclusion, it can result in abandoning various opinions.

Participants in the experiment's discussions believed that they had reached a conclusion on the items to be decided; however, and especially related to the items concerning the election of a sub-leader and the events themselves, no single conclusion was drawn. Most facilitators in the experiment listened to the discussion and drew their conclusions based on the opinion of the majority. However, two of the facilitators

who listened to the discussion while also using KMDS did not name a specific participant as sub-leader. One of the facilitators presented the ideal characteristics of leaders and sub-leaders, such as "people who live near the university" and "leaders and sub-leaders who are in different years." An examination of the keywords in the box makes it clear that the participants' opinions were not unified; moreover, Facilitator F must have noticed that the leader and sub-leader candidates were in different years. That fact may have led to the idea that leaders from different years are better for staging a successful event. Another facilitator concluded that "everyone takes turns at each event." He did not think that the participants' consent was much obtained about the leader and the sub-leader, unlike other facilitators.

The keywords that Participant U left in the box were not significantly different from that of other participants. However, she simply put each keyword in the box in the direction of the discussion of the entire participant, and did not seem to actively express her own opinions in the keywords. By comparing the timing at which a participant puts each keyword in a box with other participants, it may be possible to estimate the participant's degree of participation and agreement in the discussion.

In this way, the facilitators using KMDS could confirm the true opinions of all the participants based on the keywords they placed in the boxes. The facilitators could also know that the participants did not reach an agreement. Hence, it was suggested that the facilitators may be able to draw conclusions in a different direction from those available to the participants.

Although the results of the experiment did not show much difference based on the conditions (with/without KMDS), if the participants have more diverse ideas and a firmer hierarchical relationship, the usefulness of KMDS may have been further demonstrated. It is also undeniable that there were some differences in the qualities of individual facilitators.

Currently, the AI facilitator of our research may simply draw conclusions based on the majority keywords. In the future, the AI facilitators should also consider the lack of consensus, the presence of minority keywords, and the keywords that were mentioned only at the beginning of the discussion [18]. In some cases, the AI facilitator may need to present some new keywords that encourage a change in thinking. For example, when the sub-leader was not easily decided, the idea of "Everyone takes turns being in charge" was given by Facilitator H. In this way, skillful facilitators can also lead the participants to a desired conclusion [19, 20].

V. Conclusion

This paper proposed a Keyword Movement Disclose System (KMDS) that displays when and who puts a keyword into which box in a discussion support system (DBS). Ten facilitators listened to a discussion with/without KMDS. In some results, there were subjective differences in the conclusions drawn by the discussion participants and the facilitators, and between the facilitators with/without access to KMDS. The facilitators using KMDS could see the true opinions of all participants being expressed as they moved keywords into the boxes.

In the future, we will develop an AI facilitator which is able to appropriately advance the discussions based on various viewpoints and encourage consensus building, not just show the results of a majority vote.

References

[1] C. Sasaki, T. Oyama, S. Kajihara, C. Oshima, and K. Nakayama, "Online Discussion Support System with Facilitation Functio," International Journal of Advanced Computer Science and Applications, vol. 12, no. 8, paper 37, 2021.

[2] S. L. Marlow, C. N. Lacerenza, and E. Salas, "Communication in virtual teams: A conceptual framework and research agenda," Human Resource Management Review, vol. 27, no. 4, pp. 575-589, 2017

[3] A.C. Homan, D. Van Knippenberg, G.A. Van Kleef, and C.K. De Dreu, "Bridging faultlines by valuing diversity: diversity beliefs, information elaboration, and performance in diverse work groups," Journal of Applied Psychology, vol. 92, no. 5, pp. 1189, 2007.

[4] T. Mori, Facilitator training course. Japan: Diamond, Inc., 2007. (in Japanese).

[5] Y. Ikeda, and S. Shiramatsu, "Generating questions asked by facilitator agents using preceding context in web-based discussion," In 2017 IEEE International Conference on Agents, pp. 127-132, July 2017.

[6] T. Ito, S. Suzuki, N. Yamaguchi, T. Nishida, K. Hiraishi, and K. Yoshino, "D-Agree: Crowd Discussion Support System Based on Automated Facilitation Agent," Proc. of the AAAI Conference on Artificial Intelligence, vol. 34, no. 09, pp. 13614-13615, 2020.

[7] J. Haqbeen, T. Ito, S. Sahab, R. Hadfi, S. Okuhara, N. Saba, et al., "A contribution to covid-19 prevention through crowd collaboration using conversational AI & social platforms," arXiv preprint arXiv:2106.11023, 2021.

[8] R. Schroeter, J. Hunter, and D. Kosovic, "Vannotea: A collaborative video indexing, annotation and discussion system for broadband networks," Proceedings of the Second International Conference on Knowledge Capture: K-Cap, 2003.

[9] H. Tomobe and K. Nagao, "Discussion ontology: knowledge discovery from human activities in meetings," In Annual Conference of the JSAI, pp. 33-41, Springer, Berlin, Heidelberg, June 2006.

[10] T. Oyama, C. Sasaki, C. Oshima, and K. Nakayama: AI Facilitator Allows Participants to Conduct a Friendly Discussion and Contribute to Feasible Proposals, Communications in Computer and Information Science, vol. 1420, pp.523-530, 2021.

[11] H. Siahaan, H. Mawengkang, S. Efendi, A. Wanto, and A. P. Windarto, "Application of classification method C4. 5 on selection of exemplary teachers," Journal of Physics: Conference Series, IOP Publishing, vol. 1235, no. 1, pp. 012005, 2019.

[12] M. Mohammadi, and S. Jafari, "An expert system for recommending suitable ornamental fish addition to an aquarium based on aquarium condition," arXiv preprint arXiv:1405.1524, 2014.

[13] S. Fekri-Ershad, H. Tajalizadeh, and S. Jafari, "Design and Development of an Expert System to Help Head of University Departments," International Journal of Science and Modern Engineering, vol. 1, no. 2, pp. 45–48, 2013.

[14] W. Zhu, J. Zang, and H. Tobita, "Wordy: Interactive Word Cloud to Summarize and Browse Online Videos to Enhance eLearning," 2020 IEEE/SICE International Symposium on System Integration, IEEE, pp. 879-884, 2020.

[15] K. Nagao, K. Kaji, D. Yamamoto, and H. Tomobe, "Discussion mining: Annotation-based knowledge discovery from real world activities," Pacific-Rim Conference on Multimedia, Springer, Berlin, Heidelberg, pp. 522-531, 2004.

[16] R. Kirikihira, and K. Shimada, "Discussion map with an assistant function for decision-making: A tool for supporting consensus-building," International Conference on Collaboration Technologies, Springer, Cham. pp.3-18, 2018.

[17] C. K. De Dreu, and M. A. West, "Minority dissent and team innovation: The importance of participation in decision making," Journal of applied Psychology, vol. 86, no. 6, pp. 1191, 2001.

[18] K. Nishimoto, Y. Sumi, R. Kadobayashi, K. Mase, and R. Nakatsu, "Group thinking support with multiple agents," Systems and Computers in Japan, vol. 29, no. 14, pp. 21-31,1998.

[19] T. Proctor, "Creative problem-solving techniques, paradigm shift and team performance," Team Performance Management: An International Journal, vol. 26, no. 7/8, pp. 451-466, 2020.

[20] S. Ikari, Y. Yoshikawa, and H. Ishiguro, "Multiple-Robot Mediated Discussion System to support group discussion," 29th IEEE International Conference on Robot and Human Interactive Communication, IEEE, pp. 495-502, 2020.

A New Flipped Learning Engagement Model to Teach Programming Course

Ahmad Shaarizan Shaarani, Norasiken Bakar
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya
76100 Durian Tunggal, Melaka, Malaysia

***Abstract*—Online learning education at higher learning institutions has changed over the years as technology evolves. The main purpose of this study was to propose a new Flipped Learning Engagement (FLE) model. User testing to measure students' achievement was carried out in four separate groups namely Control Technology (CT) group, Experimental Technology (ET) group, Control Engineering (CE) group and Experimental Engineering (EE) group by using t-Test. The findings yielded that the experimental group (ET and EE) that underwent learning and teaching process by using the proposed FLE model obtained higher results or level of achievements as compared to the control groups (CT and CE) undergoing the conventional approach of teaching and learning. The study contributes mainly to the design and development of FLE model. FLE model proposed in this study can be beneficial to guide not only programming related educators but also for all educators that use flipped learning approach in their learning and teaching process. Future study should examine the proposed model in depth to improve it by adding new entities, hence, enabling its application in any related courses at various levels.**

Keywords*—Flipped learning engagement model; online programming course; student achievement; blended learning; technical based higher learning institutions

I. Introduction

The integration of new mobile technologies and online learning provides highly effective ways to help universities meet the expectations of the 21st century learners while addressing the challenges of limited resources and special needs of many students. One of the methods for mobile and online learning used is flipped learning (FL) [1].

In FL, the content of the subject matter is given to students in advance of school hours [2]. Through the traditional lecture method approach, lecturers usually go into the hall or classroom and present the lecture content by using teaching aids such as PowerPoint slides, projector screen and notes that have been printed in advance or use whiteboard and marker pen. Students will listen and understand the contents of the topics and copy the contents of the lectures or read the notes of the lectures after the sessions [3]. The method is known as teaching and learning approach. By FL method, however, the approach is different where lecturers have teaching materials in advance and ask students to first read and try to understand the content before entering the classroom [4]. This method is called learning and teaching approach. The learning process will take place before the face-to-face (F2F) meeting.

The most significant change in this new method is the initial exposure of students to lecture topics or course content. Students who enter the classroom or training room already have the basic knowledge of the course to be learned. The instructors have more time to work with the students in the classroom. In addition, instructors can focus on specific topics and assess the extent to which learners understand the topics involved [5]. By doing so, instructors have ample time to examine the effectiveness of the FL approach and compare students' performance for any subject matter.

During the class session, students will then take part in a variety of assignments and activities that are deemed appropriate for them in a given time. Doing these activities with peers will lead to team spirit. In groups, they will help each other to better understand the issues or topics, and are free to share and discuss ideas with their peers and lecturers. This approach is not new in the educational system. However, the approach is not yet fully applied in Malaysia, especially at higher level of education. There are few techniques for encouraging students and how they can be involved in FL.

In order to create an effective approach of FL in the classroom, this study has been undertaken to propose an appropriate model for building student engagement through FL. This study examined the elements for the development of a student engagement model in FL-based learning in programming course. The researchers suggested a new model to be used when implementing the FL approach for Information and Communication Technology (ICT) and engineering students at the Universiti Teknikal Malaysia Melaka (UTeM), Malaysia.

In addition, it is essential to develop the learning experience by providing a new model that can be used to prepare students for an active and collaborative learning style. This requires an appropriate model to guide educators and students through a blended learning approach when implementing FL. Student engagement is one of the key elements in FL approach [6]. This study applied FL through a blended learning approach by using online programming course. Later, the test was conducted to enhance the engagement model. Some background studies with few preliminary analyses have been conducted to propose the new model.

The findings of this study are to provide input to designers, educators and developers in their learning and teaching process using FL approach. The proposed Flipped Learning

Engagement (FLE) model can be used not only for programming related courses but for all courses implementing the FL approach. All the course contents such as lecture notes, instructions, assignments, and activities are provided online in FL.

II. Literature Review

There are many researchers studying the flipped approach to learning with positive results. In a study [7], all respondents agreed that learning has become more active after the class is being exposed to FL. Besides, over 90% stated that positive interaction with their students increases; students have greater access to course materials and instruction, students could work at their own pace; students are more likely to engage in critical thinking; and instruction becomes more differentiated and personalised. Almost 80% reported that positive interaction among students increases; that students are more likely to engage in collaborative decision-making; and that students are more likely to have choices on the ways to demonstrate their understanding on the subject matter. Over 50% agreed that students are more likely to have a choice about which learning tasks to undertake.

Students' engagement in the online programming course occurs when the students are able to browse self-study videos and activities on a weekly basis [8]. Students' behaviour, responsibilities and integrity are very important elements to ensure that the learning process takes place. Students' involvement consists of individual attitudes, thoughts, and behaviours and communication with others [9].

According to the UNESCO report in 2013, e-learning practices have evolved significantly with the increasing use of ICT in education as well as the development of network technologies [10]. The convergence of technology and educational development has led to e-learning innovation in higher education which is highlighted in the 21st century. Teaching and learning through online learning is deemed effective, as it is one way of optimising the effectiveness of teaching and learning processes.

In order to provide quality e-learning content as stated in the higher institutes' blueprint in Malaysia, there is a need to provide useful inputs to the implementers. The National Education Policy or known as *Dasar e-Pembelajaran Negara* (DePAN) is one of the initiatives aimed at providing a suitable framework for quality higher education learning. DePAN has targeted more than 30% of online learning after the year of 2021 [11]. The demand is even more necessary with the Covid-19 pandemic situation where teaching and learning processes are mostly conducted online [12]. In recent years, the concept of online and blended learning has been used to include an increasing number of online learning opportunities.

A. Research Gap

The study aimed to determine whether a FL approach using the FLE model could improve the overall student achievement. Other researchers have come up with a few models of student engagement through mixed learning [6][13][14]. However, thus far, no study has proposed a FL engagement model to improve student engagement and enhance student achievement especially in programming course.

In the initial inquiry, two courses namely mathematics and programming were identified as having the lowest student achievement in the ICT related field. Based on the results of the preliminary analysis, programming is considered to be the worst of these two courses. This is unacceptable indicator for the ICT and engineering faculties and the problem has to be addressed as programming is a basic course for students in both fields [15]. The result from the inquiry also indicated that there are several factors leading to this problem, such as; (i) students have difficulties in understanding the programming content that they consider too complex for them, (ii) students cannot link the concept of programming to the real-world problems and they have difficulty in finding solutions by using the programming technique.

Flipped learning approach is claimed to be a better approach in some other fields of study, but there is no specific study measuring the engagement of students in programing related courses using FL approach in programming course.

III. Study Description

This research applies online flipped classroom technique with a blended learning approach that involves the application of both F2F and online learning. Students must also undergo a personalized learning environment approach to learning in a flexible environment during the course of their studies. Students apply mobile learning, which means that they can learn at anytime, anywhere and using any device [16] [17]. Although many studies have mentioned the use of a FL approach, previous studies do not include FL with student involvement, especially in the programming course. Online programming course has been executed where students with internet access can learn on their own.

This online course has been developed using multimedia materials such as videos, animations, images and audio materials used by students in their learning process [18]. This research used a FL approach with the help of the lecturers who taught the course. In addition, social media communication channels were also used outside the classroom where the lecturers played their role as guides. Throughout the learning process, students used laptops, tablets and smartphone devices outside the classroom. As stated, this research has implemented a blended learning approach with F2F and online learning. Blended courses consist of between 30% and 79% of online activities F2F courses can include up to 29% of online activities, and full online courses include 80% to 100% of online activities [19].

This online programming course uses the WhatsApp application and other social media platforms, which connect lecturers and students to ensure student engagement. Lecturers will build the WhatsApp group for all students who take part in the online course. Then, the lecturers will guide the students through the group before, during and after a FL approach. This study uses a commitment model for student learning that has three phases, i.e. before class, during class and after class for the online programming course.

This study also contributes to student improved learning. The study also has an impact on the teaching method for educators. With ease of accessibility, students are likely to prefer to communicate through social media channels rather than F2F communication, reflecting their learning styles as well [20]. Rather than traditional lecture-presentations, they prefer a learning experience from the comfort of their time using online materials provided.

In addition, the results of this study are expected to assist higher education educators in how to implement blended learning which is a combination of F2F learning and online learning. A flipped model of learning is to guide educators. Blended learning is the combination of teaching and learning methods from both F2F, mobile and online learning which includes elements of both synchronous and asynchronous online learning options. Integrating new mobile technologies and online media provides highly effective ways to help universities meet 21st century learners' expectations while addressing the challenges of limited resources and many students' special needs. Hence, FL approach is one of the methods used in blended learning for mobile and online learning.

A. Research Objectives

The main objectives of this study are to develop a FLE model, measure the effectiveness use of this model to teach programming course. The study aims to develop a FLE model that can be used by ICT and engineering students. These include the following aspects:

- Develop an online programming course.
- Identifying and designing the FL approach for the programming course.
- Develop a new FLE model.
- Enhance student engagement on the basis of the FLE model for the programming course.

However, this paper only elaborate on the development of the new FLE model and the result of applying the proposed model to ICT and engineering students at UTeM.

B. Problem Statement

Computer programming is one of the most important courses offered to ICT and engineering students. At faculty of ICT, this course is a core program and for other engineering faculties at UTeM, this programming course which uses C++ language is also a compulsory course for the students [2]. Since C++ computer language is essential for both technical and non-technical studies, this programming course is deemed suitable for this study. In addition, from the pilot test which has been conducted earlier, the results of students in programming course using C++ language are among the lowest compared to other courses in the field of computer science.

Researchers believe that this problem could be reduced by applying the FLE model to teach programming course. Theoretically, this FL method through the proposed FLE model is feasible, but few problems would arise, such as how to ensure that students study those learning materials before the classroom schedule and how to enhance student engagement [21]. In order to overcome the issues of teaching and learning of programming course, this study is to design the FLE model programming course using C++ language for both ICT and engineering students.

C. Research Questions

In order to achieve the objectives of the study, two research questions were developed as follows:

RQ1: Can the proposed FLE model approach improve students' achievement in programming courses?

RQ2: Can student achievement for problem-solving skills in programming course be improved using FLE model approach compared to the traditional F2F learning?

IV. Development of FLE Model

The FLE model was a new model designed and developed to improve student engagement through a FL approach. Both the new model of engagement that was modified from the model of engaging online students organised around Self-Determination Theory (SDT) [22] and the flipped learning procedures were combined in this new FLE model.

The online programming course included an e-content such as videos, teaching materials, activities, links to materials and discussion forums for learning enrichment. All elements of programming were included in the online programing course for students of the Faculty of Information and Communication Technology (FTMK) and the Faculty of Electrical Engineering (FKE) at UTeM. In addition, this study also carried out a case study on user testing of online programming course among ICT and engineering students at UTeM.

A. Research Methodology

The study included two aspects: (i) design and development of the FLE model and online programming course with a FL approach that apply the FLE model for ICT and engineering students, and (ii) a case study on online programming course user testing with flipped classroom approach.

The discussion on the methodology of the study focuses on the design and development of the FLE model, the development of online programming course and the user test methodology based on the case study.

The engagement model was added to the elements of educational theory, instructional design (ID) elements, and interaction between lecturers and students. Each menu in the FLE model had a stated goal to improve the effectiveness of FL. The menus available in the models designed for this project included course information, course resources, interaction, active learning, frequent learning monitoring and making meaningful connections.

1) Course *information:* Course information was described as the objective of the course, the course period, and the total workload of the course.

2) Course resources: Course resources offered a video describing each subject weekly in just 5 to 10 minutes. These videos covered all important materials. They are the student resource; hence, the students should enjoy watching those videos.

3) Interaction: This is the interactive part between the lecturer-student-faculty. Any interaction that occurred was considered very important in the development and success of FL. Interaction occurred by means of student-to-student interaction, student-to-lecturer interaction, and student-to-faculty interaction.

4) Active learning: Active learning adapted few strategies such as small projects, online activities and self-assessment activities.

5) Frequent monitoring of learning: Frequent monitoring of learning was used to evaluate grades on a weekly basis and provide training on weekly topics.

6) Making meaningful connection: Making meaningful connection provided illustrative examples or case studies. All the assignments to the students need to be related with what the students' had learned.

The engagement model developed for this research involved several elements such as (i) behavioural involvement where students participated in learning activities such as completing assignments, attending classes, or contributing to the discussion, (ii) affective involvement which involved student emotional response or feelings (positive or negative) towards teachers, peers, learning, and school, and (iii) cognitive involvement which catered to the specific thought of students while participating in activities [23]. FL approach used a hybrid concept which was a combination of F2F and online learning. In addition, FL approach used active learning techniques and online technology to attract students. Using this flipped approach to learning and engagement with students, this study proposed a new model namely FLE.

The FLE model consisted of four phases: (i) instructional content, (ii) in-class non-technology activities, (iii) in-class technology activities, and (iv) evaluation and wrapping-up activities. All the suggested activities to be carried out were listed in the proposed model within each phase. The model offered an opportunity for in-class activities so that educators could use a digital approach if classrooms were equipped with technology and students could access the internet, otherwise educators could switch to a traditional approach and F2F activities with students without internet connections or digital tools.

The questionnaire was distributed to the experts in the field of online content and e-learning, comprising of professors and senior lecturers from various higher education institutions in Malaysia. A total of 13 experts participated in the validation process. Among the feedbacks received from the experts were as follows: (i) the need to clearly define between each category within the proposed FLE model; (ii) the need to properly define the criteria for FL; (iii) the need to clearly define the systematic approach of the model to student engagement; and (iv) the proposed model flow was useful in practice. All the relevant feedbacks obtained from the experts were taken into account for the improvement of the proposed FLE model.

The FL approach had three defining components: (i) moving lectures outside the classroom; (ii) delivering online courses including doing online assignments; and (iii) enhancing learning with classroom activities [24]. The format of the lectures in the suggested online course varied and evolved from slides to videos that included animations, infographics and other multimedia content. The online programming course was used to test this FLE model, allowing two-way communication between lecturers and students. Students learned independently through the online course prior to the lecture session. Online C++ programming course engaged students with assignments or quizzes that were resolved after learning from videos. Students should be more responsible for their learning through this approach where they could learn this online programming course at any time or anywhere.

The approach was expected to bring enjoyment to students and that they would not be bored to learn via the videos in the online C++ programming course because each video lesson was short, between 5 and 10 minutes. Furthermore, the videos also included animation and multimedia elements for learning programming comprehension. The new FLE model is presented in Fig. 1.

The new FLE model included a number of elements covering three main phases: (i) before the class phase involving instructional content; (ii) during the class phase involving in-class activities; and (iii) after the class phase involving evaluation and wrap-up. The components for each phase are as follows:

1) Before class phase

- Communication via the social media discussion channel.
- Lecturer prepares the development of online courses including learning materials (images, videos, teaching slides, modules, activities, links and references).
- Students gain and learn all the materials of online course.
- Students participate in interaction (either with lecturers or peers).

All of the elements in this phase are categorised as student-centered learning.

2) During class phase: As mentioned earlier, the class phase is divided into two with or without technology. In classrooms equipped with technology and internet access, the lecturer can use digital in-class activities. On the other hand, for the classroom without the technology, lecturers may use activities that do not require any digital tools or internet connection.

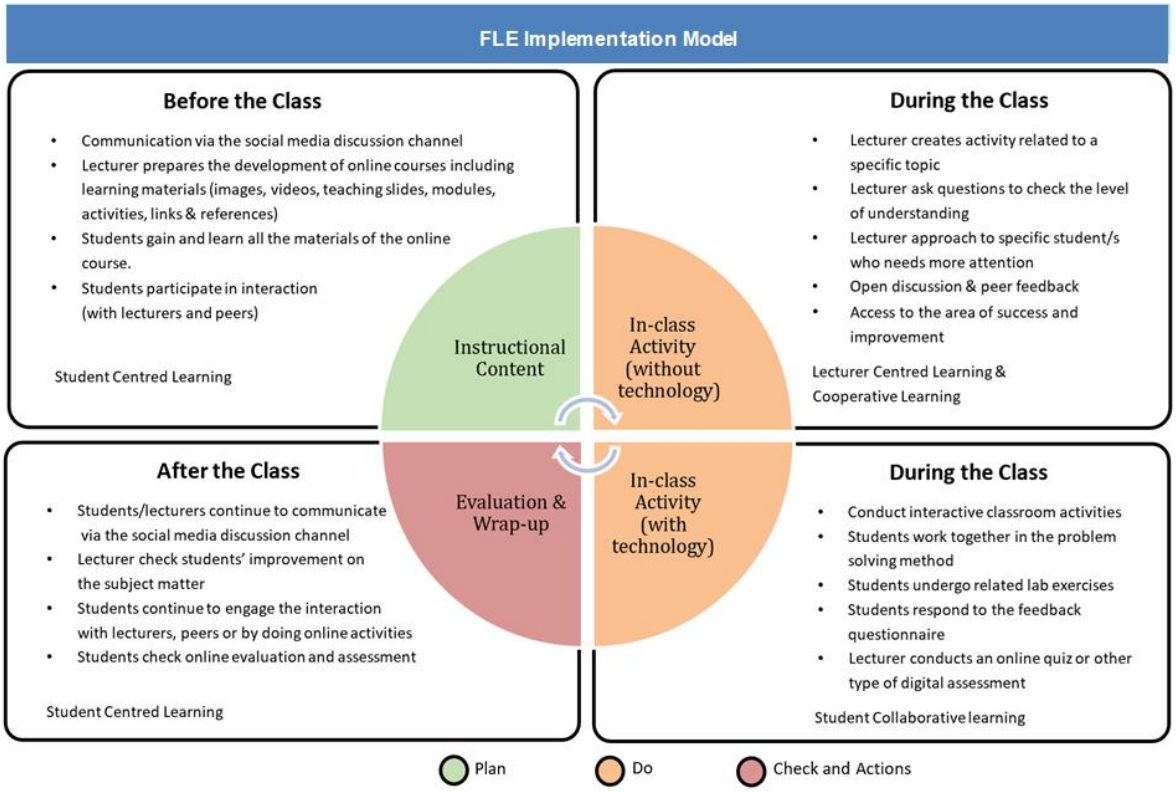


Fig. 1. The New Proposed FLE Model.

In-class activity (with technology):

- Conduct interactive classroom activities.
- Students work together on problem-solving method.
- Students undergo related lab exercises.
- Students respond to the feedback questionnaire.
- Lecturer conduct an online quiz or other type of digital assessment.
- Collaborative student learning.

The whole element in this phase is categorized as collaborative learning.

In-class activity (without technology):

- Lecturer creates activities related to particular topics.
- Lecturer asks questions to check understanding level.
- Lecturer's approach to specific student/s who need more attention.
- Lecturer-centered learning & cooperative learning.
- Open discussion and peers' feedback.
- Access to the area of success and improvement.

The whole element in this phase is categorised as a lecturer centred learning and cooperative learning.

3) After the class activity

- Students/lecturers continue to communicate via the social media discussion channel.
- Lecturer check students' enhancement on subject matter.
- Students continue to engage in interaction with lecturers, peers or online activities.
- Students check online evaluation and assessment.

All the element in this phase is categorized as student centered learning the whole element in this phase is categorised as student-centred learning.

V. Test Results for the Proposed FLE Model

The test results of user testing for online programming course as a whole were based on a case study conducted through quasi-experiments at UTeM. The students of the Faculty of Information and Communication Technology (FTMK) and the Faculty of Electrical Engineering (FKE) were chosen as respondents.

The distribution of the samples concerned is shown in Table I. For the sample list, a control group comprised 36 students (CT for FTMK samples and CE for FKE samples) and an experimental group 36 students (ET for FTMK samples and EE for FKE samples). Both the experimental and control groups were of the same class and at the same level. Students were selected equally from both groups based on their mid-term outcome. The analysis was carried out using simple percentage and average of student achievements.

TABLE I. SAMPLE DISTRIBUTION

Faculty	Group	Number of Samples
FTMK	Experiment Technology (ET) Control Technology (CT)	12 12
FKE	Experiment Engineering (EE) Control Engineering (CE)	24 24
	Total samples	72

Pre-test and post-test questions were developed to assess the student achievement. These measurements were used to test the effectiveness of the online programming course construct developed in comparison to the conventional teaching method used by the lecturers.

The questionnaire comprised three parts; Part A with 10 True or False types of questions of 20 marks. Part B with 10 multiple choice types of questions of 20 marks and part C with two structured questions of 60 marks, accumulating a total of 100 marks for the test. Part A and Part B measured the student overall cognitive level.

A combination of low cognitive level covering fact-related questions and high cognitive level category of questions covering applications, analysis, and synthesis question was utilized. Additionally, Part C also measured student higher order thinking comprising measurement type application, analysis and synthesis.

The C++ programming course was a three-credit course, with two- hour lecture per week and two-hour lab practical session per week. This course ran for a total of 14 weeks. Our quasi-experiment took four weeks starting from the 11th week until the 14th week of the semester. The e-content, and e-activities setup for the quasi-experiment comprised slides, six videos, seven e-activities, one quiz, and one tutorial.

The quasi -experiment involved the following stages: (i) Stage I (week 11): Before conducting any treatment, all students took a pre-test by using a set of questions to test their existing knowledge patterns for the control and experimental groups, (ii) Stage II (week 12 and week 13): Experimental groups went through self-directed learning process by using online programming course with the guide of instructors through social media channels. During this stage, students were required to communicate with instructors and peers. On the other hand, the control groups were taught using F2F conventional teaching and learning methods for two weeks, and (iii) Stage III (week 14): After the completion of each treatment process, all students underwent class activities with the same instructors to enhance their under-standing.

The class activities were conducted by using the F2F approach. At the end of week 14, all students were to do post-test questions on the same topic. Fig. 2 shows a summary of the three quasi-experimental stages that use the pedagogical approaches implemented during each stage for the quasi-experiment. To examine the level of change in terms of the variables measured in the quasi-experimental conditions, student achievement was measured by using the total marks in the pre-tests and post-tests. The total mark was standardized on a 0 to100 point scale, summing the scores from parts A, B and C. The results of these were used to answer RQ1 of the study. To answer RQ2, analysis of student achievement on the two programming questions in Part B, and the two structured (programming) questions in Part C of the pre-tests and post-tests were used to measure problem-solving skills of the students before and after the experiment.

SPSS 19.0 was used in analysing the quantitative data collected from the pre-tests and post-tests. Initially, descriptive statistics were produced to explore the frequency, mean, and standard deviation. Later, independent t-Tests were applied to analyse the differences in the two groups for overall student achievement. The F2F classes were recorded to observe students' experience.

Two types of data were collected: (i) student achievement data collected via pre-tests and post-tests taken by the control and experimental groups, (ii) student achievements in problem solving skills via pre-tests and post-tests. RQ1 was to find out whether the proposed FLE model approach improved student achievement in computer programming courses. To answer this question, the overall pre-tests and post-tests marks for both control and experimental groups were analysed by using the independent t-Tests. The results of the pre-tests are shown in Table II.

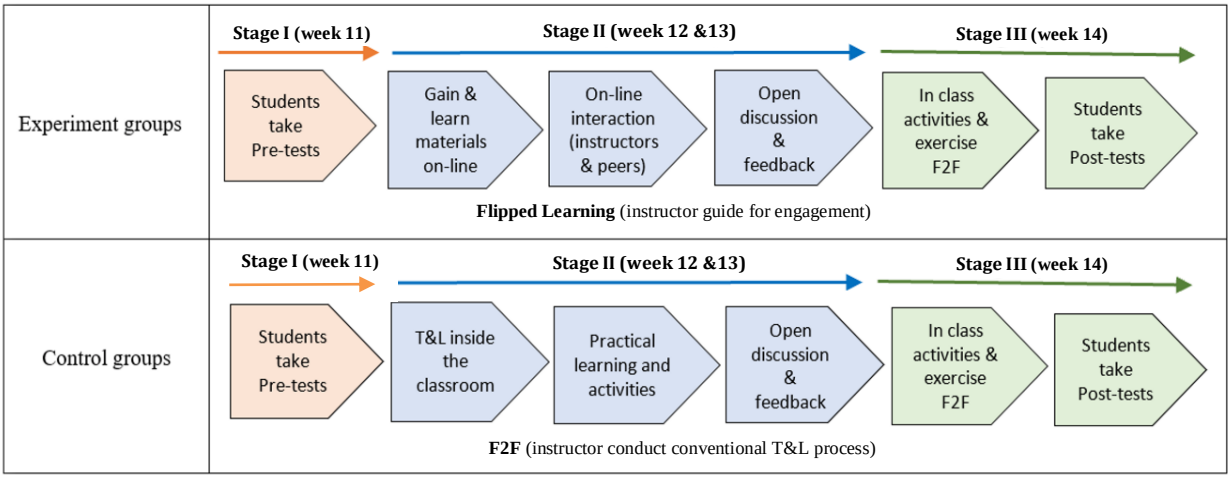


Fig. 2. Quasi-Experiment Stages for Experiment and Control Groups.

TABLE II. INDEPENDENT T-TEST RESULTS OF PRE-TESTS FOR EXPERIMENT AND CONTROL GROUPS

Group Statistics					
	Test Group	N	Mean	Std. Deviation	Std. Error Mean
PRE TEST	experiment	37	30.9189	11.06139	1.81848
	control	35	31.3143	12.00924	2.02993

Independent Samples Test						
		Levene's Test for Equality of Variances		t-Test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
PRE TEST	Equal variances assumed	.121	.728	-.145	70	.885
	Equal variances not assumed			-.145	68.688	.885

TABLE III. INDEPENDENT T-TEST RESULTS OF POST-TESTS FOR EXPERIMENT AND CONTROL GROUPS

Group Statistics					
	Test Group	N	Mean	Std. Deviation	Std. Error Mean
POST TEST	experiment	37	51.0811	12.13438	1.99488
	control	35	42.2286	11.94617	2.01927

Independent Samples Test						
		Levene's Test for Equality of Variances		t-Test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
POST TEST	Equal variances assumed	.041	.841	3.117	70	.003
	Equal variances not assumed			3.119	69.884	.003

The results were not significant (t=-.145,df=70, p > 0.05 = 0.885). The control group had a mean score of 31.31 whereas the mean score for the experimental group was 30.92. An independent t-Test showed that students in the two groups had no significant difference in their pre-test scores (t=-.145,df=70, p > 0.05 = 0.885). This result suggested that the students' prior knowledge of the Pointer topic were quite similar before the experiment. In other words, there was no significant difference between the experimental group (M=30.91; SD= 11.66) and the control group (M=31.31; SD= 12.00) for the pre-test scores.

The results of the independent t-Test were significant (t=3.117, df=70, p < 0.05 = 0.003). Table III shows the post-test results. The results showed there was a significant difference between the experimental group (M=51.08; SD= 12.13) and the control group (M=42.22; SD= 11.94) for the post-test scores.

The experimental group that used the FLE model and followed the process of online FL for the programming course achieved a better result compared to the control group that followed the conventional teaching and learning process. There was a significant difference in terms of student achievement between students using the FLE model approach as compared with students using conventional, F2F teaching method.

VI. DISCUSSION

The results of this study indicate that students that study programming course by using the FLE model approach demonstrate better achievement as compared to those in the conventional learning group. This finding is consistent with other findings [25][26]. This is probably related to the engagement in the online programming course and e-activities before the F2F classroom. The course design enables students to do self-paced learning outside of class to process the information introduced in the online content. Students benefit from the lecture videos that they watch before the F2F class, hence, highlighting supports for micro-lecture videos usage which are valid means for achieving desired learning goals [27]. In addition, e-activities provided in the online programming course allow students to consolidate their knowledge after watching the lecture videos in a timely manner.

Moreover, this study has an added advantage since the online programming course is built in-house and the F2F FL activities are personalized and adapted to the students' needs. The finding is consistent with other studies that claim students perform better and have a better understanding of the concepts when classes are personalized and adapted to individual needs [6][14][15][28]. Observation from the F2F classroom videos reveals that the FL method leads to increased student preparedness for the classes. This finding is consistent to claims by other researchers [26][29]. To ensure that a flipped

classroom is effective, this study follows guidelines and recommendation made for instructors suggested by a study [30] which consist of: (i) very organized pre-class assignments, (ii) tools for responsibility to guarantee that students will complete the pre-and post-class assignments, (iii) well planned and attractive activities for students to engage during lecture time, and (iv) all correspondence lines should be open for students to communicate with their instructors.

To indicate the importance of this study, the FLE model proposed is a new model intended to improve student engagement and enhance student achievement especially for technical based students at any higher learning institutions.

VII. Conclusion

In conclusion, the online programming course that was planned and developed on the basis of the FLE model user test is carried out using quasi-experiment. The study has successfully answered the research questions of this study. Based on the case studies, the samples are divided into two groups namely experimental (E) and control (C) groups. The findings are summarised as follows:

1) There is a significant difference in student achievement between students using the FLE model approach as compared to students who learn using the conventional F2F teaching method.

2) Assessment of online programming course by students provides a positive view on online programming course for learning and teaching using FL approach.

The second aspect of this study is user testing for the online programming course, being carried out on the basis of the effectiveness construct. The finding demonstrates that the use of online programming course based on this construct can improve the problem-solving skills of the experimental group compared to the control group. The overall achievement for the experimental group using the online programming course with FL approach is higher than the control group.

The results of this study also show that the FLE model has been successfully implemented to address problem-solving skills, hence, improving the overall performance. The findings of a learning construct based on learnability show that online programming course with FL approach succeeds in helping students gain confidence and problem-solving skills in handling C++ programming which has been identified as one of the most difficult courses for ICT and engineering students. The results of the study also support the ease construct usage where students do not have much trouble learning programming course through online realm. The results also demonstrate that online programming course that follows the FLE model can help students adapt learning and complete assignments based on student achievement capabilities through built-in modules. Finally, in general, the students are positive about the application used in learning online C++ programming course using FL approach. The overall result shows that the proposed FLE model used to design and develop the online programming course will enhance student performance for both ICT and engineering students in terms via improved results and programming problem-solving capabilities. This new proposed FLE model can be used by other programming related courses to help students understand and engage.

VIII. Future Work

Based on the findings of this research, a number of recommendations for future related studies are suggested as follows:

- Researchers should study this new FLE model in depth and make improvements by adding new entities that will allow this model to be used in the development of the prototype for other programming- based courses.
- The new proposed FLE model should be exploited fully in terms of its ability to provide a more comprehensive platform of performance reports through this online course usage for the purpose of monitoring the performance and progress of each user at university level in terms of student access to coursework.
- Researchers should study the adoption of a more flexible FLE model which can be changed by users especially educators to enhance activity-based assessments and web link additions to increase students' understanding of course sub-topics so developers can create data banks based on the Taxonomy Bloom.

References

[1] I. Amosa, T. Ahmed, O. Olufunmilola, and R. Olourotimi, “Effectiveness of blended learning and elearning modes of instruction on the performance of undergraduates in Kwara State, Nigeria,” Malaysian Online Journal of Educational Sciences, vol 5(1), 2017.

[2] R. Owston, R, D.N. York, and T. Malhotra, “Blended learning in large enrolment courses: student perceptions across four different instructional models,” Australasian Journal of Educational Technology, 2019.

[3] B. Sugeng, and A.W. Suryani, “Presentation-based learning and peer evaluation to enhance active learning and self-confidence in financial management classroom,” Malaysian Journal of Learning and Instruction, vol 15(1), pp. 173–201, 2018.

[4] H. Serin, and A. Khabibullin, “Flipped classrooms in teaching method courses at universities,” International Journal of Academic Research in Business and Social Sciences, vol 9(1), pp. 573–585, 2019.

[5] N. Falchikov, and J. Goldfinch, “Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks,” Review of Educational Research, vol 70(3), pp. 287–322, 2000.

[6] T. Baranova, L. Khalyapina, A. Kobicheva, and E. Tokareva, “Evaluation of students’ engagement in integrated learning model in a blended environment,” Education Sciences, vol 9(2), p. 138, 2019.

[7] N. Hamdan, P. McKnight, K. McKnight, and K.M. Arfstrom, “The flipped learning model: A white paper based on the literature review titled a review of flipped learning,” Flipped Learning Network, 2013.

[8] A. Rahman, S. Fatimah, M.M. Yunus, and H. Hashim, "An overview of flipped learning studies in Malaysia," Arab World English Journal (AWEJ), vol 10(4), 2019.

[9] R. Azvedo, “Defining and measuring engagement and learning in science: conceptual, theoretical, methodological, and analytical issues,” Educational Psychologist, vol 50(1), pp. 84–94, 2015.

[10] C. Depover, and F. Orivel, Developing countries in the e-learning era, Unesco, 2013.

[11] Ministry of Education, Executive Summary Malaysia Education Blueprint 2015-2025 (Higher Education), Homepage, https://www.um.edu.my/docs/defaultsource/about_um_document/media-centre/um-

magazine/4-executive-summary-pppm-2015-2025.pdf, last accessed: 22/03/2021.

[12] L. Mishra, T. Gupta, and A. Shree, "Online teaching-learning in higher education during lockdown period of COVID-19 pandemic," International Journal of Educational Research Open 1, vol. 100012, 2020.

[13] C.L. Bae, and M.H. Lai, "Opportunities to participate in science learning and student engagement: A mixed methods approach to examining person and context factors," Journal of Educational Psychology, vol 112(6), p. 1128, 2020.

[14] A. Bray, and B. Tangney, "Enhancing student engagement through the affordances of mobile technology: a 21st century learning perspective on realistic mathematics education," Mathematics Education Research Journal, vol. 28(1), pp. 173–197, 2016.

[15] H. Hasmy, N.S. Ibrahim, and A.Y. Kapi, "Analysis on C++ topic difficulties ranking: a case study on mechanical engineering students in UiTM Pasir Gudang," International Journal of Human and Technology Interaction (IJHaTI), vol. 4(2), pp. 5–12, 2020.

[16] I.M.F. Christensen, C. Kjaer, and P.S. Hansen, "Can self-paced, online learning provide teachers with the competences needed to successfully implement learning technologies?" Blended and Online Learning, p. 44, 2018.

[17] L.F.M.G. Pedro, M.MO.B. Cláudia, and M.N.S. Carlos, "A critical review of mobile learning integration in formal educational contexts," International Journal of Educational Technology in Higher Education, vol 15(10), 2018.

[18] S. Yu, "Online education and blended learning practice at Tsinghua university," Proceedings of EMOOCs 2019, unpublished.

[19] A.A. Aida, "Students satisfaction on blended learning in the school of sport sciences," Annals of Applied Sport Science, vol 8(1), 2020.

[20] P. Shea, "Introduction to online learning," Online Learning Journal, vol. 23(1), 2019.

[21] C.K. Lo, and K.F. Hew, "Developing a flipped learning approach to support student engagement: A design-based research of secondary school mathematics teaching," Journal of Computer Assisted Learning, vol. 37(1), pp. 142–157, 2021.

[22] K.F. Hew, "Towards a model of engaging online students: lessons from moocs and four policy documents," International Journal of Information and Education Technology, vol. 5(6), 2015.

[23] T. Baranova, L. Khalyapina, A. Kobicheva, and E. Tokareva, "Evaluation of students' engagement in integrated learning model in a blended environment," Education Sciences, vol. 9(2), p. 138, 2019.

[24] B.H. Shraddha, N.C. Iyer, S. Kotabagi, P. Mohanachandran, R.V. Hangal, N. Patil, and J. Patil, "Enhanced learning experience by comparative investigation of pedagogical ap-proach: Flipped classroom," Procedia Computer Science, vol. 172, pp. 22–27, 202.

[25] P.J. Muñoz-Merino, J.A. Ruipérez-Valiente, K.C. Delgado, M.A. Auger, S. Briz, V. de Castro, and S.N. Santalla, "Flipping the classroom to improve learning with MOOCs technology," Computer Applications in Engineering Education, vol.25(1), pp. 15–25, 2017.

[26] K. Wang, and C. Zhu, "MOOC-based flipped learning in higher education: students' participation, experience and learning performance," International Journal of Educational Technology in Higher Education, vol. 16(1), p. 33, 2019.

[27] R. Boateng, S.L. Boateng, R.B. Awuah, E. Ansong, and A.B. Anderson, "Videos in learning in higher education: assessing perceptions and attitudes of students at the University of Ghana," Smart Learning Environments, vol. 3(1), pp. 1–13, 2016.

[28] R.S. Jamuna, M.S. Ashok, and K. Palanivel, "Adaptive content for personalized e-learning using web service and semantic web," International Conference on Intelligent Agent and Multi-Agent Systems (IAMA), 2009.

[29] S. Findlay-Thompson, and P. Mombourquette, "Evaluation of a flipped classroom in an undergraduate business course," Business Education & Accreditation, 2014.

[30] I. Karabulut, Aliye, J.C. Nadia, and T.J. Charles, "A systematic review of research on the flipped learning method in engineering education," British Journal of Educational Technology, vol. 49(3), pp. 398–411, 2018.

Classifying Familial Hypercholesterolaemia: A Tree-based Machine Learning Approach

Marshima Mohd Rosli[1], Jafhate Edward[2], Marcella Onn[3]
Yung-An Chua[4], Noor Alicezah Mohd Kasim[5], Hapizah Nawawi[6]
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA[1, 2, 3]
40450 UiTM, Shah Alam, Selangor, Malaysia[1, 2, 3]
Institute for Pathology, Laboratory and Forensic Medicine (I-PPerForM) [4, 5, 6]
University Teknologi MARA, 47000 UiTM, Sungai Buloh, Selangor, Malaysia[4, 5, 6]
Faculty of Medicine, University Teknologi MARA, 47000 UiTM, Sungai Buloh, Selangor, Malaysia[4, 5, 6]

***Abstract*—Familial hypercholesterolaemia is the most common and serious form of inherited hyperlipidaemia. It has an autosomal dominant mode of inheritance, and is characterised by severely elevated low-density lipoprotein cholesterol levels. Familial hypercholesterolaemia is an important cause of premature coronary heart disease, but is potentially treatable. However, the majority of familial hypercholesterolaemia individuals are under-diagnosed and under-treated, resulting in lost opportunities for premature coronary heart disease prevention. This study aims to assess performance of machine learning algorithms for enhancing familial hypercholesterolaemia detection within the Malaysian population. We applied three machine learning algorithms (random forest, gradient boosting and decision tree) to classify familial hypercholesterolaemia among Malaysian patients and to identify relevant features from four well-known diagnostic instruments: Simon Broome, Dutch Lipid Clinic Criteria, US Make Early Diagnosis to Prevent Early Deaths and Japanese FH Management Criteria. The performance of these classifiers was compared using various measurements for accuracy, precision, sensitivity and specificity. Our results indicated that the decision tree classifier had the best performance, with an accuracy of 99.72%, followed by the gradient boosting and random forest classifiers, with accuracies of 99.54% and 99.52%, respectively. The three classifiers with Recursive Feature Elimination method selected six common features of familial hypercholesterolaemia diagnostic criteria (family history of coronary heart disease, low-density lipoprotein cholesterol levels, presence of tendon xanthomata and/or corneal arcus, family hypercholesterolaemia, and family history of familial hypercholesterolaemia) that generate the highest accuracy in predicting familial hypercholesterolaemia. We anticipate machine learning algorithms will enhance rapid diagnosis of familial hypercholesterolaemia by providing the tools to develop a virtual screening test for familial hypercholesterolaemia.**

Keywords*—Familial hypercholesterolaemia; predicting FH; machine learning algorithms; tree-based classifier

I. Introduction

Familial hypercholesterolaemia (FH) is the most common and serious form of inherited hyperlipidaemia, and is characterised by severely elevated low-density lipoprotein cholesterol (LDL-C) levels. It is an important cause of premature atherosclerosis and coronary heart disease (CHD), but is potentially treatable [1], [2]. Globally, the prevalence of heterozygous FH has been estimated at 1:200–1:500 [3]. However, the majority of FH individuals remain under-diagnosed and under-treated, resulting in lost opportunities for preventing premature CHD (pCHD).

In Malaysia, the prevalence of hypercholesterolaemia and severe hypercholesterolaemia is approximately 60% and 3%, respectively, and we have recently reported a high community prevalence of clinically diagnosed FH of 1:100 [4]. Further, FH was detected in about 35% of patients with pCHD [5]. With an estimated Malaysian population of 32 million, it is projected that at least 64,000–160,000 individuals are affected, the majority of whom are likely to be undiagnosed or inadequately treated. However, the prevalence of confirmed FH is not well established in Malaysia because DNA testing is costly and not commonly available in primary care clinics. Screening based on the lipid profile and LDL-C related measures is a reasonable alternative approach to assess the risk present, but contends with problems [6].

FH is usually diagnosed using four well-known diagnostic instruments: Simon Broome (SB;[7]), Dutch Lipid Clinic Criteria (DLCC; [8]), US Make Early Diagnosis to Prevent Early Deaths (US MEDPED;[9]) and Japanese FH Management Criteria (JFHMC; [10]). In Malaysia, the reports of FH are highly varied in terms of diagnostic method [11], due to lack of consensus in usage FH diagnostic criteria for screening of FH. Additionally, the input variables and the outcome of each diagnostic criteria are different, therefore, any attempt to combine multiple diagnostic criteria into one diagnostic criteria is not possible. According to the national standard guideline for management of dyslipidaemia, clinicians may use the DLCC, SB and US-MEDPED tools to diagnose patients [12]. A handful of Malaysian FH study groups [13], [14] already reported their research findings based on these diagnostic criteria [15], [16].

The above-mentioned FH diagnostic instruments are traditionally paper-based, and the diagnostic outcomes are manually scored by healthcare providers. This practice, however, has various well-known shortcomings that are typical of paper-based data collection systems, such as the expense of paper and space constraints for printing and storage. In addition, as diagnostic criteria specifically designed for Malaysians are still not available, the need to

choose among multiple instruments of diagnostic criteria means that diagnosing FH has become time-consuming and laborious.

Machine learning techniques have been widely applied in the field of medical diagnostic applications because they can perform large-scale data analysis and predict a potential outcome efficiently [17], [18] [19], [20]. These techniques incorporate the use of artificial intelligence, which learns the dataset's patterns, and subsequently designs and trains a predictive model. The model seeks to make predictions on new data and is commonly used for classification, decision-making and rule-mining. Using these techniques can help predict and identify FH individuals who are at risk of developing pCHD, which in turn opens a major opportunity in healthcare.

Therefore, the goal of our research is to determine the most relevant features of the above-mentioned four diagnostic instruments that are useful in the diagnosis of FH in Malaysian patients, using machine learning models. We apply three classification models (random forest, gradient boosting and decision tree classifiers) with a recursive feature elimination (RFE) algorithm to perform feature selection by iteratively training a model, ranking features, and then removing the lowest ranking features. We anticipate that the pertinent features selected by the three classifiers will assist Malaysian FH study groups to construct a set of population-based diagnostic criteria for FH screening in upcoming studies.

The contributions of this paper are:

- We present a range of different tree-based machine learning approach with Recursive Feature Elimination method for detection of FH in Malaysian population.
- We use the largest number of primary health care records that contain a diagnosis of FH according to four well-known diagnostic instruments (DLCC, SB, JFHMC and US MEDPED) conducted in Malaysia.
- We determine the novel predictive features that are useful in the diagnosis of FH in Malaysian patients, using machine learning models.

II. Related Work

In this section, we start with related work that discuss studies on the prediction and classifications of FH using machine learning techniques. Then, we discuss recent studies that predict the presence of FH-causing genetic mutations. Finally, we discuss the importance of tree-based machine learning techniques that provides important insights to this research.

Several studies on the prediction and classifications of FH using machine learning techniques have been conducted by various researchers. For example, Shi et al. (2014) used logistic regression [21] to estimate the prevalence of FH and its treatment for adults in a random Chinese population and to assess the associated risk factors. They found that there was a high prevalence of phenotypic FH among those aged $\geqslant 50$ years, which suggests that FH is common and remains under-detected among Chinese population. Their findings were consistent with other researchers showing under-detection and under-treatment of FH in other countries [22], [23].

A group of researchers used random forest as a machine learning approach, with electronic health record data from Stanford Health Care and random forest classification for identification of potential FH patients [19], [24]. Their aims were to promote early diagnosis and timely intervention for high-risk pCHD patients with undiagnosed FH by using random forest for performing features of FH score.

Weng et al. (2015) used a stepwise logistic regression method [25] to predict FH, involving nine variables. The stepwise logistic regression was used to improve the identification of individuals in primary care settings who could be prioritised for further clinical assessment. The study also removed one of the variables, family history, which eventually resulted in significant improvement in discrimination.

Later, the same group of researchers published a new study of identifying and managing possible FH using SB criteria in primary care setting [26]. The study used six variables (demographic data, family medical history, physical signs, lipid characteristics and statin used in medication habits) and two methods (descriptive analysis and Wald's method). Their results showed 118 of 831 patients who were at least 18 years of age had blood total cholesterol levels >7.5 mmol/L, and 32 of them were without previous diagnosis of FH.

Pina et al. (2020) used three machine learning algorithms to predict the presence of FH-causing genetic mutations in two independent FH cohorts: a classification tree (CT), a gradient boosting machine (GBM) and a neural network (NN) [27]. They found that the three machine learning algorithms performed better than the clinical DLCC in predicting carriers of FH-causative mutations by evaluating the area under receiver operating curve (AUROC) parameter. This indicates that machine learning techniques may help the confirmation of FH, especially in the context of primary care or specialist clinics such as specialist lipid, cardiology or endocrinology clinics, which may prompt family cascade screening for detection of more FH among family members.

Although several techniques have been proposed to resolve the challenges associated with the prediction and classification of FH, we found that there is still a lack of research in predicting and classifying FH patients with machine learning techniques to determine important features of FH diagnostic criteria to diagnose FH. As mentioned earlier, only a few groups of researchers have apparently used random forest to predict FH, and none appear to have utilised other tree-based machine learning techniques such as decision tree and gradient boosting, which generally involve human-like algorithms that are compatible with all four diagnostic instruments. Moreover, there is a scarcity of reports on the use of different machine learning models in predicting FH in the local Malaysian population.

The tree-based machine learning techniques were widely used for solving classification problem in prediction of disease due to ability to deal with many clinical predictors of disease. Decision tree model is the most fundamental of the tree-based

approach that able to generate human-understable rules without requiring much computational effort. Random forest model is an ensemble of decision trees, which utilise bagging aggregation approach to gain many trees and average over multiple trees for reducing the possibility of overfitting. Gradient boosting model is another variation of an ensemble method, which uses subsets of the original data to generate a series of average performing models and then "boosts" their performance by merging them using a specific cost function. Hence, the decision tree, gradient boosting and random forest models were explored in this study to detect FH. We expect that the outcome from the best classification model can be used to identify the relevant features that generate the highest accuracy in predicting FH, which potentially facilitate the development of Malaysian-based FH diagnostic criteria in future.

III. Materials and Methods

A. Study Design and Population

In this study, we used a secondary dataset containing 5248 individuals from all states in Malaysia, who were recruited from community health screening programmes and specialist lipid clinics in Malaysia, such as the Universiti Teknologi MARA (UiTM) Specialist Lipid Clinic, UiTM Cardiology Clinic and National Heart Institute (IJN), from 2011 to 2019. Individuals with secondary causes of hypercholesterolaemia, such as nephrotic syndrome, hypothyroidism, chronic kidney disease and cholelithiasis, were excluded from the study.

The dataset consists of 24 raw features, with 54.05% of the dataset having complete fields. Because of the low percentage of complete fields, we applied univariate and multiple imputation methods to replace the quantitative missing values to overcome the limitation of missing values in the dataset. After the missing data were successfully imputed, the dataset was further processed to reduce the number of features with weak relations with the target feature.

The cleaned dataset comprised 16 features describing the patients' demographic data and clinical characteristics: age; gender; smoking habit; patient history of pCHD, cerebrovascular accident (CVA) or peripheral vascular disease (PVD), and diabetes; lipid profile including high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC) and triglycerides (TG); family history of FH, hypercholesterolaemia and pCHD; patient's physical symptoms of corneal arcus and tendon xanthomata; and whether the patient was on lipid-lowering therapy. Table I shows the demographic and clinical characteristics of the study population.

B. Decision Tree Approach and Algorithm

The experiments were conducted using SPSS Modeler 18 and Python. Three classification models were used to train and test the dataset: random forest, gradient boost and decision tree. The cleaned dataset was partitioned into 70:30 ratios for training and testing; 70% (3674 instances) of the overall dataset were labelled X_train and used to train the classification model, and 30% (1574 instances) of the dataset were labelled X_test and used to test the model.

TABLE I. Demographic and Clinical Characteristics of the Study Population (n = 5248)

Feature	Total
Age, mean (SD)	41.41±15.404
Gender	
Male	(2009) 38.3%
Female	(3238) 61.7%
High-density lipoprotein cholesterol, mean (SD)	1.29±0.40
Baseline low-density lipoprotein cholesterol, mean (SD)	3.27±1.14
Triglycerides cholesterol, mean (SD)	1.69±1.17
Total cholesterol, mean (SD)	5.32±1.43
Tendon xanthomata	(22) 0.4%
Corneal arcus	(263) 5.0%
Lipid-lowering therapy	(383) 7.3%
Smoking	(630) 12.0%
Diabetes	(342) 6.5%
History of coronary heart disease	(104) 2.0%
History of cerebrovascular accident or peripheral vascular disease	(64) 1.2%
Family history of familial hypercholesterolaemia	(84) 1.6%
Family history of hypercholesterolaemia	(728) 13.9%
Family history of coronary heart disease	(682) 13.0%

In this study, we used multi-class classification for the DLCC and SB because these diagnostic instruments involve classifying into one of more than two classes. We used binary classification for the JFHMC and US MEDPED diagnostic criteria because these diagnostic instruments classify into one of two classes. We applied an RFE algorithm with the three classification models to select a subset of the most relevant features for the dataset and to eliminate weak features identified as noises, which might affect the performance of the models. The RFE approach consisted of three steps: (a) training the classification model to determine initial importance scores, (b) removing the bottom features with the lowest importance scores from the dataset, and (c) assigning ranks to remove features according to the sequence of their most recent importance scores. These steps were executed iteratively until the specified number of remaining features rounded to zero.

We evaluated the performance of each classification model according to accuracy, sensitivity, specificity and precision values. The accuracy values were calculated using Eq. (1).

$$\text{Accuracy} = \frac{TP+TN}{(TP+FP+TN+FN)} \quad (1)$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative based on a confusion matrix. We used sensitivity, specificity and precision values to support the accuracy values. The sensitivity and specificity methods are described in Eq. (2) and Eq. (3), respectively.

$$\text{Sensitivity} = \frac{TP}{(TP+FP)} \quad (2)$$

where TP is true positive, TN is true negative and FP is false positive based on a confusion matrix.

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

where TN is true negative and FP is false positive based on a confusion matrix.

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (4)$$

where TP is true positive and FP is false positive based on a confusion matrix.

IV. Results

The best model was based on the highest accuracy value supported by sensitivity, specificity and precision values. Table II shows the results for predictive accuracy values for each diagnostic instrument and model.

Overall, results show that all the models can be used for classifying the FH dataset. However, in overall performance, the decision tree model produced the highest accuracy value. The model recorded an impressive average accuracy value of 99.72% compared with the other models (random forest, 99.54%, and gradient boosting, 99.52%). Of note, the accuracy values obtained by the decision tree model for the DLCC and JFHMC diagnostic instruments were the main contributors to its overall performance. The model obtained perfect accuracy values of 100% for the DLCC diagnostic instrument, in which it outperformed the other models because of some advantages of the decision tree model, such as splitting criteria and the pruning method. The multi-way splitting tree of the decision tree model was advantageous when dealing with multi-classification involving more than two classes.

Table III shows the results for sensitivity, specificity and precision values for each model across the four diagnostic instruments. Overall, results show that all the models can be used to classify FH patients correctly according to the DLCC, US MEDPED and JFHMC diagnostic tools. However, for SB diagnostic criteria, the models encountered a problem caused by two factors: (1) multi-classification involving three classes and (2) high similarity of data.

According to the sensitivity results for the DLCC, the decision tree model demonstrated the perfect value (100%). The gradient boosting model was fairly close, with a value of 75%, while the random forest model was rated 43.75%. For the JFHMC, all the models demonstrated the perfect value (100%) for sensitivity. For the US MEDPED, the random forest model achieved the highest sensitivity value with 99.81% compared with the gradient boosting model (99.48%) and random forest model (99.55%).

TABLE II. Classification of Accuracy Values for Machine Learning Models Across the Four Diagnostic Instruments

Accuracy (%)			
Diagnostic instrument	Decision tree	Random forest	Gradient boosting
DLCC	100.00	99.36	99.49
SB	99.75	99.81	99.74
JFHMC	100.00	99.94	100.00
US MEDPED	99.11	99.05	98.86
Average	99.72	99.54	99.52

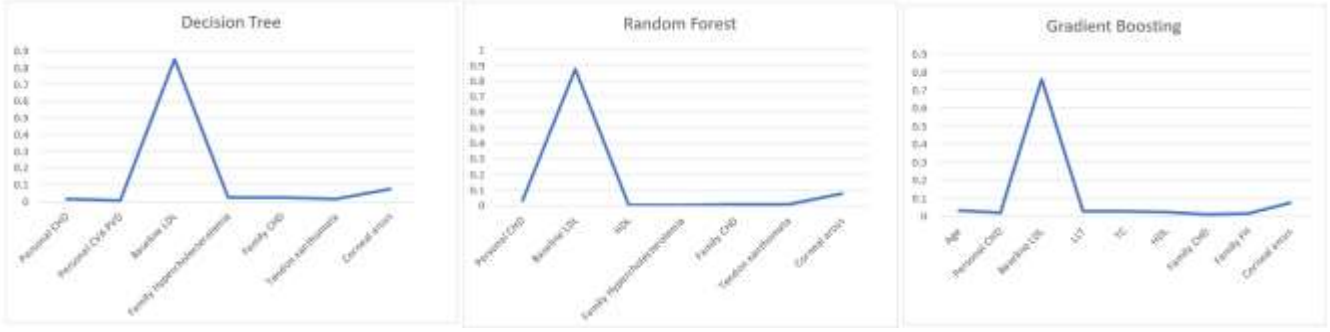
SB: Simon Broome diagnostic criteria; DLCC: Dutch Lipid Clinic Criteria; JFHMC: Japanese FH Management Criteria; US MEDPED: US Make Early Diagnosis to Prevent Early Deaths.

TABLE III. Classification of Sensitivity, Specificity and Precision Values for Machine Learning Models Across four well-known Diagnostic Instruments

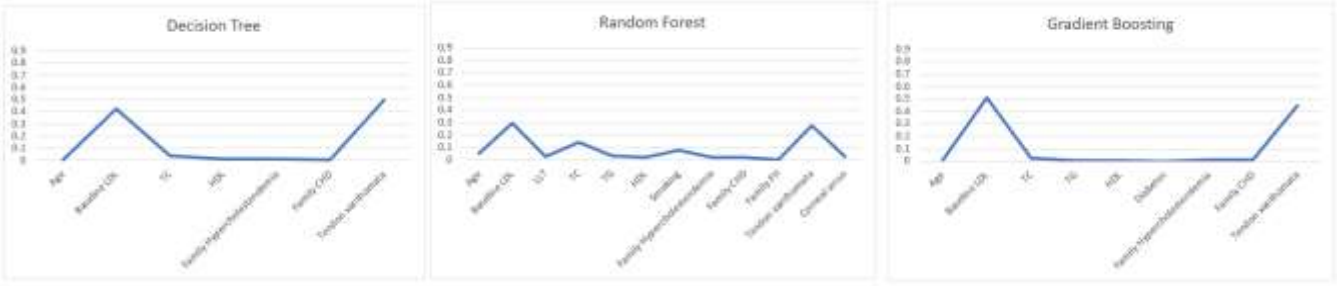
Machine learning model	Accuracy	Sensitivity	Specificity	Precision	No. of features
DLCC					
Random forest	99.36%	43.75%	99.94%	87.50%	7
Gradient boosting	99.49%	75.00%	99.74%	75.00%	9
Decision tree	**100%**	**100%**	**100%**	**100%**	**7**
SB					
Random forest	99.81%	25.00%	100%	100%	12
Gradient boosting	99.74%	100%	99.74%	50.00%	9
Decision tree	99.75%	0%	100%	0%	7
JFHMC					
Random forest	99.94%	100%	98.63%	99.93%	7
Gradient boosting	100%	100%	100%	100%	4
Decision tree	**100%**	**100%**	**100%**	**100%**	**4**
US MEDPED					
Random forest	99.05%	99.81%	65.71%	99.23%	7
Gradient boosting	98.86%	99.48%	71.43%	99.35%	9
Decision tree	**99.11%**	**99.55%**	**80.00%**	**99.55%**	**10**

Specificity values ranged from 65% for random forest to 80% for decision tree for the US MEDPED. Decision tree had perfect specificity values (100%) for the DLCC, SB and JFHMC, and gradient boosting had perfect specificity for the JFHMC. A precision value of 100% was obtained by decision tree for the DLCC and JFHMC, random forest for the SB, and gradient boosting for the JFHMC.

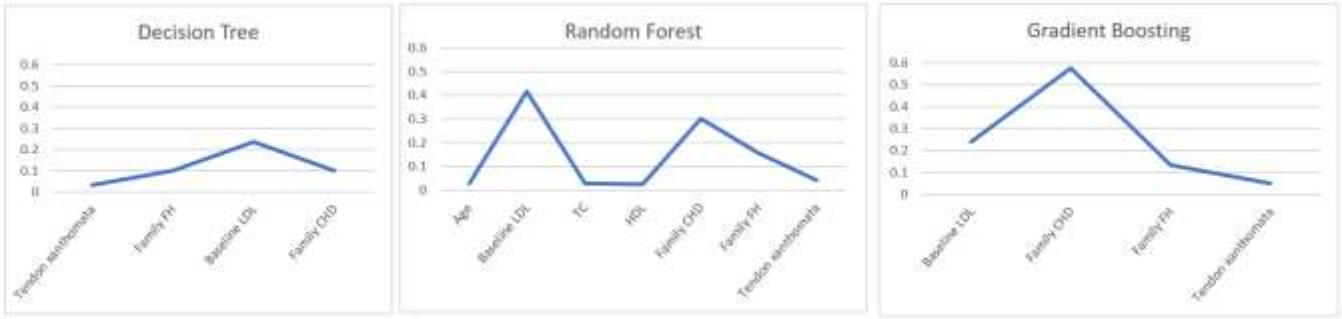
Based on accuracy, sensitivity, specificity and precision values, decision tree is the best model for classifying the FH dataset according to the diagnostic criteria of the DLCC, JFHMC and US MEDPED instruments. For further verification, Fig. 1 shows the clinical feature ranking by feature importance using RFE for the four diagnostic tools (DLCC, SB, JFHMC and US MEDPED) across the three classification models. Each classification model was run on RFE, which was initiated with one clinical feature and increased the number of clinical features until it reached the maximum number. The best model was mainly based on the highest accuracy value and the minimum number of clinical features for the specific tree-based model.



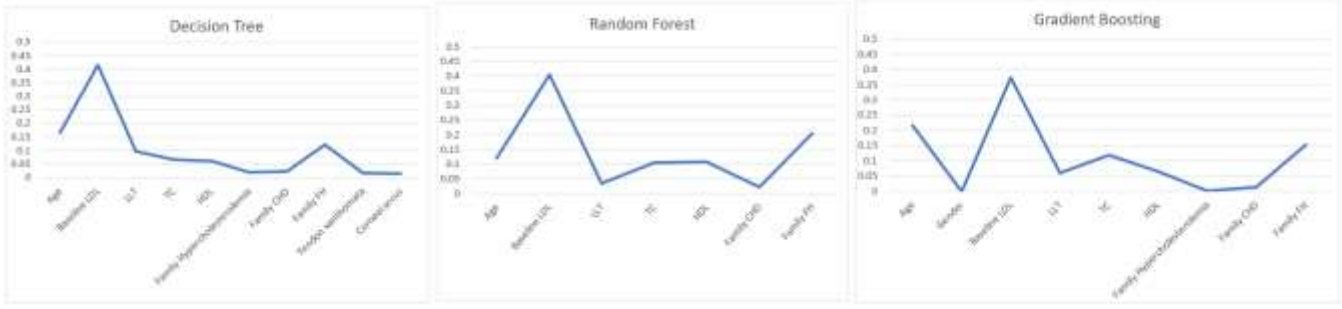
(a) DLCC.



(b) SB.



(c) JFHMC.



(d) US MEDPED.

Fig. 1. Feature Ranking by Feature Importance for Four Diagnostic Instruments across Three Classification Models.

In Fig. 1(a), the DLCC shows perfect accuracy (100%) reached by the decision tree classifier with seven clinical features (history of CHD, history of CVA or PVD, family history of hypercholesterolaemia, family history of CHD, presence of tendon xanthomata, and presence of corneal arcus, LDL-C level) included in the model. Fig. 1(b) shows that the maximum accuracy for the SB reached by the random forest classifier is 99.81%, with 12 clinical features (age, total cholesterol, triglycerides cholesterol level, LDL-C level, lipid-lowering therapy, smoking habit, HDL-C level, family history of hypercholesterolaemia, family history of CHD, family history of FH, presence of tendon xanthomata, and presence of corneal arcus) included in the model.

In Fig. 1(c), the JFHMC shows the maximum accuracy reached by the decision tree classifier and gradient boosting is 100%, with four clinical features (family history of CHD, family history of FH, LDL-C level, presence of tendon xanthomata) included in the model. In Fig. 1(d), the US MEDPED shows the maximum accuracy reached by the random forest classifier is 99.11%, with seven features (age, lipid-lowering therapy, total cholesterol, HDL-C level, family history of CHD, family history of FH, LDL-C level) included in the model. Overall results show that the decision tree classifier (Fig. 1) outperformed the other classifiers in terms of accuracy and minimum numbers of clinical features selected.

In terms of number of features in the dataset, the random forest classifier showed the most selected features (12) for SB criteria compared with the other models, which mostly had seven features selected across the four diagnostic instruments. An increase in features is indicative of a longer period taken for the model to process. Therefore, fewer features are preferable for significant improvement of accuracy performance, comprising the strongest features identified by RFE. The three classification models with RFE selected six common features: family history of CHD, LDL-C level, presence of tendon xanthomata and/or corneal arcus, family hypercholesterolaemia, and family history of FH. Overall, we found that the decision tree classifier is the best model for classification as it demonstrated the highest accuracy and selected the minimum number of features among the classification models. Based on our results, the best diagnostic instrument is the one that includes the maximum number of the six relevant features that can help to accurately classify FH patients. Table IV shows the presence or absence of the selected six features across each diagnostic instrument.

TABLE IV. PRESENCE OF SELECTED FEATURES IN EACH DIAGNOSTIC INSTRUMENT

Selected features	SB	DLCC	US MEDPED	JFHMC
Family history of CHD	√	√	X	√
LDL-C level	√	√	X	√
Family history of hypercholesterolaemia	X	√	X	X
Family history of FH	X	X	√	√
Presence of tendon xanthomata	√	√	X	√
Presence of corneal arcus	X	√	X	X

From Table IV, the DLCC instrument includes five of the six selected features (except family history of FH), and the JFHMC instrument includes four of the features (except corneal arcus and family history of hypercholesterolaemia). None of the features were present in the US MEDPED criteria, except family history of FH. This indicates that the DLCC instrument is the most suitable for Malaysian patients on the basis of the relevant features selected by the best classification model.

V. DISCUSSION

This study is the first to report on detection of FH by applying machine learning models (random forest, gradient boosting and decision tree) with RFE to over 5000 primary health care records that contain a clinically diagnosis of FH according to four well-known diagnostic instruments (DLCC, SB, JFHMC and US MEDPED). Machine learning models provide an additional effective way of screening patients and do not replace the clinical evaluation using diagnostic criteria.

In our study, results showed that the three machine learning models had similar high predictive accuracy in classifying FH patients (accuracy > 99.00%). This is consistent with prior findings using a random forest algorithm in health data [19] and other prior findings using random forest, gradient boosting, deep learning and ensemble learning algorithms in primary care data [28]. The decision tree model outperformed the other machine learning models, with the highest accuracy to determine the likelihood of FH.

Despite the similar accuracy, this study found minimal differences for other performance values between machine learning models. Our analysis highlights specificity values were consistently high across all machine learning models for DLCC, SB and JFHMC that indicate the proportion of patients without actual FH were correctly classified. However, results for sensitivity and precision values varied between machine learning models. For example, random forest model for DLCC identified small proportion of patients with actual FH due to the low sensitivity value (43.75%), but the model would be efficient in having a higher detection rate of FH (high precision value 87.5%).

This study further highlights variations in the selected clinical features identified by the different machine learning models used. For example, decision tree for the DLCC identified seven clinical features (history of CHD, history of CVA or PVD, LDL-C level, family history of hypercholesterolaemia, family history of CHD, presence of tendon xanthomata, presence of and corneal arcus), which is in line with the SB and DLCC diagnostic criteria to systematically identify those who are likely to have FH. Gradient boosting and decision tree for the JFHMC identified four clinical features (family history of CHD, family history of FH, LDL-C level and presence of tendon xanthomata), and random forest for the US MEDPED identified seven features (age, lipid-lowering therapy, total cholesterol, HDL-C level, family history of CHD, family history of FH and LDL-C level). Taken together, these results suggest six relevant clinical features across four diagnostic instruments that can predict FH in Malaysian population: family history of CHD,

LDL-C level, presence of tendon xanthomata, presence of corneal arcus, family history of hypercholesterolaemia and family history of FH.

The findings of this study have important implications for developing FH diagnostic criterion specific for Malaysian population. Our study suggest that machine learning models allow the identification of novel predictive features for detecting FH in Malaysian population. For instance, five out of the six relevant features are well-established criteria in the DLCC diagnostic instrument [8] which previous studies on FH in Malaysia applied DLCC as the main reference diagnostic criteria and it is widely recommended globally [4], [15]. Future studies, which take these novel predictive features into account, will be undertaken.

This study recommends several strengths. We evaluate a range of different tree-based machine learning approach with Recursive Feature Elimination method for detection of FH in Malaysian population. We used the largest number of primary health care records that contain a diagnosis of FH according to four well-known diagnostic instruments (DLCC, SB, JFHMC and US MEDPED), compared to other previous FH studies conducted in Malaysia. This study also assessed the clinical features of the abovementioned four diagnostic instruments to identify the novel predictive features that are useful in the diagnosis of FH in Malaysian patients, using machine learning models.

Compared with relying on multiple FH diagnostic criteria, as being practised currently, the use of machine learning techniques allows healthcare providers to conduct early testing for the presence of FH in patients. It simplifies the current labour-intensive and time-consuming process in the diagnosis of FH in Malaysia by streamlining and focusing on important features of diagnostic criteria that are relevant and pertinent to the procedure. The machine learning techniques offer major opportunities to increase diagnosis of FH and to prevent pCHD and early death.

However, we acknowledge several study limitations, which are common in other research using healthcare data. The limitations include the potential for information bias due to missing data. Missing data may introduce bias in the performance of prediction models. However, we used mean or mode imputation methods to replace quantitative missing values with the mean of the attribute or qualitative missing values with the mode of the attribute to overcome these effects. Another potential information bias in the dataset is that some patients could potentially be misclassified because of inaccurate reporting of family history. Future studies should validate and replicate our machine learning models with the implementation of RFE in other populations to confirm the findings of this study. Further, additional evaluation of the feasibility of machine learning applications in clinical practice is required to support the computational capacity of healthcare systems.

VI. Conclusion

The decision tree classifier performs best in identifying the relevant features for the DLCC, SB, US MEDPED and JFHMC. Family history of CHD, family history of hypercholesterolemia, family history of FH, LDL-C level, presence of tendon xanthomata and presence of corneal arcus, are the relevant features for diagnosing FH among DLCC, SB, US MEDPED and JFHMC diagnostic criteria that give the highest accuracy in the classification model. Future research should include these six relevant features, which have potential to be developed into an efficient FH prediction model to assist clinicians in identifying FH patients.

Overall, this study highly suggests that machine learning algorithms may help the diagnosis of FH in classifying FH among patients, leading to effective identification of high-risk patients with FH. The three classifiers used in this study embody the most important features in predicting patients with FH. These features also contribute to unify the population-based diagnostic criteria, constituting a first step towards development of more relevant, locally adjusted and tested Malaysian FH diagnostic criteria for early diagnosis of FH in the local community. This is also particularly important in family contact tracing for indexed cases. Efficient, locally adjusted diagnostic criteria will improve early and overall detection, hence anticipating early treatment and prevention of pCHD.

Acknowledgment

The authors would like to thank the Universiti Teknologi Mara for their financial support of this project under FRGS Grant No. 600-IRMI/FRGS 5/3 (212/2019).

References

[1] T. Phuong Kim, L. Thuan Duc, and H. Le Thuy Ai, “The Major Molecular Causes of Familial Hypercholesterolemia,” Asian J. Pharm. Res. Heal. Care, vol. 10, no. 2, pp. 60–68, Aug. 2018, doi: 10.18311/ajprhc/2018/20031.

[2] A. Wiegman, S. S. Gidding, G. F. Watts, M. J. Chapman, H. N. Ginsberg, M. Cuchel, L. Ose, M. Averna, C. Boileau, J. Borén, E. Bruckert, A. L. Catapano, J. C. Defesche, O. S. Descamps, R. A. Hegele, G. K. Hovingh, S. E. Humphries, P. T. Kovanen, J. A. Kuivenhoven, L. Masana, B. G. Nordestgaard, P. Pajukanta, K. G. Parhofer, F. J. Raal, K. K. Ray, R. D. Santos, A. F. H. Stalenhoef, E. Steinhagen- Thiessen, E. S. Stroes, M.-R. Taskinen, A. Tybjærg-Hansen, and O. Wiklund, “Familial hypercholesterolaemia in children and adolescents: gaining decades of life by optimizing detection and treatment,” Eur. Heart J., vol. 36, no. 36, pp. 2425–2437, Sep. 2015, doi: 10.1093/eurheartj/ehv157.

[3] B. G. Nordestgaard, M. J. Chapman, S. E. Humphries, H. N. Ginsberg, L. Masana, O. S. Descamps, O. Wiklund, R. A. Hegele, F. J. Raal, J. C. Defesche, A. Wiegman, R. D. Santos, G. F. Watts, K. G. Parhofer, G. K. Hovingh, P. T. Kovanen, C. Boileau, M. Averna, J. Borén, E. Bruckert, A. L. Catapano, J. A. Kuivenhoven, P. Pajukanta, K. Ray, A. F. H. Stalenhoef, E. Stroes, M.-R. Taskinen, A. Tybjærg-Hansen, and European Atherosclerosis Society Consensus Panel, “Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease: consensus statement of the European Atherosclerosis Society.,” Eur. Heart J., 2013, doi: 10.1093/eurheartj/eht273.

[4] Y.-A. Chua, A. Z. Razman, A. S. Ramli, N. A. Mohd Kasim, and H. M. Nawawi, “Familial Hypercholesterolaemia in the Malaysian Community: Prevalence, Under-Detection and Under-Treatment,” J. Atheroscler. Thromb., 2021, doi: 10.5551/jat.57026.

[5] S. A. Nazli, Y. A. Chua, N. A. Mohd Kasim, Z. Ismail, A. B. Md Radzi, K. S. Ibrahim, S. Kasim, A. Rosman, and H. M. Nawawi, “Familial Hypercholesterolaemia among Patients with Coronary Angiogram-Proven Premature Coronary Artery Disease,” Strait Circ. J., vol. 1, no. 2, p. 44, 2019, doi: https://doi.org/10.6907/SCJ.201909/SP_1(2).0037.

[6] J. C. Defesche, "Defining the challenges of FH Screening for familial hypercholesterolemia," J. Clin. Lipidol., vol. 4, no. 5, pp. 338–341, Sep. 2010, doi: 10.1016/j.jacl.2010.08.022.

[7] K. E. Heath, S. E. Humphries, H. Middleton-Price, and M. Boxer, "A molecular genetic service for diagnosing individuals with familial hypercholesterolaemia (FH) in the United Kingdom," Eur. J. Hum. Genet., vol. 9, no. 4, pp. 244–252, Apr. 2001, doi: 10.1038/sj.ejhg.5200633.

[8] S. W. Fouchier, J. C. Defesche, M. A. Umans-Eckenhausen, and J. J. Kastelein, "The molecular basis of familial hypercholesterolemia in The Netherlands," Hum. Genet., vol. 109, no. 6, pp. 602–615, Dec. 2001, doi: 10.1007/s00439-001-0628-8.

[9] R. R. Williams, S. C. Hunt, M. C. Schumacher, R. A. Hegele, M. F. Leppert, E. H. Ludwig, and P. N. Hopkins, "Diagnosing heterozygous familial hypercholesterolemia using new practical criteria validated by molecular genetics," Am. J. Cardiol., vol. 72, no. 2, pp. 171–176, Jul. 1993, doi: 10.1016/0002-9149(93)90155-6.

[10] M. Harada-Shiba, H. Arai, S. Oikawa, T. Ohta, T. Okada, T. Okamura, A. Nohara, H. Bujo, K. Yokote, A. Wakatsuki, S. Ishibashi, and S. Yamashita, "Guidelines for the management of familial hypercholesterolemia," J. Atheroscler. Thromb., 2012, doi: 10.5551/jat.14621.

[11] A. Al-Khateeb and H. Al-Talib, "Genetic Researches Among Malaysian Familial Hypercholesterolaemia Population," J. Heal. Transl. Med., vol. 19, no. 2, pp. 1–11, Dec. 2016, doi: 10.22452/jummec.vol19no2.1.

[12] R. Jeyamalar, W. A. Wan Azman, H. Nawawi, G. H. Choo, W. K. Ng, M. A. Rosli, O. Al Fazir, K. Sazzli, M. Oteh, and D. K. L. Quek, "Updates in the management of Dyslipidaemia in the high and very high risk individual for CV risk reduction.," Med. J. Malaysia, vol. 73, no. 3, pp. 154–162, 2018, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29962499.

[13] K. L. Khoo, P. Van Acker, H. Tan, and J. P. Deslypere, "Genetic causes of familial hypercholesterolaemia in a Malaysian population.," Med. J. Malaysia, vol. 55, no. 4, pp. 409–18, Dec. 2000, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11221151.

[14] S.-H. Lye, J. K. Chahil, P. Bagali, L. Alex, J. Vadivelu, W. A. W. Ahmad, S.-P. Chan, M.-K. Thong, S. M. Zain, and R. Mohamed, "Genetic Polymorphisms in LDLR, APOB, PCSK9 and Other Lipid Related Genes Associated with Familial Hypercholesterolemia in Malaysia," PLoS One, vol. 8, no. 4, p. e60729, Apr. 2013, doi: 10.1371/journal.pone.0060729.

[15] S. Abdul-Razak, R. Rahmat, A. Mohd Kasim, T. A. Rahman, S. Muid, N. M. Nasir, Z. Ibrahim, S. Kasim, Z. Ismail, R. Abdul Ghani, A. R. Sanusi, A. Rosman, and H. Nawawi, "Diagnostic performance of various familial hypercholesterolaemia diagnostic criteria compared to Dutch lipid clinic criteria in an Asian population," BMC Cardiovasc. Disord., vol. 17, no. 1, p. 264, Dec. 2017, doi: 10.1186/s12872-017-0694-z.

[16] A. Al-Khateeb, M. K. Zahri, M. S. Mohamed, T. H. Sasongko, S. Ibrahim, Z. Yusof, and B. A. Zilfalil, "Analysis of sequence variations in low-density lipoprotein receptor gene among Malaysian patients with familial hypercholesterolemia," BMC Med. Genet., vol. 12, no. 1, p. 40, Dec. 2011, doi: 10.1186/1471-2350-12-40.

[17] A. Khan, J. P. Li, A. U. Haq, I. Memon, S. H. Patel, and S. ud Din, "Emotional-physic analysis using multi-feature hybrid classification," J. Intell. Fuzzy Syst., vol. 40, no. 1, 2021, doi: 10.3233/JIFS-201069.

[18] M. H. Memon, I. Memon, J. P. Li, and Q. A. Arain, "IMRBS: image matching for location determination through a region-based similarity technique for CBIR*," Int. J. Comput. Appl., vol. 41, no. 6, 2019, doi: 10.1080/1206212X.2018.1468643.

[19] K. D. Myers, J. W. Knowles, D. Staszak, M. D. Shapiro, W. Howard, M. Yadava, D. Zuzick, L. Williamson, N. H. Shah, J. M. Banda, J. Leader, W. C. Cromwell, E. Trautman, M. F. Murray, S. J. Baum, S. Myers, S. S. Gidding, K. Wilemon, and D. J. Rader, "Precision screening for familial hypercholesterolaemia: a machine learning study applied to electronic health encounter data," Lancet Digit. Heal., vol. 1, no. 8, pp. e393–e402, Dec. 2019, doi: 10.1016/S2589-7500(19)30150-5.

[20] Z. Obermeyer and E. J. Emanuel, "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine," N. Engl. J. Med., vol. 375, no. 13, pp. 1216–1219, Sep. 2016, doi: 10.1056/NEJMp1606181.

[21] Z. Shi, B. Yuan, D. Zhao, A. W. Taylor, J. Lin, and G. F. Watts, "Familial hypercholesterolemia in China: Prevalence and evidence of underdetection and undertreatment in a community population," Int. J. Cardiol., 2014, doi: 10.1016/j.ijcard.2014.04.165.

[22] M. Benn, G. F. Watts, A. Tybjaerg-Hansen, and B. G. Nordestgaard, "Familial hypercholesterolemia in the Danish general population: Prevalence, coronary artery disease, and cholesterol-lowering medication," J. Clin. Endocrinol. Metab., 2012, doi: 10.1210/jc.2012-1563.

[23] G. F. Watts, S. Gidding, A. S. Wierzbicki, P. P. Toth, R. Alonso, W. V. Brown, E. Bruckert, J. Defesche, K. K. Lin, M. Livingston, P. Mata, K. G. Parhofer, F. J. Raal, R. D. Santos, E. J. G. Sijbrands, W. G. Simpson, D. R. Sullivan, A. V. Susekov, B. Tomlinson, A. Wiegman, S. Yamashita, and J. J. P. Kastelein, "Integrated guidance on the care of familial hypercholesterolaemia from the International FH Foundation.," Int. J. Cardiol., vol. 171, no. 3, pp. 309–25, Feb. 2014, doi: 10.1016/j.ijcard.2013.11.025.

[24] J. M. Banda, A. Sarraju, F. Abbasi, J. Parizo, M. Pariani, H. Ison, E. Briskin, H. Wand, S. Dubois, K. Jung, S. A. Myers, D. J. Rader, J. B. Leader, M. F. Murray, K. D. Myers, K. Wilemon, N. H. Shah, and J. W. Knowles, "Finding missed cases of familial hypercholesterolemia in health systems using machine learning," npj Digit. Med., 2019, doi: 10.1038/s41746-019-0101-5.

[25] S. F. Weng, J. Kai, H. Andrew Neil, S. E. Humphries, and N. Qureshi, "Improving identification of familial hypercholesterolaemia in primary care: Derivation and validation of the familial hypercholesterolaemia case ascertainment tool (FAMCAT)," Atherosclerosis, vol. 238, no. 2, pp. 336–343, Feb. 2015, doi: 10.1016/j.atherosclerosis.2014.12.034.

[26] S. Weng, J. Kai, J. Tranter, J. Leonardi-Bee, and N. Qureshi, "Improving identification and management of familial hypercholesterolaemia in primary care: Pre- and post-intervention study," Atherosclerosis, vol. 274, pp. 54–60, Jul. 2018, doi: 10.1016/j.atherosclerosis.2018.04.037.

[27] A. Pina, S. Helgadottir, R. M. Mancina, C. Pavanello, C. Pirazzi, T. Montalcini, R. Henriques, L. Calabresi, O. Wiklund, M. P. Macedo, L. Valenti, G. Volpe, and S. Romeo, "Virtual genetic diagnosis for familial hypercholesterolemia powered by machine learning," Eur. J. Prev. Cardiol., p. 204748731989895, 2020, doi: 10.1177/2047487319898951.

[28] R. K. Akyea, N. Qureshi, J. Kai, and S. F. Weng, "Performance and clinical utility of supervised machine-learning approaches in detecting familial hypercholesterolaemia in primary care," npj Digit. Med., vol. 3, no. 1, p. 142, Dec. 2020, doi: 10.1038/s41746-020-00349-5.

Development of Star-Schema Model for Lecturer Performance in Research Activities

M. Miftakul Amin[1], Adi Sutrisman[2], Yevi Dwitayanti[3]
Department of Computer Engineering, Politeknik Negeri Sriwijaya, Palembang, Indonesia[1, 2]
Department of Accounting, Politeknik Negeri Sriwijaya, Palembang, Indonesia[3]

***Abstract*—In this study, the researchers developed a multidimensional data model to investigate the activities of lecturers in universities in carrying out research activities as part of the Three Pillars of Higher Education. Information about lecturers' research activities has been managed using spreadsheet (excel) documents. Thus, access and analysis of the information were limited. Data warehouse development was carried out through several stages, namely requirement analysis, data source analysis, multidimensional modeling, ETL process, and reporting. The information generated in this data warehouse (DW) can be used as one of the business intelligence (BI) models in universities. In this study, the star-schema model was used in designing dimension tables and fact tables to facilitate and speed up the query process. The information generated in this study can be used by management in universities to make decisions and strategic planning. The results of this study can also be used as one of the important information in the preparation of institutional accreditation data and study program accreditation.**

Keywords*—Data warehouse (DW); star-schema; multidimensional data; business intelligence (BI)

I. Introduction

Lecturers as one of the intellectual assets in higher education have one of the activities that are part of the Three Pillars of Higher Education, namely carrying out research and community service. Every year there is a lot of funding to carry out research and community service activities. This funding source comes from the ministries of education, culture, research, and technology, as well as from local governments, and internal universities.

This research and community service information is used as one of the assessment criteria in the preparation of study program accreditation instruments and university accreditation. Currently, at Politeknik Negeri Sriwijaya, research data and community service are documented at the Research and Community Service Center (Pusat Penelitian dan Pengabdian kepada Masyarakat abbreviated as P3M) Politeknik Negeri Sriwijaya, in the form of spreadsheet (excel) file data, without being further managed into useful information for upper management in the decision-making process.

Referring to the research and community service handbook published by the Ministry of Education, Culture, Research, and Technology in 2020, there are 13 types of research funding schemes, and 10 community service funding schemes [8]. With this large funding opportunity, it is appropriate for universities to carry out adequate data management in managing research information and community service.

Currently, universities have quite large data and are spread across several sub-units within it. These data will continue to grow over time. Upper management needs tools to generate information and to assist the decision-making process [1]. Data Warehouse (DW) technology can be used to extract important information from data scattered across information system management units into centralized integrated storage, to provide management information needs, view data from various perspectives, detailed information, and historical data.

A Data Warehouse is an integrated repository of information, making it possible to query and analyze the data. The basic idea is to carry out the process of extracting, filtering, and integrating relevant data. The development of data warehouses in universities is still rarely carried out, even though universities are very rich in the information contained in them. One of the factors is that business transactions in universities are non-commercial [2]. This is also reinforced by Yu [3] which states that the implementation of a data warehouse is underestimated in a university environment. Because many people think that universities are non-profits organizations. Whereas with the increasing number of study programs, lecturers, employees, and students in the future, a university must consider the integration of a data warehouse-based decision support system to make better decisions.

In this study, the researchers developed a multidimensional data model to investigate various perspectives and points of view related to research activities carried out by lecturers in universities. This research is important to do in order to provide relevant information for top management at the Politeknik Negeri Sriwijaya in decision making and strategic management.

This research is organized as follows. Section 2 describes some of the theoretical concepts that underlie this research, such as the concept of a data warehouse and multidimensional data modeling. Section 3 describes the methodology and stages in model development. Section 4 describes the experimental setup, research results, and discussion. Section 5 contains conclusions.

II. Literature Review

A. Implementation of Data Warehouse in Higher Education

Based on research conducted by Bassil [4] it is mentioned that data warehouse development in universities can be implemented by transforming operational databases into data warehouses that can be used in the decision-making process and perform data analysis, prediction, and forecasting. The

development of this data warehouse can be done through several stages, namely data extraction, data cleansing, data transforming, and data indexing and loading. An operational database is a regular database that is intended to run a business on a database and support daily transactions [5].

Bogdanova [6] developed a model known as CaMeLOT as an Educational Framework for Conceptual Data Modelling. By using Bloom's taxonomy modeling is carried out as a part of software engineering. This study proposes that the model can be used in the preparation of a curriculum that can adapt the use of technology in the implementation of learning in universities.

A previous study conducted by Santoso [1] mentions that the development of data warehouses in universities can be categorized into 2 groups, namely traditional and modern. A modern data warehouse is characterized by the use of big data technology in its implementation. This data can be taken from various information spreads on the internet such as social media, sensors, blogs, videos, and audio as data sources. Meanwhile, the traditional data warehouse sources are only limited to the transaction and operational data that exist in the university environment.

Yulianto [7] has developed a multidimensional data warehouse model for the Integrated Academic Fee (IAF) in universities. This research follows 4 stages in the development of Business Intelligence (BI), namely preparation, integration, analysis, and visualization. The results of the study can be used as part of the admission DSS.

Asroni et.al [16] investigated the implementation of data warehouses in universities to manage alumni data through data tracer studies. The output of this research is to produce a reporting system using the SQL Server Analysis Service (SSAS) tool to view various dimensions such as alumni profile, department, faculties, and salaries. This information is presented in the form of graphs, tables, and diagrams.

B. Data Warehouse Concept

A data warehouse is a collection of integrated databases which is subject-oriented and designed to support decision-making functions [9]. The data flow in the data warehouse comes from the operational level which is transformed into the data warehouse [10]. According to Thakur [13] mentions that data warehouses have data needs that change from time to time. Thus, this will cause dynamic changes in data storage. A data warehouse is a database designed to perform analysis of decision making, where data and information are generated from the ETL (Extract, transform, Load) [14].

Seen from the infrastructure aspect, the data warehouse consists of several technical components that can be grouped into two categories, namely operation infrastructure and physical infrastructure, such as server hardware, operating system, network software, database software, LAN, WAN, vendor resources, persons, procedure and training [15].

C. Multidimensional Data Modeling

A multidimensional data can be implemented into a star-schema model and can use join operations to relate the tables that exist in it during the query process. The star-schema model consists of a fact table in the middle, then surrounded by a dimension table [11]. A good multidimensional data in data warehouse development should have a simple database structure. It aims to speed up the query process which will be carried out in the analytical stage. The fact table contains facts or measures that are used as business parameters, while the dimension table contains descriptions for query processing.

Several advantages are obtained when implementing a star-schema in data warehouse development, including simplifying query, simplifying reporting logic, improving query, performance, and accelerating data aggregation [17].

In the star-schema model, the fact table and dimension table are connected by a key known as a surrogate key which acts as the primary key of the dimension table and becomes a foreign key in the fact table [18]. This relation occurs logically and can be used to perform the JOIN process in the Query command.

III. METHOD

This study involved data on the activities of lecturers in research at the Politeknik Negeri Sriwijaya within a span of 3 years, from 2018 to 2020. The activity data comes from research activities organized by the Ministry of Education, Culture, Research, and Technology, and activities organized by the university internally.

Meanwhile, the software used in building the system in this research is the framework Codeigniter, with MariaDB/MySQL as the database engine, PHP as the scripting engine, and Apache as the webserver.

Fig. 1 is a stage in the development of a multidimensional database. The stages of system development in this study refer to what was conveyed by Zea [12] who built a data warehouse system in several stages as follows:

- Requirement analysis, in this stage some information is formulated which will later be presented in the data warehouse.
- Data source analysis, the data source is taken from the existing database in the university's internal environment.
- Multidimensional modeling, this stage is carried out to formulate a data warehouse design to describe the relationship between fact table data and dimension tables.
- ETL process, this stage is used to carry out the data retrieval process, transformation process, and data storage into the data warehouse.
- Reporting, this stage presents data in the form of summary information and other important information that acts as output in system design.

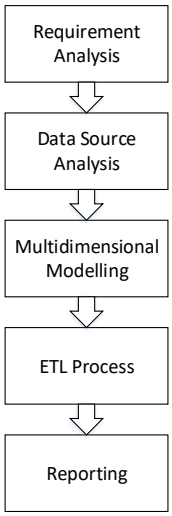


Fig. 1. Step of Design System.

IV. Result and Analysis

A. Requirement Analysis

This study aims to determine for the availability of data and information as shown in Table I. The information presented in the dashboard as an application on the end-user side can be displayed in the form of tables and graphs. A data table is used to display data in tabular form in which there are facilities for sorting, searching, and paging. Meanwhile, the information in the form of graphs is presented to display a summary of the research data that makes it easier to understand the information.

B. Data Source Analysis

Fig. 2 is an architectural model developed in this study. The data source comes from data sources originating from the academic information system (Sistem Informasi Akademik, abbreviated as SISAK) to retrieve lecturer and department data information. Meanwhile, data on lecturers' research activities come from research information system data (SIMP3M) to obtain data on research schemes, research contracts, and research funding.

Furthermore, at the data warehouse stage, the staging process is carried out to store the extracted data from the data source which has been modified successively to finally be loaded into the multidimensional database in the data warehouse (DW). Then, through the web browser application, there is a dashboard to process reporting, query, visualization, and analysis of the previously formed multidimensional data.

Data derived from the data source can be taken from tables, or with data in the form of csv files. This csv file type is used to facilitate the ETL process which will later be implemented in the system.

C. Multidimensional Modeling

Researchers logically designed a multidimensional data model using a star schema as can be seen in Fig. 3. This star-schema model is used logically to facilitate the query process later. This star-schema consists of a number of dimension tables and fact tables.

In Fig. 3, there is surrogate_key (SK) in each table, both fact table, and dimension table as a unique record marker in the table and facilitates the data query process. Based on the need for data source analysis that has been defined in the previous stage, there are several dimension tables and fact tables as follows:

1) lecturer_dim: The lecturer_dim table is used to store information on the lecturer who acts as a researcher in research activities.

2) department_dim: The department_dim table is used to store information on departments that are part of research activities and the home base of lecturers or researchers.

3) contract_dim: The contract_dim table is used to store research contract information in every research activity carried out by lecturers.

4) schema_dim: The schema_dim table is used to store research scheme data that can be followed by a lecturer in conducting research activities.

5) date_dim: The date_dim table is used in the system to store date information and hierarchical information such as month name, month number, quarter, and year.

6) funding_fact: The funding_fact table is used to store information on research conducted by lecturers. This table is a fact table that contains the research history.

In each table, both dimension table and fact table, there are effective_date and expiry_date columns to indicate whether a record is active or not used in the system for query and data retrieval processes.

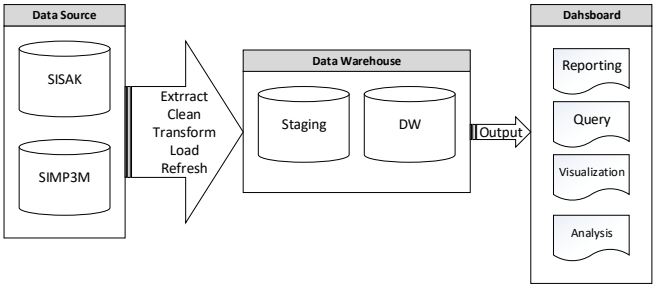


Fig. 2. Architecture of System.

TABLE I. Information Requirement

No.	Requirement		
	Information	***Format***	***Timeframe***
1	All Lecturer/ Researcher	Table	All Period
2	All Research Funding	Table	All Period
3	All Research Contract	Table	All Period
4	All Research Scheme	Table	All Period
5	All Departement	Table	All Period
6	Top 5 Lecturer/ Researcher	Chart	All Period
7	Top 5 Department Research	Chart	All Period
8	Top 5 Research Scheme	Chart	All Period
9	Research By All Department	Chart	All Period
10	Research By Department and Year	Chart	Custome Filter
11	Trend Research Funding	Chart	Custome Filter
12	Funding Amount	Chart	Custome Filter

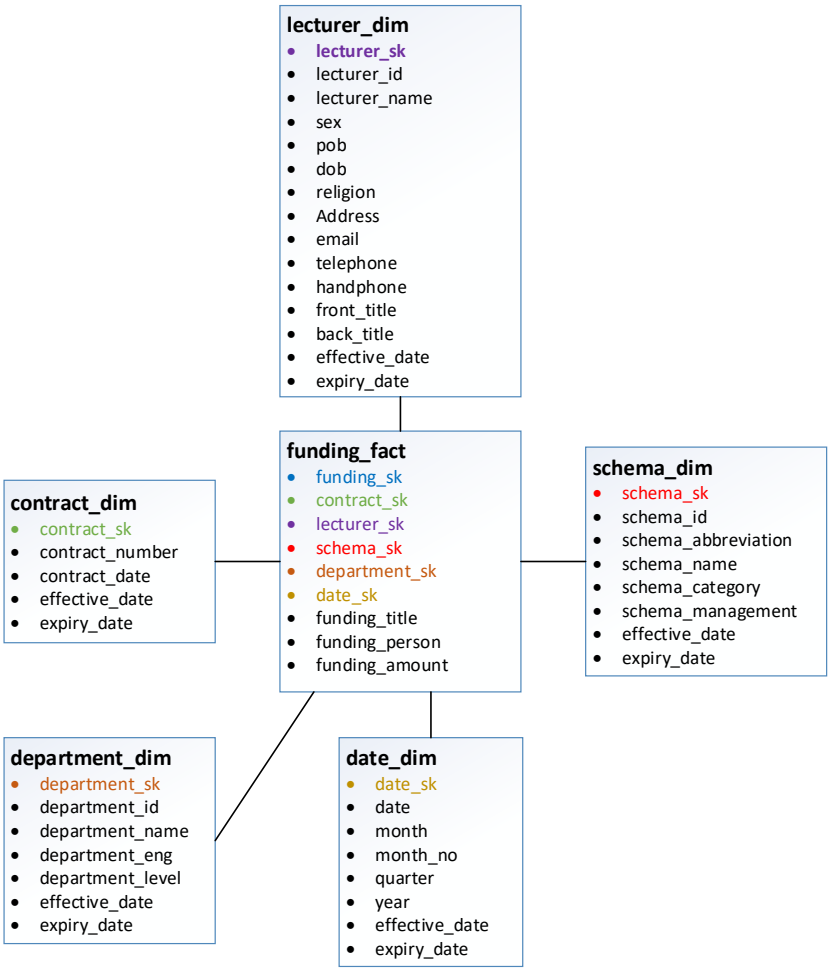


Fig. 3. Star-Schema Model.

D. ETL Process

In the development of this data warehouse, the activity that is quite time-consuming is the ETL (Extraction, Transformation, and Loading). This activity includes the source extraction and performing data population. This activity is to ensure that the data taken from the data source Table I is guaranteed its integrity and validity. Table I is a matrix that describes the ETL (Extract, Transform, Load) process mechanism in the system design. The extract is a step to get data from a data source which is then stored in the data warehouse, while the transform process is a process to prepare data, and load is a process to store data into the data warehouse. In practice, this ETL stage is not a separate stage, but sometimes it is a stage that is an integrated series of processes.

Referring to Table II, there are two mechanisms for the data extraction process from a data source, namely, push and pull. Pull mode is a process to pull data from the data source which is carried out by the data warehouse system. In the design of this system, all data retrieval processes use pull mode. Meanwhile, push mode is a data extraction process carried out by the data source to send data to the data warehouse.

Judging from the data, whether to be sent or withdrawn into the data warehouse, there are two approaches [11], namely, whole part and change data capture (CDC). The whole part is a process to retrieve overall data from data sources when filling out the data warehouse. These data are master data or reference data, such as lecturer data, department data, and schemas which in the timeline review rarely add or update data. Meanwhile, data on contract and funding is processed by the CDC because they are transaction data and there is often be a process of adding and updating data. Pulling data with pull mode is done with the last data change made from the previous pull mode stage.

TABLE II. SOURCE EXTRACTION MATRIX

Data Source	Source Extraction		
	Data Warehouse Table	***Extraction Mode***	***Loading Type***
lecturer	lecturer_dim	Whole, Pull	SCD
department	department_dim	Whole, Pull	SCD
schema	schema_dim	Whole, Pull	SCD
contract	contract_dim	CDC, Pull	unique contract number
funding	funding_fact	CDC, Pull	unique contract number
n/a	date_dim	n/a	pre-population

The process of loading data for the lecturer_dim, department_dim, and schema_dim tables is carried out with SCD (Slowly Changing Dimension) type 1 by replacing the data in certain columns where the data is updated. Meanwhile, the contract_dim and funding_fact tables are carried out with the CDC when the data update process occurs based on the unique contract number. Meanwhile, the date_dim table is not contained in the data source but is generated in the data warehouse system with pre-loading population mode by retrieving and extracting the date parameters during the ETL process.

As an additional feature in the development of this project, part of the data is taken from data sources that come from RESTful web services. This allows two heterogeneous data sources to communicate through a web services interface. The data taken from these web services is used to display summary information in the form of graphic visualizations, making it easier to understand the information presented.

E. Reporting

At this reporting stage, a dashboard is provided in the form of a web application that can present data in the form of tables and graphs. The information presented in the dashboard is generated from the existing SQL commands in the system. The use of star-schema was chosen in this study to facilitate and speed up the query process.

Fig. 4 provides information on the dashboard display that is run on the end-user, side, at the top, there are main menu options consisting of the dashboard, academic, research, human resource, and student. Meanwhile, at the bottom, there is summary information about the number of lecturers, the number of students, the number of departments, the number of research schemes, the number of research funding, and the number of contract funding. In the next section, there is ranking information in the form of top 5 for department, researcher, and funding scheme from lecturer research activities carried out. The following is an example of a query command that is used to produce output in the form of a top 5 research (lecturer) graph as shown in Fig. 4.

```
elect `b`.`lecturer_name` AS
`lecturer_name`,`c`.`department_level` AS
`department_level`,`c`.`department_eng` AS
`department_eng`,count(0) AS `qtyresearch` from
(((((`funding_fact` `a` join `lecturer_dim` `b`) join
`department_dim` `c`) join `schema_dim` `d`) join
`contract_dim` `e`) join `date_dim` `f`) where `a`.`lecturer_sk`
= `b`.`lecturer_sk` and `a`.`department_sk` =
`c`.`department_sk` and `a`.`schema_sk` = `d`.`schema_sk`
and `a`.`contract_sk` = `e`.`contract_sk` and `a`.`date_sk` =
`f`.`date_sk` group by `b`.`lecturer_name` order by count(0)
desc limit 5
```

Meanwhile, in Fig. 5 there is information in the form of graphs that describe the distribution of research activities of lecturers spread across several departments within Politeknik Negeri Sriwijaya. This information is generated in the all-period time range according to the number of data records in the database. Information is presented in a bar chart where the x-axis is the department's data, while the y-axis is the number of studies conducted by lecturers as researchers in the department.

F. Performance Evaluation

To see the extent of the performance generated from the developed model, a performance evaluation is carried out by looking at the payloads (bytes) and response time (milli second) of the system. By using network monitoring, a snippet of data is obtained from a web page which is displayed as shown in Table III. It can be seen in Table III that regardless of the number of payloads or data transferred from the server to the client, judging from the response time, the execution time is not much different. As discussed in the ETL Process section, the data source used in this model comes from an internal domain and a RESTful Web Services from a different web domain (cross domain).

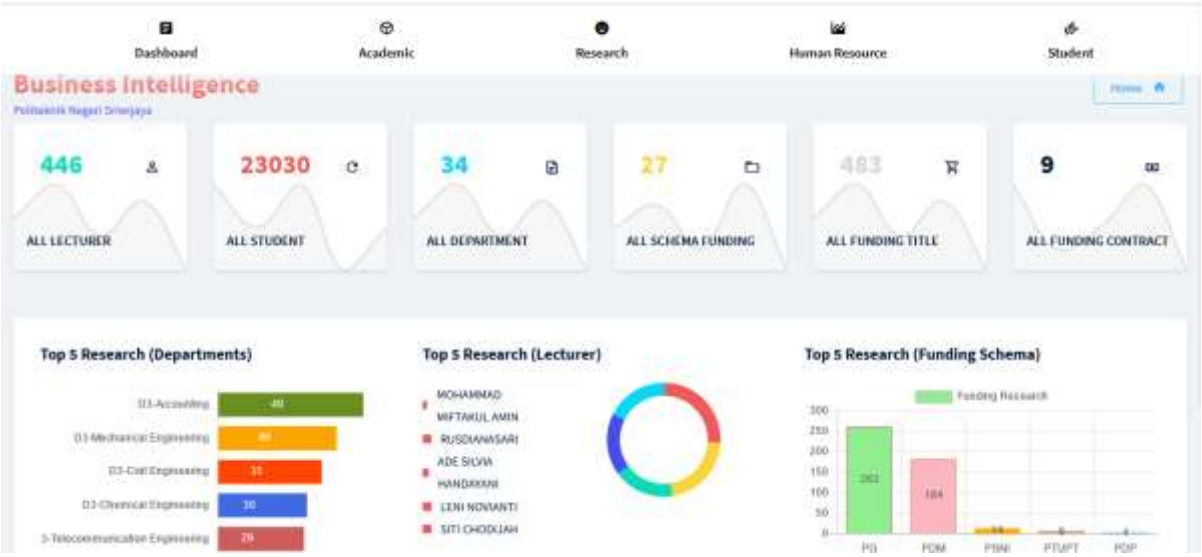


Fig. 4. Dashboard of System.

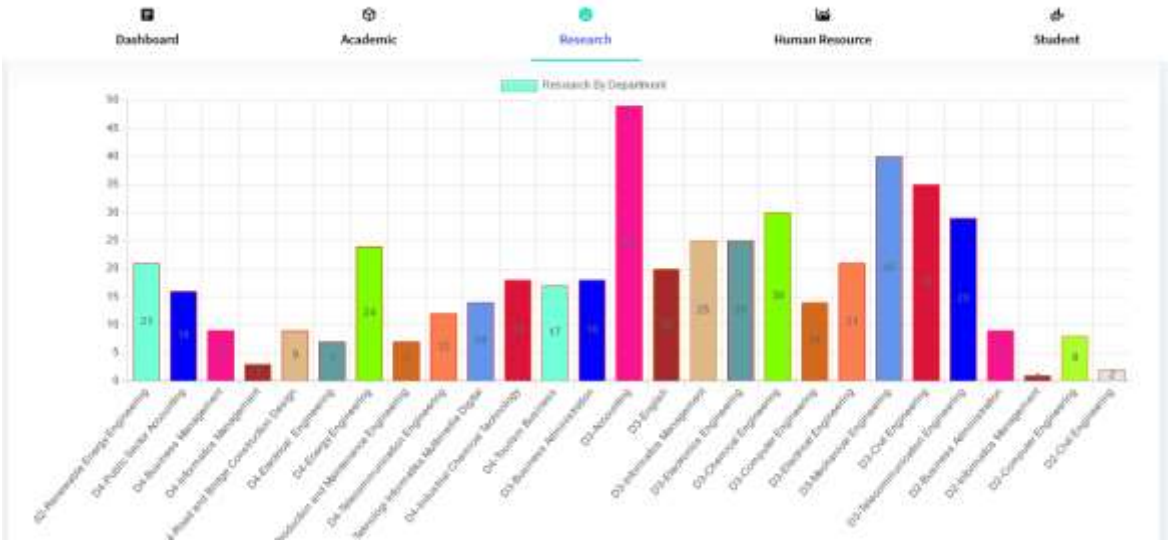


Fig. 5. Information of Research by all Departments.

TABLE III. PERFORMANCE OF APPLICATION MODEL

Services Number	Evaluation Performance		
	Domain	*Payload (Bytes)*	*Time (ms)*
1.	Cross Domain	1670	62
2.	Cross Domain	972	46
3.	Internal Domain	983	61
4.	Internal Domain	814	73
5.	Internal Domain	645	71
6.	Internal Domain	710	61
7.	Internal Domain	743	65
8.	Cross Domain	652	61
9.	Cross Domain	654	60
10.	Cross Domain	651	63
11.	Internal Domain	527	63
12.	Internal Domain	528	64
13.	Internal Domain	526	62
14.	Internal Domain	573	60
15.	Cross Domain	748	54
16.	Cross Domain	719	63
17.	Cross Domain	711	63
18.	Cross Domain	1060	62

By looking at Fig. 6 provides an overview of the distribution of payloads and response times executed by web browsers. By looking at the fast and relatively stable response time for any amount of payloads, the model of this system development can be utilized in developing business intelligence applications by using star schemes and RESTful web services in the application architecture.

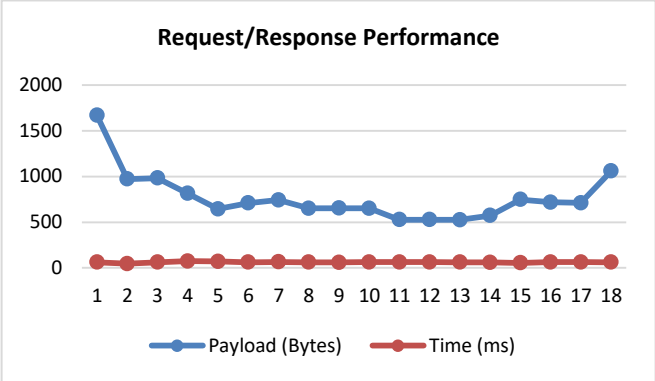


Fig. 6. Payloads and Response Time of System.

V. CONCLUSION

Through this research, a multidimensional data model has been built that contains information about the research activities of lecturers in the Politeknik Negeri Sriwijaya. This study includes 446 lecturer data records, 34 department data records, 483 record research data records spread over the period of the study 2018, 2019, and 2020. The information generated from the developed model can help higher education management carry out strategic planning. and decision-making. In addition, the resulting information can be used in the preparation of data in accreditation instruments, both for university or institutional accreditation and study program accreditation. The star-schema model was chosen in this study to facilitate the multidimensional database modeling process and to speed up the query process carried out in the reporting and presentation stages of data in a graphical form that can be displayed in a web browser application. Further research from this research is optimizing the data analysis process, and developing data mining applications from data that has been successfully processed.

REFERENCES

[1] L. W. Santoso and Yulia, "Data Warehouse with Big Data Technology for Higher Education", Procedia Computer Science, vol. 124, pp. 93-99, 2017.

[2] M. Baranovic, M. Madunic, I. Mekterovic, "Data warehouse as a part of the higher education information system in Croatia", in Proceedings of the 25th International Conference on Information Technology Interfaces (ITI), 2003, pp. 121 - 126.

[3] X. Yu, "The Application of Data Warehouse in Teaching Management in Colleges and Universities", J. Phys.: Conf. Ser. 1738 012090, 2021.

[4] Y. Bassil, "A Data Warehouse Design for A Typical University Information System", Journal of Computer Science & Research (JCSCR), vo. 1, no. 6, 2012, pp. 12 – 17.

[5] W. Inmon, "Building the Operational Data Store", 2nd ed. John Wiley & Sons, 1999.

[6] D. Bogdanova and M. Snoeck , "CaMeLOT: An Educational Framework For Conceptual Data Modelling", Information and Software Technology (2019), doi: https://doi.org/10.1016/j.infsof.2019.02.006.

[7] A. A. Yulianto, Y. Kasahara, "Data Warehouse System for Multidimensional Analysis of Tuition Fee Level in Higher Education Institutions in Indonesia", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 6, 2020, pp. 541-550.

[8] DRPM, "Panduan Peneltian dan Pengabdian Kepada Masyarakat Edisi XIII", Kementrian Pendidikan, Kebudayaan, Riset, dan Teknologi, 2020.

[9] Rudy, E. Miranda, E. Suryani, "Implementation of Data Warehouse, Datamining and Dashboard for Higher Education", Journal of Theoretical and Applied Information Technology, vo. 64, no. 3, 214, pp. 710-717.

[10] W. H. Inmon, "Building the Data Warehouse 4th ", Wiley Publishing inc, 2005.

[11] D. Darmawikarta, "Dimensional Data Warehousing with MySQL: A Tutorial", Brainy Software Corp. 2007.

[12] O.M. Zea, J.P. Gualtor, S.L. Mora, "A Holistic View of Data Warehousing in Education", IEEE Access, vo. 6, p. 64659-64673, Okt 2018, DOI:10.1109/ACCESS.2018.2876753.

[13] G. Thakur, A. Gosain, "A Comprehensive Analysis of Materialized Views in a Data Warehouse Environment", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 5, 2011, pp. 76-82.

[14] A. Chakir, H. Medromi, A. Sayouti, "Actioins for Data Warehouse Success", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No. 8, 2013, pp. 130-133.

[15] A.F. Neamah, "Adoption of Data Warehouse in University Management: Wasit University Case Study", 2021 J. Phys.: Conf. Ser. 1860 012027.

[16] Asroni, T. B. Arimbi, S. Riyadi, "Implementing of Data Warehouse Data Alumni using the Single Dimensional Data Store Method", J. Phys.: Conf. Ser. 1471 012021.

[17] K.A. Shobirin, A.P.S. Ikandar, I.B.A. Swamardika, "Data Warehouse Schemas using Multidimensional Data Model for Retail", International Journal of Engineering and Emerging Technology, vol. 2, no. 1, January-June 2017, pp.84-86.

[18] Y. Peng, "Metadata Based Visualization System for Multidimensional Data", Advances in Intelligent Systems Research, vol. 134, 2017, pp.564-567.

Empirical Analysis of Feature Points Extraction Techniques for Space Applications

Janhavi H. Borse[1]
Department of Computer Engineering
Smt. Kashibai Navale College of Engineering
Savitribai Phule Pune University, Pune, India

Dipti D. Patil[2]
Department of Information Technology
MKSSS's Cummins College of Engineering for Women
Savitribai Phule Pune University, Pune, India

***Abstract*—Recently, space research advancements have widened the scope of many vision-based techniques. Computer vision techniques with manifold objectives require that valuable features are extracted from input data. This paper attempts to analyze known feature extraction techniques empirically; Scale Invariant Feature Transform (SIFT), Speeded up robust features (SURF), Oriented fast and Rotated Brief (ORB), and Convolutional Neural Network (CNN). A methodology for autonomously extracting features using CNN is analyzed in more detail. The autonomous process demonstrates the use of convolutional neural networks for feature extraction. Those techniques are studied and evaluated empirically on lunar satellite images. For analysis, a dataset containing different affine transformations of a video frame is generated from a sample lunar descent video. The nearest neighbor algorithm is then applied for feature matching. For an unbiased evaluation, a similar process of feature matching is repeated for all the models. Well-known metrics like repeatability and matching scores are employed to validate the studied techniques. The results show that the CNN features showed much better computational efficiency and stable performance concerning matching accuracy for lunar images than other studied algorithms.**

Keywords*—Artificial intelligence; convolutional neural network; computer vision; feature extraction; machine learning; satellite images; space research

I. Introduction

Recent advances in space exploration have opened doors for many research challenges. Processing real-time videos and images captured through spacecraft cameras is one of such challenging tasks. Extracting features useful for further space exploration and navigation tasks is at the primary stage. A spacecraft, once injected into a planet's orbit, keeps on orbiting around the planet. While in orbit, it keeps on capturing videos and images through its onboard cameras. Such kinds of motion result in spatially transformed images of the same scene majority of times. Detecting points of interest from such images or videos in real-time is of paramount importance and a challenging task indeed.

Many proven systems exist which work as image pre-processing techniques for computer vision tasks. There are few areas like image retrieval, medical imaging, object detection, and recognition where these techniques are extensively used. But in this era of automation and artificial intelligence, manual pre-processing of images needs to be avoided. Hence many new systems have been developed with automated feature detection procedures in these domains using deep CNNs. Still, the area of space research has a scope to enter into this automation. This paper intends to analyze a few feature extraction techniques concerning their suitability and sustainability in space applications.

II. Related Work

There are many state-of-the-art algorithms available in the literature as feature detectors & descriptors. Still, their computational complexity does not allow them to be used for real-time tasks. Few comparisons amongst them are available in the reviews [1]–[8]. Many of these algorithms are proposed for detecting and describing points of interest from an image. Initial emphasis was only detecting points of interest or edges from raw images for object detection tasks. Later the focus was shifted to object recognition tasks by taking care of spatial transformations. For keypoint extraction, remarkable work is brought in by the Harris Corner detector [9] and Scale Invariant Feature Transform (SIFT) [10]. Harris Corner detector can extract key points valid for feature tracking algorithms, while SIFT addresses invariance's challenge to affine transformations. But these algorithms were computationally intensive, and hence the next challenge was to speed up the feature detection process. Researchers eventually discovered the new developments like Speeded up robust features (SURF) [11], [12], Features from accelerated segment test (FAST) [13], [14], Binary robust independent elementary features (BRIEF) [15], and Oriented FAST and rotated BRIEF (ORB) [16]. Through the literature, SURF is found to deliver quality features and is computationally efficient as well. ORB, which is a combination of FAST & BRIEF, is computationally speedy than SURF, but the features it extracts are not suitable for image matching tasks. Moreover, these algorithms are standalone versions, and their real-time applicability is questionable. Few works demonstrated the use of feature extraction techniques specific to application domains like medical imaging [17], image retrieval systems [18], and gesture recognition [19], [20].

In the last decades, few deep learning techniques and convolutional neural network techniques[21], [22], [2] are also developed with an abundance in data availability and computationally powerful resources in recent years. The ultimate target of these techniques is image recognition and computer vision task. Many of these techniques rely on already built and tested deep neural network models like Inception [23], VGG [24], XCeption [25], ResNet [26]. Many

researchers have used transfer learning by finetuning ready models for reaching their goals. Two things must be clear in transfer learning before using a particular base model; the first is the dataset on which the model is trained, and the second is the intended application domain. Most of these well-known models are trained on a generalized IMAGENET (http://www.image-net.org) dataset. Hence the knowledge gained through its training is adapted in current research, aiming to deal with data consisting of satellite images and videos.

The real-time requirements to address space research challenges and the existing methods discussed so far motivate us to empirically analyze a few feature extraction techniques like SIFT, SURF, ORB, and CNN. An automated feature extraction process using a convolutional neural network (CNN) is also designed and implemented for experimental analysis. Hence, this work's primary purpose is to study these techniques empirically and analyze their performance concerning time-critical space applications. For this purpose, a dataset consisting of lunar images is constructed from videos captured by a spacecraft's onboard cameras. Each image is spatially transformed with the known transformation matrix. Features are extracted from each image and its transformed versions using all the studied techniques. Image matching using the nearest neighbor algorithm is performed for each image tuple (reference image, transformed image). Ideally, when an image is spatially transformed, its transformed versions show many similarities in detected features as long as a downward-looking camera captures the video with minimal frame delay. One can efficiently compute the ground truth feature vector with known transformations by applying the same transformations to the reference image's features. Finally, their results are validated with available performance metrics and compared with each other.

The details of state-of-the-art techniques like SIFT, SURF, ORB are prevalently available in the literature. Past few years, the scope of CNN is widened due to the automation in the feature extraction process. Still, for few domains like space research, some more analysis is needed for testing its reliability. This paper is intended to perform a comparative analysis of these techniques for space research applicability. For testing the validity of the analysis, CNN features are compared with the SIFT, SURF, and ORB features.

This paper is organized as follows: Section 2 discusses an automated feature extraction process using a CNN architecture. Section 3 elaborates on experimental setup and dataset generation. The performance metrics are discussed in Section 4. Section 5 discusses results and comparisons in classical algorithms, and finally, Section 6 concludes the paper.

III. An Automated Feature Extraction Process using CNN

A. Selecting a ResNet Architecture for CNN Features

Transfer learning is used to generate features. The CNN architecture consists of a ResNet as a base model without a classification layer, as shown in Fig. 1. It is then cascaded with one flatten layer, 2 fully connected (FC) layers at the end. The output of convolutional layers is 3-dimensional maps- $(M \times N \times f)$. The first two dimensions are the size of a feature map, and the last dimension is the number of maps generated at each layer. The number of feature maps corresponds to the number of filters. After the last convolutional layer, a flattened layer is introduced to flatten a 3-dimensional tensor into a single dimension. FC layers at the end serve as an output layer for the Model. The number of computational units in this layer will decide the dimensions of the feature vector. A general deep learning network model expects softmax or other nonlinear functions at the output layer for targeting classification or another more vital task. The goal is to extract features, so the nonlinear function interface at the output layer is removed from the ResNet block.

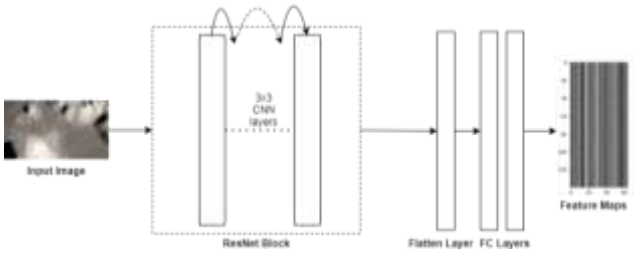


Fig. 1. CNN Model with ResNet base Architecture. Only 64 Features are Shown for Simplicity.

It is assumed that each CNN layer consists of an f-number of $m \times n$ sized high-level convolutional filters. Each input image $I^{i}_{(M,N)}$ is padded to preserve the size of the original image. Then the padded image is passed into the convolution layer to get an output image as,

$$I^{o}_{(M,N)} = I^{p}_{(M+p,N+p)} * f_{(m,n)} \quad (1)$$

In (1), $I^{p}_{(M+p,N+p)}$ does zero-padding form a padded image to an input image $I^{i}_{(M,N)}$ with pad size, p. Rectifier Linear Unit (ReLU) is used as an activation function for both layers. Max pooling is applied to this successive output after padding. A detailed procedure for extracting features through the CNN model is given in a pseudo-code described by algorithm 1.

Using Algorithm 1, CNN features are extracted. The transformed dataset contains both the reference image and its transformed versions. Training is done by applying an 8:2 train-test cross-validation split on the dataset. The details of data collection and creation are described in section 3. The features are extracted from the reference image I^{ref} first and then from its transformed versions: $I^{trans} = (I^{ref}_{R}, I^{ref}_{T}, I^{ref}_{S})$. These features represent the points of interest from each image. Training starts with transfer weights from trained ResNet model. Training is done to minimize the average loss function given by (2). K_p^{gtruth} are the ground truth feature vectors and these are computed using the known transformation parameters. K_p^{trans} are the features extracted from the CNN (Fig. 1(a)). Weights W of the network is adjusted during each epoch to minimize f_{loss} using (2).

$$f_{loss} = \frac{\sum_{D_{trans}} min_W \left\| K_p^{gtruth} - K_p^{trans} \right\|}{|D_{trans}|} \quad (2)$$

Algorithm 1: Procedure for extracting features through a CNN Model.

START

INPUT: Video V

OUTPUT: Features K (1024 x 1)

PROGRAM CNNModel

1. D_{raw} := Call: Function VideoToImage to generate images from video
2. Apply geometric transformations on raw images to generate transformed image dataset,

$$D_{trans} := T_{\text{rotate}}(D_{raw}) \cup T_{\text{scale}}(D_{raw}) \cup T_{\text{translate}}(D_{raw})$$

3. For each image $I^{i}_{(M,N)}$ in D_{trans} REPEAT:

$$I^{p}_{(M+p,N+p)} := Padding\left(I^{i}_{(M,N)}\right)$$

For all filters in a Layer REPEAT:

$$I^{o}_{j(M,N)} := I^{p}_{(M+p,N+p)} * f^{j}_{(m,n)}$$

$$I^{R}_{j(M,N)} := ReLU\left(I^{o}_{j(M,N)}\right)$$

$$I^{PR}_{j(M+p,N+p)} := Padding\left(I^{R}_{j(M,N)}\right)$$

$$I^{Pool}_{j(M,N)} := MaxPool\left(I^{PR}_{j(M+p,N+p)}\right)$$

END REPEAT

1. REPEAT step 3 for next CNN Layer
2. Flatten each image into a single-dimensional vector,

$$V^{i}_{(MxNxf,1)} := Flatten\left(I^{Pool}_{j(M,N)}\right)$$

3. Pass this vector through FC layer,

$$K^{i}_{(256,1)} := FC\left(V^{i}_{(MMxNxf,1)}\right)$$

END REPEAT

END

After training the network, it is evaluated on the rest of the test images. The working of the first algorithm is as follows. Initially, images were extracted from a lunar video to generate a raw image dataset D_{raw} . Then, the following affine transformations were applied to produce a transformed image dataset D_{trans}.

$$M_{scale} = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$M_{translate} = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix}$$

$$M_{rotate} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (3)$$

A translation factor of (t_x, t_y) pixels, a rotational angle of $\theta^0$and a scale factor of (s_x, s_y) were used as described by (3). Each image from this new dataset is fed to CNN. Image dimensions are 640×360. Each image is padded to preserve the dimensions and then convolved with f filters of size 3×3. Rectifier Linear Unit (ReLU) is the nonlinear activation function applied to the convolved output. It generates feature maps of size $640 \times 360 \times f$. It is downsampled by using a max-pooling operation. The process is repeated for each next layer with $2 * f$ filters of size 3×3 till the last layer. In the end, the 3-dimensional feature maps are converted into a 1-dimensional tensor using a flatten layer. These flattened feature maps were used as feature vectors of the CNN model.

Algorithm 2 is implemented to track similar feature points from the reference image and transformed image for evaluating the feature similarity. Initially, for each image in the dataset, the features are extracted from all four known techniques, SIFT, SURF, ORB, and CNN, to extract feature vectors K_p^{ref} *and* K_p^{trans} . Then these feature vectors are passed on to the similarity matching using the nearest neighbor algorithm. Scores of matchings between the two image features are used to evaluate different techniques under consideration.

Algorithm 2: Extracting Similar Feature Points from Reference Image and Transformed Image

START

INPUT: Images: Reference I^{ref} and transformed image I^{trans}

OUTPUT: Matched keypoints $K_p^{matched}$

PROGRAM TRACKeypoints

For each image pair (I^{ref}, I^{trans}) in D_{trans} DO:

1. Apply any of the feature detectors like SIFT, SURF, ORB, and CNNModel to get extract keypoints K_p^{ref} and K_p^{trans}
2. Apply image matching technique to find matched key points,

$$K_p^{matched} := NearestNeighbour\left(K_p^{ref}, K_p^{trans}\right)$$

3. Pass vector $K_p^{matched}$ to further compute repeatability and matching scores between two images.

END

IV. Experimental Setup and Dataset Generation

The CNN model was implemented on a 2.4 GHz Intel Core i7 processor with 16 GB DDR4 RAM. Code scripts were written in Python 3.7 with tensor flow framework as backend.

For this research, a python script was written for generating images from a spacecraft landing video. The video is publicly available on the website, https://svs.gsfc.nasa.gov/. This video is an animated view of the landing site of Apollo 17 - Lee Lincoln scarp. The sources created this visualization from Lunar Reconnaissance Orbiter (LRO) photographs and elevation mapping. The video's frame rate is found to be 25 fps, and hence each 25th image frame was captured and stored as an image. In all, 915 grayscale images were generated, which forms the raw image dataset. In the raw image dataset, 200 images were selected at random, and geometric

transformations are applied to create the transformed image dataset containing 800 images. The size of each image in the dataset is 640x360 pixels. Out of the whole dataset, 640 images were used for training, and the remaining 160 are used for testing. Sample images from the dataset are shown in Fig. 2. The complete procedure for dataset generation and description can be found in our prior work [22].

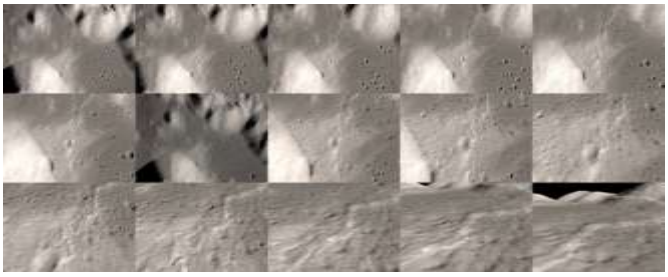


Fig. 2. Montage of Sample Images from the Raw Image Dataset.

The performance of the implemented CNN is evaluated using the metrics repeatability score and matching score [3], [6], [27]–[29], which are widely used for the evaluation of feature detectors. For validating the results of CNN, known feature detectors like SIFT, SURF, and ORB are implemented in the python and Opencv environments. SIFT Lowe's implementation [10] was directly used with few modifications. The ORB algorithm is implemented with two different variations corresponding to the number of feature points extracted equally to 1000 and 500.

The following procedure is followed for computing these metrics to perform an unbiased evaluation of studied feature detectors. Initially, features are extracted from the reference image and then from its transformed version. The descriptors of both images were passed to a Fast Library for Approximate Nearest Neighbours (FLANN) matching algorithm to find a matched point from a transformed image similar to the reference image. This algorithm tends to see the overlapped region from both the images and then returns all the points that match the points in the reference image using a known homography. All the key points in this common region are called correspondences between the two images. After computing the maximum correspondences, the algorithm tries to find the correct matches using some threshold. These interim computations help to calculate the scores of performance parameters.

V. Performance Metrics

For evaluating and comparing the performances of the feature detectors, performance metrics, namely repeatability, matching score, and time taken for feature extraction, are employed.

Image correspondences are key points in common regions between two images using known homography. Repeatability is the measure of the robustness of a feature detector to the external image transformations. The repeatability score is calculated by using (4).

$$\text{Repeatability} = \frac{C^+}{F_{ref}} \tag{4}$$

In (4), C^+ is maximum image correspondences, F_{ref} is the number of features of the reference image.

A matching score measures accuracy while matching the descriptors of logically the same key points from two different images. The matching score is calculated by the formula given by (5).

$$Matching\ Score = \frac{C^+ \cap C^*}{F_{ref}} \tag{5}$$

As in (4), C^* is a number of correct matches.

VI. Results and Discussion

A. Performance Evaluation using Repeatability

The automated feature extraction process results using CNN are compared with conventional methods like SIFT, SURF, and ORB. The distribution of repeatability scores on a percentage scale for all the studied algorithms is shown in Fig. 3. Both versions of SIFT are having almost similar distributions of repeatability ranging between an interval [40,100]. SURF values are found to lie within a range of [55,100], while the range for ORB is [40,100]. ORB features seem to be more minor variants to the transformations, while SIFT & SURF features are highly variant to the input transformations. Graphs show that the CNN model for all the test samples has retained the constant repeatability score of 100%. It means that CNN features are more efficient in finding repeated regions of interest. In the transformations like rotation and scaling, most image regions are repeated. But during a translational shift, few new image regions are added while few regions are subtracted from the original image. An ideal feature detector must take such changes into account. But CNN has neglected the translational shift. The extracted descriptors for a common region of the transformed image always find a proper match for reference descriptors with minimum losses. Most of the features are matched between the two images. As the same procedure is followed for finding key point matches between the two images, it can be concluded that the CNN features are more robust to external transformations than the other classic methods. As it shows no variation in the repeatability scores, it is one of the stable feature detectors. When tuned to generate a more significant number of features (ORB1000), it results in more variations in repeatability than ORB500 and hence can generalize better.

Fig. 4 shows the average repeatability score obtained by all the studied techniques. The score is high for ORB-500 and CNN64. For ORB-1000 also is comparable. But for SURF and SIFT, it is less than 90%. Overall, ORB and CNN are found more robust and hence showed stable performance in terms of repeatability of features. SIFT and SURF are quite unstable as far as this dataset is concerned. CNN features that showed the highest average repeatability show that it is invariant to the external factors such as camera position, angle, motion while extracting the image features. Such transformations are widespread in real-time captured data and hence need a stable feature detector. Thus, CNN model can generalize better if tuned for a more significant number of output features and trained on more data samples.

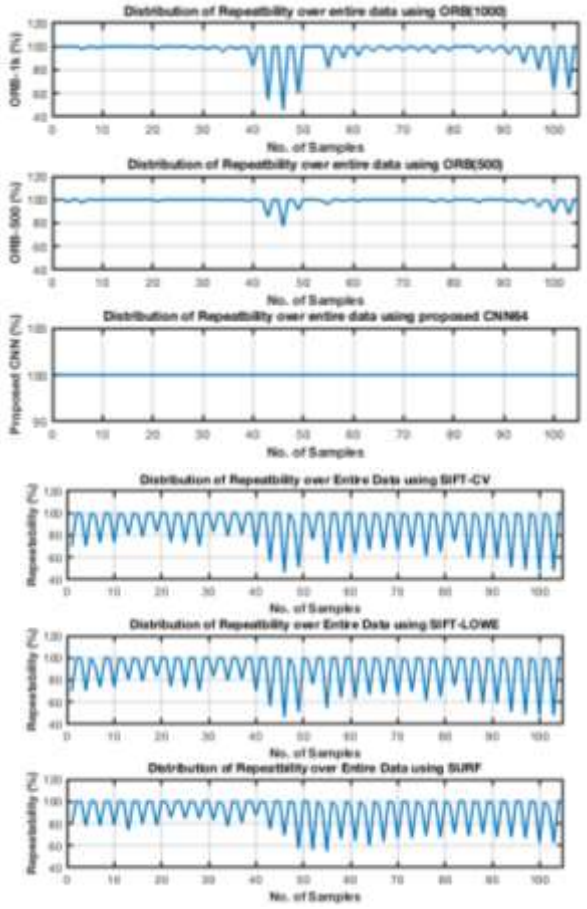


Fig. 3. Distribution of Repeatability Score obtained from Studied Techniques.

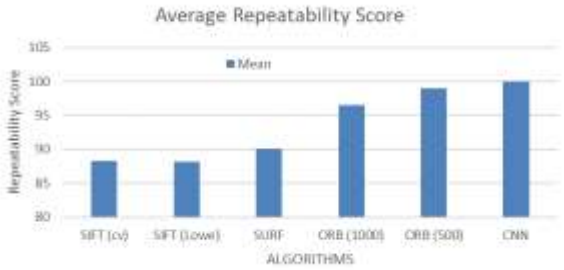


Fig. 4. Average Repeatability Score obtained from Various Algorithms.

B. Performance Evaluation using Matching Score

In addition to repeatability, a matching score is yet another evaluation parameter employed to quantify the performance of the studied techniques. Fig. 5 shows a box plot for matching scores from the features obtained through CNN and conventional algorithms. SIFT and ORB-generated features show skewness in the results, which does not seem stable and reliable data for image recognition or classification tasks. The skewness in the data might bias the model for the following tasks. SURF seems reasonably reliable but has introduced greater variance to changing inputs. CNN boxplot is concentrated near the mean value, and it does not show any skewness in the results. Hence CNN based model seems unbiased, and hence more stable.

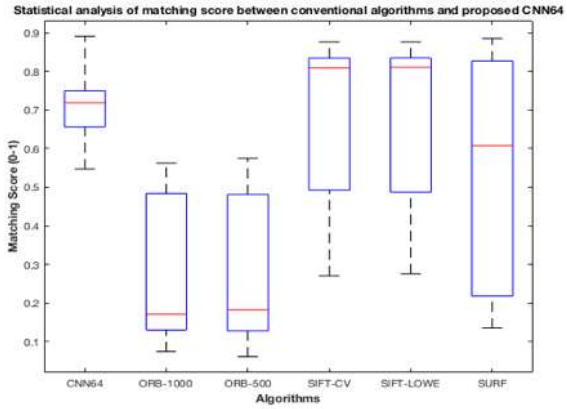


Fig. 5. Statistical Analysis of Matching Score for different Algorithms.

Table I shows the values of computed statistics. The average matching score obtained through ORB is the least of all. The highest score is obtained through CNN. Although the average matching score obtained through SIFT is not far less than that obtained through CNN, SIFT generated score shows the more significant variance in computation. Ideally, the matching score should be on a higher side with minor variance as the images are part of a shorter duration video, mainly capturing the same ground scene.

TABLE I. PRIMARY STATISTICS OF MATCHING SCORE OBTAINED FROM DIFFERENT ALGORITHMS

Algorithms/Stats	SIFT-cv	SIFT-Lowe	SURF	ORB-1000	ORB-500	CNN
Mean	0.6917	0.6893	0.5495	0.2655	0.2696	0.7107
Median	0.8090	0.8101	0.6079	0.1713	0.1826	0.7187
Variance	0.0367	0.0369	0.0716	0.0290	0.0294	0.0037

On the other hand, SURF and ORB have shown significantly lesser matching scores than the CNN model. ORB shows consistency in the matching score computation with minor variation compared to SIFT and SURF, but it is more significant than CNN. Once again, the CNN model has shown invariance against the transformations. Overall, concerning matching scores, the CNN features have shown better performance than others.

To measure consistent and robust performance, we run a one-way Analysis of Variance (ANOVA) test to prove our hypothesis for matching scores computed through the application of all the algorithms. For testing the hypothesis, we selected 100 random samples out of the whole dataset with replacement. We run the ANOVA test for 10 such samples. We assumed that a stable and robust detector would always show negligible between-group variance, and hence its sample means are more equivalent to the grand mean of the population. We rigorously tested each sample mean against its grand mean for each algorithm listed in Table I. We run the test with a 95% of a confidence interval. For CNN, we found our assumption held throughout all samples. The assumption did not hold in the case of other algorithms. Few sample means were far away from the grand mean as in Table I. It proved the robustness and consistent behavior of the CNN compared to other feature detectors.

C. Performance Evaluation using Computational Time

The most critical evaluation parameter that needs to take care of for real-time data is the computation time required to detect and extract points of interest. Fig. 6 shows the joint bar graph, which describes the time needed for processing each image by the proposed CNN model and its companion algorithms. The red bar indicates the number of descriptors detected by an algorithm, and the blue bar shows the time required to perform that task.

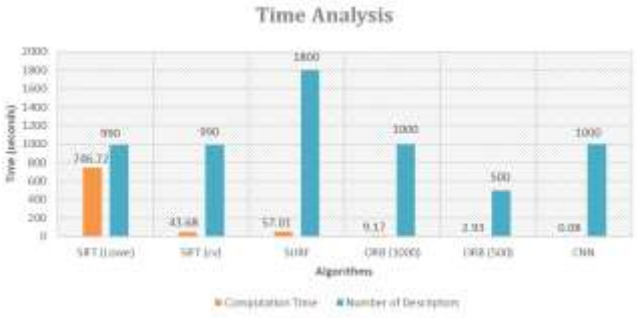


Fig. 6. Algorithmic Analysis of Average (Per Image) Processing Time.

The bar graph of Fig. 6 shows the number of descriptors extracted by each algorithm along with computation time. CNN shows the shortest average processing time amongst all. Both ORB versions show the following smallest time requirement. Then comes SURF, SIFT(cv), and at last SIFT(Lowe). SIFT Lowe's version consumes the highest time compared to others because it processes 2-dimensional data without pre-processing it to any time-efficient form. SIFT and SURF attempt to extract all possible key points from an image and lag in time performance. During tensor flow implementation of CNN, each image is converted into the most efficient tensor representation and then processed further; hence the time required for feature extraction gets drastically reduced. Overall, the evaluation of the CNN is better than the listed algorithms for generated lunar image data.

VII. Conclusion

In this paper, state-of-the-art algorithms for feature extraction are implemented and analyzed on lunar descent image data in detail. A similar process is followed for unbiased evaluation, and known metrics of repeatability, matching score accuracy, and extraction time are used to compare implemented algorithms.

From the detailed analysis of results, it is observed that the CNN model has outperformed the studied conventional algorithms based on suggested performance metrics. The CNN model is capable of handling real-time data with less time requirement. Once a few network parameters are decided, CNN does its job automatically using the input data. No hand-crafted tasks such as image pre-processing, image localization, segmentation are needed as in conventional algorithms.

In effect, the overall performance of the CNN architecture, when compared to existing algorithms, showed much better computational efficiency and stability. The analysis shows that CNN's more profound architecture with transfer learning can be used to meet the real-time demands of space research. But, vigorous training and validation using extensive data are necessary to generalize the model to a greater extent. Extension to work is validating the model by using generated features for object detection tasks.

Acknowledgment

Firstly, we are grateful to Dr. Vinod Kumar, Division Head, U R Rao Satellite Centre ISRO, India, for advising and guiding us in understanding space-related issues.

We want to thank ISRO, India, for supporting & funding this research. We want to express special thanks to Smt. Kashibai Navale College of Engineering and Savitribai Phule Pune University India, for providing us opportunity and support.

References

[1] Ali Ismail Awad and Mahmoud Hassaballah, Image Feature Detectors and Descriptors, vol. 630, no. February. 2016.

[2] P. Fischer, A. Dosovitskiy, and T. Brox, “Descriptor Matching with Convolutional Neural Networks: a Comparison to SIFT,” pp. 1–10, 2014, [Online]. Available: http://arxiv.org/abs/1405.5769.

[3] B. Istanbul, “Analysis of Feature Detector and Descriptor Combinations with a Localization Experiment for Various Performance Metrics Ertugrul BAYRAKTAR* , Pınar BOYRAZ.”

[4] K. Lenc and A. Vedaldi, “Large scale evaluation of local image feature detectors on homography datasets,” Br. Mach. Vis. Conf. 2018, BMVC 2018, 2019.

[5] S. Li, “A review of feature detection and match algorithms for localization and mapping,” IOP Conf. Ser. Mater. Sci. Eng., vol. 231, no. 1, 2017, doi: 10.1088/1757-899X/231/1/012003.

[6] K. Mikolajczyk et al., “A comparison of affine region detectors,” Int. J. Comput. Vis., vol. 65, no. 1–2, pp. 43–72, 2005, doi: 10.1007/s11263-005-3848-x.

[7] G. M. Moura and R. L. D. S. Da Silva, “Analysis and evaluation of feature detection and tracking techniques using OpenCV with focus on markerless augmented reality applications,” J. Mob. Multimed., vol. 12, no. 3–4, pp. 291–302, 2017.

[8] E. Salahat and M. Qasaimeh, “Recent advances in features extraction and description algorithms: A comprehensive survey,” Proc. IEEE Int. Conf. Ind. Technol., pp. 1059–1063, 2017, doi: 10.1109/ICIT.2017.7915508.

[9] M. Harris, C. and Stephens, “A Combined Corner and Edge Detector,” in In C. J. Taylor, editors, Proceedings of the Alvey Vision Conference, 1988, pp. 23.1-23.6, doi: 10.5244/C.2.23.

[10] D. G. Low, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., pp. 91–110, 2004, [Online]. Available: https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf.

[11] H. Bay, T. Tuytelaars, and L. Van Gool, “LNCS 3951 - SURF: Speeded Up Robust Features,” Comput. Vision–ECCV 2006, pp. 404–417, 2006, [Online]. Available: http://link.springer.com/chapter/10.1007/11744023_32.

[12] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-Up Robust Features (SURF),” Comput. Vis. Image Underst., vol. 110, no. 3, pp. 346–359, 2008, doi: 10.1016/j.cviu.2007.09.014.

[13] E. Rosten and T. Drummond, “Machine Learning for High-Speed Corner Detection,” pp. 430–443, 2006.

[14] E. Rosten, R. Porter, and T. Drummond, “Faster and better: A machine learning approach to corner detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 1, pp. 105–119, 2010, doi: 10.1109/TPAMI.2008.275.

[15] S. Leutenegger, M. Chli, and R. Y. Siegwart, “BRISK: Binary Robust invariant scalable keypoints,” Proc. IEEE Int. Conf. Comput. Vis., no. November, pp. 2548–2555, 2011, doi: 10.1109/ICCV.2011.6126542.

[16] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," Proc. IEEE Int. Conf. Comput. Vis., pp. 2564–2571, 2011, doi: 10.1109/ICCV.2011.6126544.

[17] A. Reema Matthew, A. Prasad, and P. Babu Anto, "A review on feature extraction techniques for tumor detection and classification from brain MRI," 2017 Int. Conf. Intell. Comput. Instrum. Control Technol. ICICICT 2017, vol. 2018-January, pp. 1766–1771, 2018, doi: 10.1109/ICICICT1.2017.8342838.

[18] S. Dhingra and P. Bansal, "Experimental analogy of different texture feature extraction techniques in image retrieval systems," Multimed. Tools Appl., vol. 79, no. 37–38, pp. 27391–27406, 2020, doi: 10.1007/s11042-020-09317-3.

[19] A. Sharma, A. Mittal, S. Singh, and V. Awatramani, "Hand Gesture Recognition using Image Processing and Feature Extraction Techniques," Procedia Comput. Sci., vol. 173, no. 2019, pp. 181–190, 2020, doi: 10.1016/j.procs.2020.06.022.

[20] H. Dino et al., "Facial Expression ssRecognition based on Hybrid Feature Extraction Techniques with Different Classifiers," TEST Eng. Manag., vol. 83, no. 22319, pp. 22319–22329, 2020.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention, 2015, pp. 234–241.

[22] J. H. Borse, D. D. Patil, and V. Kumar, "Tracking Keypoints from Consecutive Video Frames Using CNN Features for Space Applications," Teh. Glas., vol. 15, no. 1, pp. 11–17, Mar. 2021, doi: 10.31803/tg-20210204161210.

[23] C. Szegedy et al., "Going deeper with convolutions," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12-June, pp. 1–9, 2015, doi: 10.1109/CVPR.2015.7298594.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–14, 2015.

[25] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-Janua, pp. 1800–1807, 2017, doi: 10.1109/CVPR.2017.195.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[27] S. Ehsan, N. Kanwal, A. F. Clark, and K. D. McDonald-Maier, "Improved repeatability measures for evaluating performance of feature detectors," Electron. Lett., vol. 46, no. 14, pp. 998–1000, 2010, doi: 10.1049/el.2010.1442.

[28] T. Mouats, N. Aouf, D. Nam, and S. Vidas, Performance Evaluation of Feature Detectors and Descriptors Beyond the Visible, vol. 92, no. 1. Journal of Intelligent & Robotic Systems, 2018.

[29] J. L. Sch¨onberger, H. Hardmeier, T. Sattler, and M. Pollefeys, "Comparative Evaluation of Hand-Crafted and Learned Local Features," pp. 01–10, 2017.

Traffic Adaptive Deep Learning based Fine Grained Vehicle Categorization in Cluttered Traffic Videos

Shobha B.S, Deepu.R
Department of Computer Science and Engineering
Maharaja Institute of Technology Mysore
Mysore, India

***Abstract*—Smart traffic management is being proposed for better management of traffic infrastructure and regulate traffic in smart cities. With surge of traffic density in many cities, smart traffic management becomes utmost necessity. Vehicle categorization, traffic density estimation and vehicle tracking are some of the important functionalities in smart traffic management. Vehicles must be categorized based on multiple levels like type, speed, direction of travel and vehicle attributes like color etc. for efficient tracking and traffic density estimation. Vehicle categorization becomes very challenging due to occlusions, cluttered backgrounds and traffic density variations. In this work, a traffic adaptive multi-level vehicle categorization using deep learning is proposed. The solution is designed to solve the problems in vehicle categorization in terms of occlusions, cluttered backgrounds.**

Keywords*—Vehicle categorization; deep learning; traffic density estimation; clutter

I. Introduction

Smart traffic management based on video feeds from traffic surveillance cameras is being proposed as a means for efficient traffic regulation at a lower cost compared to sensors based traffic management.

Smart traffic management aims to regulate the traffic conditions in peak hours, manage congestions, transport of emergency vehicles, detect and handle accidents/incidents in the road. Vehicle categorization is an important functionality in smart traffic management. Segmenting the vehicles and categorization of them based on multiple levels like vehicle type, speed, direction of travel, meta attributes like color, make etc. is important for localization and tracking of vehicles in smart traffic management.

Vehicle categorization becomes very challenging in presence of noisy cluttered background, environmental conditions (fog, rain, lighting and haze), shadow and occlusions etc. The problem becomes even more difficult in case of need for fine grained multi-level categorization like learning speed, direction and meta attributes of the vehicles from the video stream. But the applications of this multilevel categorization are innumerous in terms of regulation, description, indexing and tracking.

This work deals with this need and proposes a traffic adaptive deep learning multi-level vehicle categorization which can work in conditions of cluttered background and occlusions in the video. The approach is an integration of two different mechanisms – one designed for low density traffic and another for high density traffic. A deep learning optimized topological active net segmentation is done to segment the vehicles in case of low density traffic. For high density traffic, convolutional neural network segmentation is done. After segmentation, features are extracted and mapping is done to learn various meta-attributes of the segmented vehicles.

II. Literature Survey

Authors in [1] proposed an architecture called TPSVedet for categorization of vehicle to small, medium and large size. From the video, background model is constructed by averaging the frames over time. Background is subtracted from each frames and over the ROI region geometry based features are extracted. PCA is used for dimensionality reduction and the dimension reduced features are classified using machine learning methods like ANN, SVM and AdaBoost. The problem in this approach is that it can detect only one vehicle in the ROI region. Deep convolutional neural network along with joint fine tuning is used for vehicle classification in [2]. Deep residual network (ResNet) model is used for convolutional neural network. Drop out method is used for preventing the over fitting. The network is trained with images and their ground truth vehicle location marked as boxes. The trained model is able to localize the vehicle in the image and classify it one of 11 different types of vehicles. The method cannot work for multiple vehicles in the same image. Authors in [3] reviewed and compared the various methods for feature extraction, global representation and classification for automated vehicle classification. Most of the approaches were found to work for relatively static background and cannot work in presence of occlusions or changing lighting conditions. Geometrical feature based approaches were found to have higher misclassification rate. Texture-based approaches have high sensitivity and computational costs. Authors in [4] proposed a vehicle categorization approach based on geometric feature extraction. Geometric features like shape and size is extracted and passed to trained random forest model to classify the vehicle. This approach assumes fixed vehicle placement and it does not work for multiple vehicles in the image. Authors in [5] experimented with vehicle identification for cars with images taken in different viewpoint of “front (F)”, “rear (R)”, “side (S)”, “front-side (FS)”, “rear side (RS)”, and “All-View”. From their experiment, the convolutional neural network trained with Front side and Rear images were found to have higher accuracy compared to others. Front and Rear image had the information to detect the make for almost all the

cars. Deep learning based vehicle classification is proposed in [6]. A two layer convolution based CNN is trained with car images as input and vehicle type/color as output. The method can classify only a single image and time complexity for classification is also high. Also the approach has only 70% accuracy for color classification. Authors in [7] used texture characteristics in the headlight and grill area to classify the vehicle. Headlight and grill area is segmented from the vehicle falling in the ROI area and the GLCM texture features are extracted from it. Vehicles were then classified based on the similarity of GLCM features. Authors in [8] proposed a vehicle classification system based on side view profile of the vehicles. Side view images of vehicle are skeletonized and features such as joints and endpoints are extracted from it. The features are looked up for similarity against training image features to classify the vehicle. The accuracy of the proposed solution is very sensitive to occlusions. Authors in [9] used YOLOv3 deep learning network for vehicle detection in the images. Road surface area is split to two categories of remote area and proximal area. The vehicle object in the road area is segmented and classified using YOLOv3 to three vehicle category of bus, car and motorcycle. Though the solution can work for multiple vehicles in the image, it does not work for high density traffic. The solution assumes a larger gap between the vehicles. Author in [10] recognized vehicle logo using enhanced scale-invariant feature transform (SIFT)-based feature-matching scheme. Logo is segmented from the vehicle image applying phase congruency calculation. From the logo segment, SIFT features are extracted and matched against trained patterns to recognize the logo. Authors in [11] solved the problem of vehicle color recognition using BoW model. The ROI for identifying the dominant color is implicitly selected in this method. From the ROI region, local color features are extracted and classified using a multi-class SVM to recognize the color. Authors in [12] proposed a robust system for car make recognition from car front images even in presence of low contrast and compression based distortions. Car brand region is segmented and SIFT features are extracted from the brand region. The SIFT features are then matched to training images to recognize the car make. An unsupervised convolutional neural network is used for classification of vehicle type in [13] based on the vehicle frontal images. Sparse Laplacian filter learning is used to capture rich and discriminative information from the vehicle image. The output of the convolutional neural network is the probability of each type the vehicle belongs to. But the method can recognize only one vehicle in the image with no distortion in the frontal view. A simple convolutional neural network model is proposed in [14] for classifying six different vehicle types. Convolutional features are learnt from the low resolution input images. The convolutional features are classified with a fully connected standard network to probability of each class of vehicle. A two stage classification method for vehicle recognition is proposed in [15]. The classification uses both global and local features. An improved canny edge detection with smooth filtering is proposed to extract global features. Local features are extracted using Gabor wavelet. The vehicle is classified to small or large at first stage of classification using the global features and vehicle type is found in second stage classification using local features. Authors in [16] proposed a deformable model integrating both detection and classification into one stage. A deformable part based model is trained using annotated vehicle image for classifying the vehicle. Vehicles are extracted from the traffic image and model alignment is done on extracted image crop. SVM classifier is trained to classify the model to the vehicle type. Regression analysis was used for vehicle classification in [17]. Foreground segments having vehicles are detected first using a wrapping method. Low level features are extracted from the foreground segments and cascaded regression approach is used to classify the vehicles. A stochastic multi class vehicle classification system based on Bayesian model is proposed in [18]. Low dimensional features of vehicle tail light are classified using a Bayesian network to four different vehicle types. Author in [19] used statistical random pixel distribution features acquired from low dimensional images to recognize the logo of the vehicle. Multiscale scanning algorithm is used to jointly detect and classify logos. Author in [20] used speeded up robust features (SURF) to recognize vehicle. SURF features are extracted from front and rear view of the vehicles. The features are then classified by multi class SVM to the type of the vehicle. Author in [21] have presented a detailed review on vehicle detection, recognition and tracking. Multi view methods for vehicle detection are also discussed.

III. Fine Grained Traffic Adaptive Vehicle Categorization

Most of existing solutions are based on drawing the bounding box around the foreground objects and classifying the vehicle type of them. But in case of dense traffic, occlusion makes the drawing of bounding box difficult and the accuracy of vehicle detection difficult. Cluttered background in terms of pedestrian movement, shadows etc. causes error in boundary box localization and due to this vehicle classification becomes erroneous. In this work, the vehicle classification is handled as a three stage process. In first stage, images in the video are pre-processed by removing the background, shadows and illumination artefacts. In the second stage, the image is split to two categories – Type 1 category where bounding box estimation would be easy and a Type 2 category where bounding box estimation is difficult due to clutters in the image. In third stage, for type 1 category images, deep learning convolutional neural network is used to generate bounding box for foreground vehicles and features extracted from the bounding box segment are used for fine grained vehicle classification. For type 2 category images, integration of deep learning with topological active net deformable model is used for efficient segmentation of the vehicles. The clutters are filtered out in this step and the features collected from the mesh are used for fine grained vehicle classification. The architecture of the proposed solution is given in Fig. 1. Each stage of the proposed solution is detailed below.

A. Preprocessing

In this stage, a background model is constructed based on analysis of the frames in the video. The goal of background modeling is to find a best estimate of background so that impact of shadows and sudden illumination changes on the foreground model is minimized.

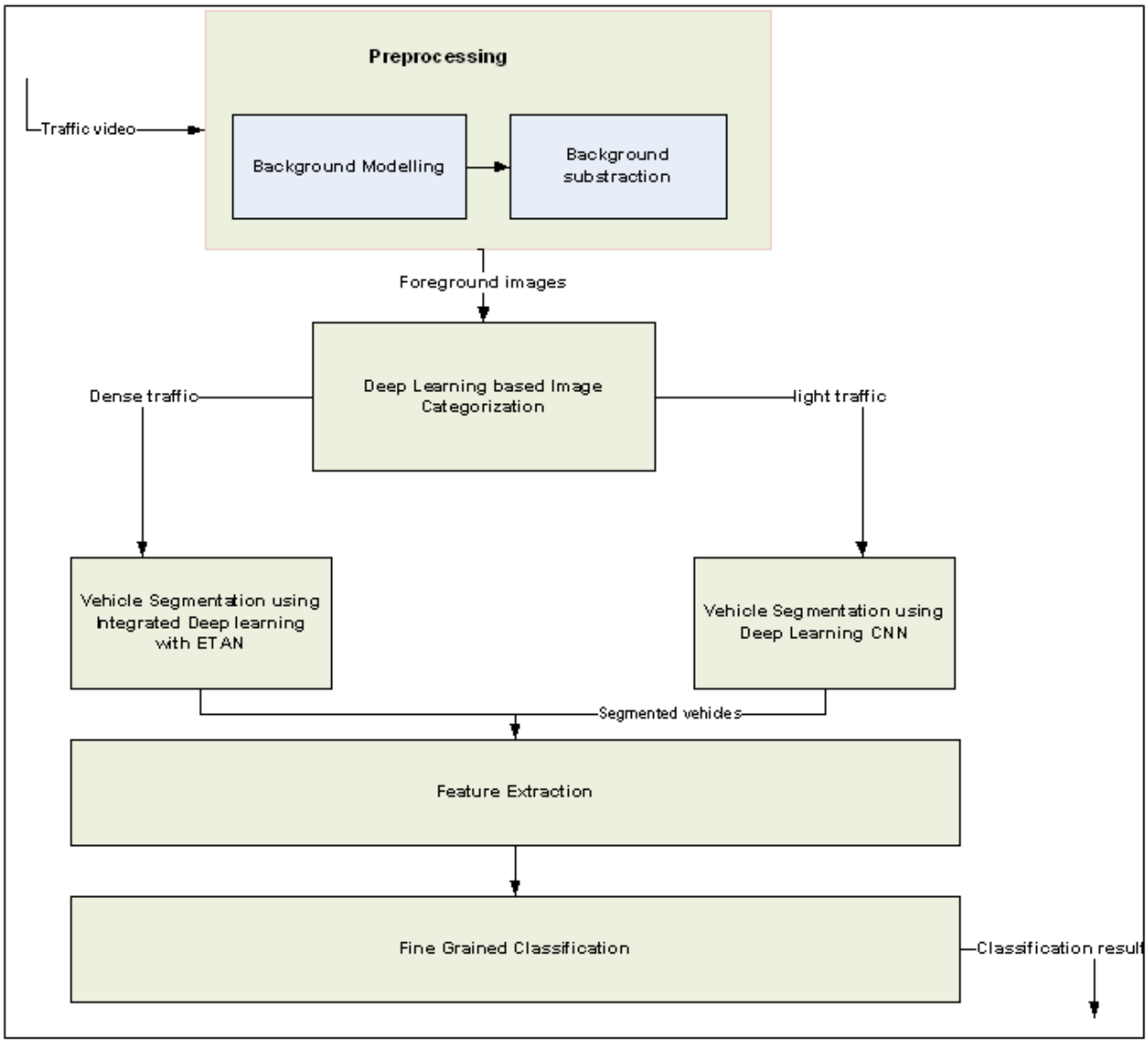


Fig. 1. Proposed Architecture.

This work proposes an adaptive background modeling where the background is first initialized by analysis of few frames and then it is continuously updated for every time foreground is extracted from subsequent frames. This is different from previous approaches of creating a fixed background model by analyzing all the frames. The advantage in this type of adaptive modeling is that unimportant backgrounds do not appear for long period of time and they disappear in subsequent background models.

The initial background model is initialized by taking the pixel values of the first frame and then the model is subsequently updated by calculating the pixel value for each pixel in the background model as

$$p_k(x) = \tilde{p}_{k-1}(x) + \frac{1}{G_k\sqrt{2\pi\sigma^2}} exp(\frac{-1}{2}(\frac{x-x_k}{\sigma})^2) \quad (1)$$

Where, x_k is the pixel value observed at the k^{th}frame.

Where N is the number of frames so far considered.

The value of $\tilde{p}_k$ is calculated as

$$\tilde{p}_k(x) = \frac{p_k(x)}{\sum_0^N p_k(x)} \quad (2)$$

Where N is the number of frames so far considered.

$G_k$is the gain parameter which control the learning rate of background modeling. The prior information slowly disappears and new information is learnt slowly by increasing the $G_k$value. The model learning can be made fast by decreasing the G_k value. $G_k$is made adaptive using sigmoid function so that learning is fast initially and slow later on and finally settles to constant after processing many frames.

$$G_k = Gain \times \frac{2}{1+exp(-\frac{cnt-\beta}{\gamma})} \quad (3)$$

The values of $Gain, \beta$ controls the inflection point of the sinusoid function and gradient is controlled by γ. cnt is the continuously increasing value proportional to the frames.

For every new frame, background model is updated and then foreground is extracted by subtraction. After obtaining the foreground, shadow and sudden illumination changes are removed and made suitable for further processing. Shadows are detected in HSV color space since it has the information of hue, saturation and brightness. The presence of shadow for a pixel x is calculated as

$$S(x) = \begin{cases} 1, if\ \alpha \leq \frac{I_K^V(x)}{B_K^V(x)} \leq \beta^{I_K^S(x) - B_K^S(x)} \leq \\ \quad \tau_S \hat{}\, |(I_K^H(x) - B_K^H(x)) \leq \tau_H \\ \quad 0, otherwise \end{cases} \quad (4)$$

I_K^V is the brightness value of current frame and B_K^V is the brightness value of the background frame. I_K^S is the saturation value of current frame and B_K^S is the saturation value of the background frame. I_K^H is the hue value of the current frame and B_K^H is the hue value of the background frame. τ_S and τ_H are the thresholds of hue and saturation. The parameters of α and β are usually between 0 and 1. α is related to the brightness and β is related to light intensity. After the shadow mask is constructed, the pixel value for the places where 1 is set in the mask is brightened. By this way shadows are removed.

For sudden brightness or darkness, fast adaptation is needed in the background model. This is done by initializing the cnt value as below.

$$cnt = \frac{\beta}{2}, \forall\ |\text{mean}(\text{intensity(x)} - \text{mv}_\text{k}| > T \quad (5)$$

$$\text{mv}_\text{k} = (\text{G}_\text{k} - 1) \times \frac{\text{mv}_{\text{k}-1}}{\text{G}_\text{k}} + mean(intensity(x))/\text{G}_\text{k} \quad (6)$$

Where T is the threshold to initialize cnt.

B. Image Categorization

The preprocessed image must be categorized to two types based on the complexity in arriving at boundary box for foreground vehicles.

Type 1 - bounding box estimation would be easy.

Type 2 - bounding box estimation is difficult.

A convolution neural network is trained with traffic images [23] of various vehicle densities and their labels (type 1 or type 2). The five layer convolutional neural network is trained with configuration detailed in Table I.

The trained convolutional neural network is used to classify the preprocessed image to type 1 or type 2.

TABLE I. CONVOLUTIONAL NEURAL NETWORK

Input	200 * 200 resized image
Convolution layers	1st and 2nd layers
Pooling layers	2
Activation function	ReLU
Dropout	1 dropout
Output	2

After resizing the image to 200*200 pixels, the resized image is passed to first convolutional layer. The output from the convolutional layer is passed to max pooling layer to reduce the dimension of features. This process is repeated for all convolutional layers. Over fitting is avoided by adding a dropout in the 4th layer. Classification is done at last layer using Softmax function. The output layer has two neurons each corresponding to one class – one for type 1 and one for type 2. The network will finally output the coordinates, confidence, and category of the object.

C. Vehicle Segmentation

This work proposes two segmentation methods. For type 1 image segmentation, YOLO deep learning network is used for segmenting the vehicles. For type 2 image segmentation, extended topological active mesh net is applied.

Type 1 Segmentation

YOLO v3 network is used for segmentation of type 1 images. YOLOv3 algorithm uses convolutional neural network adopting Darknet-53 network structure to extract features. The input image is split to equal size grids. Presence of object is detected at each grid by YOLOv3. Final bounding around the object is drawn by connecting the neighborhood grids containing the object. A novel part of YOLOv3 is that due to use of direct learning of residuals, training is simplified and detection accuracy increases.

The segmentation flow using YOLOv3 algorithm is shown in Fig. 2.

The final output of the YOLOv3 algorithm is the coordinates of the detected vehicles and the vehicle category in terms of car, bus, truck and motorcycle. From the coordinates bounding box is drawn on the input image and segmentation is done on the bounding boxes to give the vehicle images and their classified type. The vehicle image is passed to fine grained classification to extract further features like color and make.

Type 2 Segmentation

YOLOv3 algorithm alone cannot work for the case of higher density of vehicles and occlusions in the Type 2 images. This work augments YOLOv3 with deformable model based solution using extended topological active net segmentation for type 2 images.

The image is segmented using YOLOv3 as type1 and vehicle bounding box is obtained.

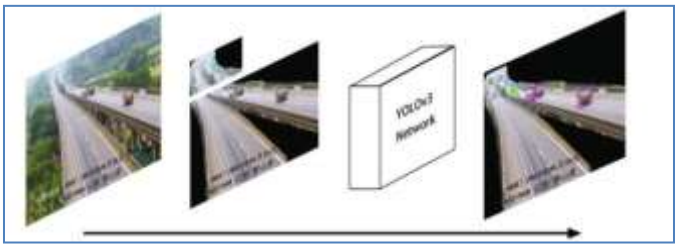


Fig. 2. Type 1 Segmentation.

A mesh is placed over the entire preprocessed image. The mesh portions lying only with the vehicle bounding box found by YOLOv3 is kept and rest of the mesh is removed. Over this image, extended topological active net segmentation is done to arrive at accurate vehicle boundaries. To achieve this, the links in the mesh can be categorized to one of following.

1) Links completely within the object.
2) Links at boundary of object and background.
3) Links at the background.

The links at the boundary must be removed, so that the remaining links represents the object. To speed the process of removing of the links, the links must be first classified. To speed up this classification a Naïve Bayesian classifier is built.

For a link following five features are extracted and they is classified to one of three labels defined above using a trained Naïve Bayes classifier. Following are the features extracted from each link.

1) Local minima of a link (f1).
2) Thickness of a probable edge (f2).
3) LoG value around the starting node (f3).
4) LoG value around the ending node (f4).
5) Difference in dominant color around the two endpoint of the link (f5).

Local minima of the link (AB) are calculated using link features shown in Fig. 3 as follows.

With D as the middle of AB, the horizontal direction HH' is split to equal spaced points along the span of the next neighbor link. If the intensity distribution along S_0,S_1,S_2… are monotonically increasing , the difference between the initial and final sampling point is taken as candidate feature for the direction of DH'. In case S_0, S1, S2… are not monotonically increasing, the candidate feature value for the direction of DH' is taken as 0. Similarly the feature values along DH, DV, DH' is taken and the maximum of these values is taken as local minima of the link.

Thickness of a probable edge is calculated as follows:

For each of axis(DH', DH, DV, DV') the maximum range of monotonically increasing or decreasing value of the sampled points is taken and the minimum of these four values is the indication of thickness of probable edge at the boundary.

LoG is calculated with Laplacian of a Gaussian filter of size $5X5$ around the starting node A and the ending node B. It indicates the presence of edges near the nodes A and B. It is calculated as

$$LoG(x,y) = -\frac{1}{\Pi\sigma^2}\left(1-\frac{x^2+y^2}{2\sigma^2}\right)e^{\frac{x^2+y^2}{2\sigma^2}} \quad (7)$$

For all links in the mesh obtained after YOLOv3, link are categorized using Naïve Bayes classifier and all the links at the boundary are removed.

Each mesh obtained after link refinement is processed for occlusion and clutter removal as follows. Following geometric features are extracted from the mesh.

- Length (l)
- Width (w)
- Area (a)

The extracted features are compared against the known clutters like humans trespassing for similarity matching. In case of similarity, the mesh is not passed to next step for vehicle classification. If not similar, an image with mesh part alone is created and passed to YOLOv3 to arrive at the vehicle category to one of following types- car, bus, truck and motorcycle. The flow of this procedure is given in Fig. 4. Due to this refinement of object boundaries using ETAN segmentation, the clutters affecting the accuracy of vehicle classification are removed in the images.

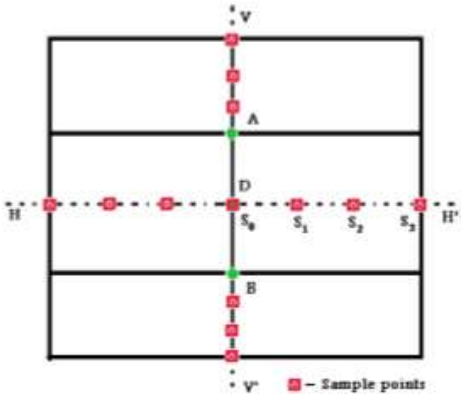


Fig. 3. Link Features.

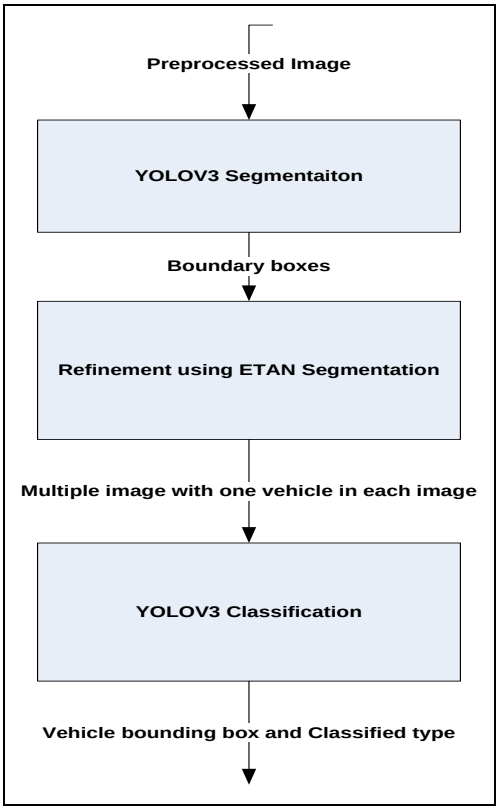


Fig. 4. Type 2 Segmentation Flow.

D. Fine Grained Vehicle Classification

From the individual vehicle detected in earlier process, features are extracted for fine grained classification like color, make of the vehicle. From the vehicle image, following color features are extracted

1) RGB color histogram.
2) Hue histogram.
3) Color moment.

A linear SVM is trained with color features as the input and the vehicle color as the output. The color features extracted from the vehicle is passed as input to the linear SVM as in Fig. 5 to classify the color of the vehicle.

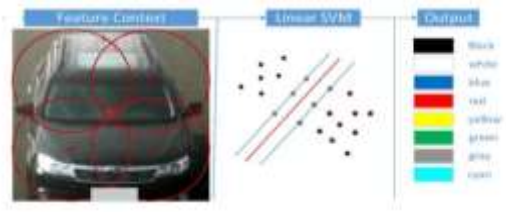


Fig. 5. Vehicle Color Classification.

Bag of SURF feature based make and model recognition approach proposed in [20] is executed on the vehicular image to classify the make and model.

IV. Contribution of the Proposed Solution

The proposed solution has following contributions.

1) A novel adaptive background model proposed in this work with variable gain control is able to remove occlusions like shadows, illuminations from foreground.

2) The vehicle segmentation model to be applied for better segmentation of vehicles is decided based on the density of vehicle distribution in the image. Compared to previous works of applying one particular segmentation model for all density, the proposed work uses different segmentation model based on density distribution.

3) Novel deformation model based segmentation is proposed to mitigate the effects of clutters and occlusion due to colliding vehicle.

V. Result

The performance of the proposed traffic adaptive fine grained vehicular classification is tested against MIO-TCD dataset [22]. The dataset has about 6 lakh images in 11 categories. The image of varying vehicle densities and occlusions are selected for testing. The performance of the proposed solution is compared against YOLO based classification method proposed in [9] and CNN based classification method proposed in [14].

The performance is compared for vehicular classification in terms of following parameters.

1) Precision
2) Recall
3) Accuracy

The performance is measured for a total of 1000 images from the dataset for four different categories of car, bus, truck and motorcycle.

Vehicle categorization accuracy for Bus in the proposed solution is 7.4% higher compared to [9] and 4.3% higher compared to [14]. The comparison of vehicle categorization for bus is shown in Fig. 6. The result of vehicle categorization performance for bus is given in Table II.

TABLE II. Vehicle Categorization Performance for Bus

	[9]	[14]	Proposed
Accuracy	87	90	94
Precision	88	91	96
Recall	89	88	93

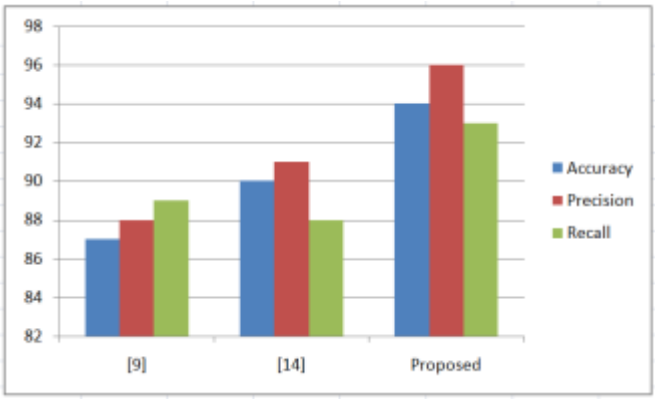


Fig. 6. Comparison of Vehicle Categorization for Bus.

Vehicle categorization accuracy for Truck in the proposed solution is 7.3% higher compared to [9] and 4.2% higher compared to [14] and is given Table III. The comparison of vehicle categorization for truck is shown in Fig. 7.

TABLE III. Vehicle Categorization Performance for Truck

		[9]	[14]	Proposed
Accuracy		88	91	95
Precision		89	92	97
Recall		89	89	94

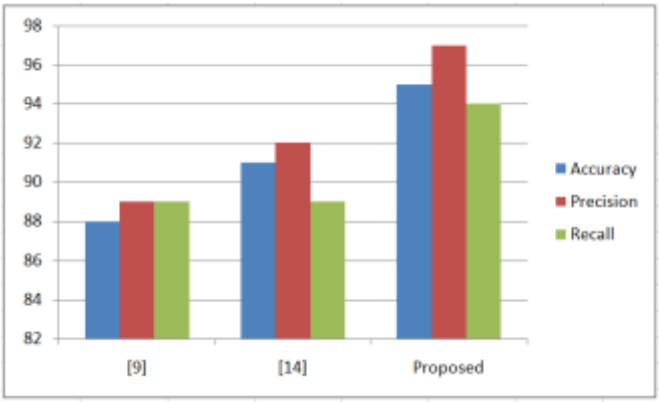


Fig. 7. Comparison of Vehicle Categorization for Truck.

Vehicle categorization accuracy for Car in the proposed solution is 10.63% higher compared to [9] and 9.57% higher compared to [14] and is given Table IV. The comparison of vehicle categorization for car is shown in Fig. 8.

TABLE IV. VEHICLE CATEGORIZATION PERFORMANCE FOR CAR

	[9]	[14]	Proposed
Accuracy	84	85	94
Precision	83	86	96
Recall	84	85	94

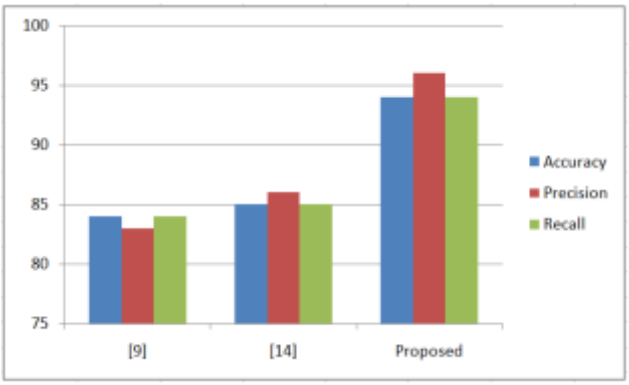


Fig. 8. Comparison of Vehicle Categorization for Car.

Vehicle categorization accuracy for Motorcycle in the proposed solution is 14.13% higher compared to [9] and 11.95% higher compared to [14] and is given Table V. The comparison of vehicle categorization for car is shown in Fig. 9.

TABLE V. VEHICLE CATEGORIZATION PERFORMANCE FOR MOTORCYCLE

	[9]	[14]	Proposed
Accuracy	79	81	92
Precision	79	81	95
Recall	78	78	91

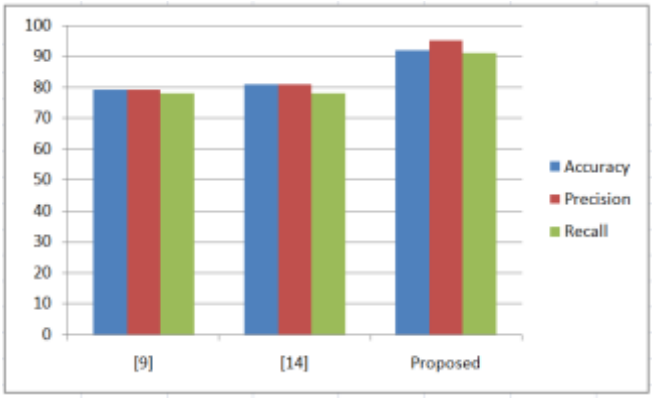


Fig. 9. Comparison of Vehicle Categorization for Motorcycle.

The vehicle categorization accuracy is higher in the proposed solution due to better segmentation of vehicles even in presence of occlusions and clutters. ETAN segmentation is able to accurately detect vehicle boundaries. Due to accurate segmentation of vehicles, the accuracy of deep learning classification has also increased. Due to this reason, the proposed solution performed better than other deep learning classification methods. The results also prove a consistent performance of proposed solution for all types of vehicles. Even for small vehicles, the classification accuracy is higher in proposed solution due to accurate segmentation with ETAN.

VI. CONCLUSION

A traffic adaptive fine grained vehicular classification using deep learning is proposed in this work. The proposed solution is able to solve the problems of shadow, clutter etc. Adaptive background modeling with shadow elimination and dynamic contrast elimination is done as preprocessing. The preprocessed image is classified into two types based on the complexity in arriving at bounding boxes around the vehicles. YOLOV3 deep learning model and its integration with extended topological active net segmentation is followed for vehicle segmentation and vehicle classification. The proposed solution has on average 7.5% more accuracy compared to existing solutions in terms of vehicle classification. Tracking the classified vehicles in successive frames in the video can be considered as the part of future work.

REFERENCES

[1] Kul, Seda&Eken, Süleyman&Sayar, Ahmet. (2017). Distributed and collaborative real-time vehicle detection and classification over the video streams. International Journal of Advanced Robotic Systems. 14. 172988141772078. 10.1177/1729881417720782.

[2] H. Jung, M. Choi, J. Jung, J. Lee, S. Kwon and W. Y. Jung, "ResNet-Based Vehicle Classification and Localization in Traffic Surveillance Systems," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, 2017, pp. 934-940, doi: 10.1109/CVPRW.2017.129.

[3] Boukerche, Azzedine& Siddiqui, Abdul &Mammeri, Abdelhamid. (2017). Automated Vehicle Detection and Classification: Models, Methods, and Techniques. ACM Computing Surveys. 50. 1-39. 10.1145/3107614.

[4] Xiao, Wen &Vallet, Bruno & Schindler, Konrad &Paparoditis, Nicolas. (2016). Street-side vehicle detection, classification and change detection using mobile laser scanning data. ISPRS Journal of Photogrammetry and Remote Sensing. 114. 166-178. 10.1016/j.isprsjprs.2016.02.007.

[5] L. Yang, P. Luo, C. C. Loy and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 3973-3981, doi: 10.1109/CVPR.2015.7299023.

[6] W. Maungmai and C. Nuthong, "Vehicle Classification with Deep Learning," *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, Singapore, 2019, pp. 294-298, doi: 10.1109/CCOMS.2019.8821689.

[7] Nam, Y., Nam, Y. Vehicle classification based on images from visible light and thermal cameras. J Image Video Proc. 2018, 5 (2018). https://doi.org/10.1186/s13640-018-0245-2.

[8] Sarikan, Selim&Ozbayoglu, Murat &Zilci, Oguzhan. (2017). Automated Vehicle Classification with Image Processing and Computational Intelligence. Procedia Computer Science. 114. 515-522. 10.1016/j.procs.2017.09.022.

[9] Song, H., Liang, H., Li, H. et al. Vision-based vehicle detection and counting system using deep learning in highway scenes. Eur. Transp. Res. Rev. 11, 51 (2019). https://doi.org/10.1186/s12544-019-0390-4.

[10] Psyllos, Apostolos&Anagnostopoulos, Christos-Nikolaos &Kayafas, Eleftherios. (2010). Vehicle Logo Recognition Using a SIFT-Based Enhanced Matching Scheme. Intelligent Transportation Systems, IEEE Transactions on. 11. 322 - 328. 10.1109/TITS.2010.2042714.

[11] P. Chen, X. Bai, and W. Liu, "Vehicle Color Recognition on Urban Road by Feature Context.", IEEE Transactions on Intelligent Transportation Systems, 2014.

[12] Pawel Badura and Maria Skotnicka. 2014. Automatic car make recognition in low-quality images. In Information Technologies in Biomedicine, Volume 3, EwaPietka, Jacek Kawa, and WojciechWieclawek (Eds.). Advances in Intelligent Systems and Computing, Vol. 283. Springer International Publishing, 235–246. DOI:http://dx.doi.org/10.1007/ 978-3-319-06593-9_21.

[13] Z. Dong, Y. Wu, M. Pei, and Y. Jia. 2015. Vehicle type classification using a semi supervised convolutional neural network. IEEE Trans. Intell. Transport. Syst. 99 (2015), 1–10.

[14] Roecker, Max & Costa, Yandre& Almeida, Joao & Matsushita, Gustavo. (2018). Automatic Vehicle type Classification with Convolutional Neural Networks. 1-5. 10.1109/IWSSIP.2018.8439406.

[15] W. Sun, X. Zhang, S. Shi, J. He, and Y. Jin, "Vehicle type recognition combining global and local features via two-stage classification," Mathematical Problems in Engineering, vol. 2017, November 2017.

[16] S. Bai, Z. Liu, and C. Yao, "Classify vehicles in traffic scene images with deformable part-based models," Machine Vision and Applications, pp. 1–11, November 2017.

[17] M. Liang, X. Huang, C. Chen, X. Chen and A. Tokuta, "Counting and Classification of Highway Vehicles by Regression Analysis," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2878-2888, Oct. 2015, doi: 10.1109/TITS.2015.2424917.

[18] M. Kafai and B. Bhanu. 2012. Dynamic bayesian networks for vehicle classification in video. IEEE Trans. Industr. Info. 8, 1 (Feb 2012), 100–109.

[19] Haoyu Peng, Xun Wang, Huiyan Wang, and Wenwu Yang. 2015. Recognition of low-resolution logos in vehicle images based on statistical random sparse distribution. IEEE Trans. Intell. Transport. Syst. 16, 2 (April 2015), 681–691. DOI:http:// dx.doi.org/10.1109/TITS.2014.2336675.

[20] A. J. Siddiqui, A. Mammeri, and A. Boukerche. 2016. Real-time vehicle make and model recognition based on a bag of SURF features. IEEE Trans. Intell. Transport. Syst. 17, 11 (Nov 2016), 3205–3219. DOI:http://dx.doi.org/10.1109/TITS. 2016.2545640.

[21] Shobha B S and Deepu R, "A Review on Video Based Vehicle Detection, Recognition and Tracking", 3rd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions 2018.

[22] https://www.kaggle.com/yash88600/miotcd-dataset-50000-imagesclassification.

[23] https://www.aicitychallenge.org/2021-data-and-evaluation/.

Wide Area Measurement System in the IEEE-14 Bus System using Multiobjective Shortest Path Algorithm for Fault Analysis

Lilik J. Awalin[1]*, Syahirah Abd Halim[2], Jafferi Bin Jamaludin[3], Nor Azuana Ramli[4]
Faculty of Advanced Technology and Multidiscipline, Airlangga University[1]
Gedung Kuliah Bersama, Kampus C Unair, Jl. Mulyorejo, Surabaya 60155, Indonesia[1]
Department of Electrical, Electronic & Systems Engineering, Faculty of Engineering & Built Environment[2]
Universiti Kebangsaan Malaysia, Selangor, Malaysia[2]
UM Power Energy Dedicated Advanced Centre (UMPEDAC), Kuala Lumpur, Malaysia[3]
Centre For Mathematical Sciences, Universiti Malaysia Pahang),Gambang, Pahang, Malaysia[4]

***Abstract*—In a large-scale interconnected power system network, few challenges exist in evaluating and maintaining the overall system stability. The power system's ability to supply all types of loads during natural disasters or faults has yet to be addressed. This work focuses on developing a wide-area measurement system to manage and control the power system under all operating conditions. The IEEE-14 bus system was modeled in PSCAD software for simulating nineteen types of fault based on multi-objective shortest path algorithm. To manage the wide area measurement, the research must comprehend the working principle of the multi-objective shortest path algorithm, whereby the proposed method will determine the new path for the IEEE-14 bus system. To evaluate the performance of the multi-objective shortest path algorithm, all sections of the IEEE-14 bus system were simulated with faults. The distances of the normal path (without simulated fault) and the new path (with simulated fault) were recorded. Based on the recorded data, it was found that the location of the fault has significant influence on the shortest path of the buses connected to each other.**

***Keywords*—IEEE-14 bus system; wide area measurement; multi-objective shortest path algorithm; fault location; PMU**

I. Introduction

SCADA system used for real-time monitoring of a power system network is not alone sufficient to provide data related to dynamic performance of the overall system in steady state condition. The actual performance of the system can be predicted by utilising any power system simulation tools, which normally have few limitations in term of parameters exactness and generalized estimation. Therefore, the system operator must be ensuring of consistent and detailed dynamic information in order to decide on how to deal with the system during specific conditions and energy security.

The technique is based on close observation on the angle and phasors with respect to time duration in dynamic state by using phasor measurement unit (PMU). PMU will show the dynamic information of individual buses in the system for the system operator. This information will be useful in determining how to respond to the system's condition in a timely manner. The PMU can also be used as an investigator apparatus to monitor performance of hardware in the system, such as generator controller. Additionally, the system operator needs to utilize the PMU in managing a wide area monitoring system as the one of the approaches in smart grid operation to provide the required data.

The operator must also understand the overall system behavior during dynamic phenomena based on the data obtained from the wide area monitoring system. The PMU, which is part of the wide area monitoring system, normally consists of standard components such as communication system, data analyst and information display. Placement of the PMU in the power system is dependent on a preliminary study of the system's constraints and stability problems that might possibly occur in the wide area monitoring system.

In [1], the system operator also supervises any event in the system for the purpose of monitoring and analyzing the dynamic behavior of the power system during different scenarios. The system protection, dependability and stability will be considered as well in the development process of the wide area monitoring system for the power grid. This demonstrates the significance of the wide area monitoring system in ensuring reliable operation of the power grid and improved system security. In return, this approach can help to improve the overall technical aspects by optimising the stability of the overall system and reducing the operational cost.

II. Literature Review

A. Monitoring System

An electric power system is made up of a number of interconnected components that serve a purpose of delivering reliable power supply at a low cost. The majority of the world's population is already reliant on electrical energy for a better quality of life and economic growth. However, the power system is always vulnerable to internal and external disturbances that can disrupt the system operational capability,

*Corresponding Author

such as lightning, electromagnetic interference, and equipment malfunctions.

However, in a very large, interconnected power system, several challenges exist in assessing and maintaining the stability of the entire system as mention in [2]. Another issue in the power system related to large penetration of renewable energy sources has recently aroused, contributing to additional risk on the stability which requires extensive monitoring on the system operation. However, the development of renewable energy sources is critical for energy security. Consequently, power system monitoring should be a key aspect in accomplishing adaptable tasks to maintain system dependability and reliability of electricity supply.

In recent years, advancements in information and communication technology have enabled greater adaptability in wide-area power system monitoring in terms of rapid and massive data transmission. As one of the smart grid innovations in the power grid, the wide-area monitoring system with phasor measurement units is a promising solution towards the improvement of overall system operation, as mention in [3].

B. Optimal Placement of PMU using Integer Linear Programming

Power engineers are becoming interested in phasor measurement units (PMU) since they can provide synchronized estimates of real-time phasors of voltage and current. Furthermore, the use of PMUs has also changed the way state estimation is being performed. A state estimator plays a vital role in the security of power system operations as it is used for online monitoring, analysis and control.

According to [4], an ideal PMU placement algorithm was developed to recover the bad data processing capability of state estimation by taking advantage of the PMU technology. Techniques for identifying placement sites for phasor measurement units in a power system based on incomplete observability are presented in [5], where simulated annealing method was used to solve the pragmatic communication-constrained in PMU placement problem. The authors in [6]-[10] developed an optimal placement algorithm for PMUs by using integer programming. However, the proposed integer programming becomes a nonlinear integer programming under the existence of conventional power flow or power injection measurements. Besides, a similar formulation of optimal PMU placement problem is proposed by integer linear programming. According to [11]-[12], there is two proposed formulation which is without conventional measurements and with conventional measurements. Referring to [11], simulation results show that the proposed algorithm demonstrates computational efficiency and can be used in actual practical system. However, the simulation was only performed on the IEEE-14 bus system due to computational limitations.

C. Dynamic Shortest Path Algorithm

Wide area monitoring system normally consists of three main elements, namely management, measurement, and communication. It is critical to plan these elements carefully for a power system to function well. According to [13]-[15], measurement and communication elements in a wide area network are planned flexibly from an administrative standpoint to achieve a suitable degree of system monitoring.

The optimal placement of phasor measurement units is determined using an integer linear programming (ILP) layout technique that takes zero-injection bus effects into consideration. The PMU location problem is solved using the integer linear programming (ILP) technique with and without conventional estimations, as well as maximum estimation redundancy across all buses. According to [16]-[19], a new decision on discernibleness limitations was proposed to address the optimal positioning issue, which in turn might reduce the number of required PMUs.

According to publications [20]-[21], a new decision method based on identifiable limits that could reduce the number of PMUs has been addressed to solve the optimal PMU position problem.

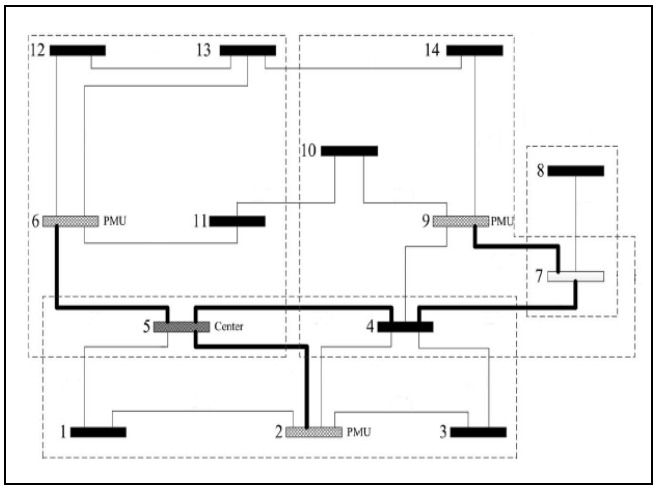


Fig. 1. Optimal Design of Wide Area Monitoring System for the IEEE 14 Bus System [22].

In [22], the PMU placement was designed as illustrated in Fig. 1. For the optimal structure of the communication framework, a dynamic multi-objective shortest path (MOSP) routing technique was implemented. As a result, the proposed technique optimizes PMU location initially, then enhances communication methods that connect all phasor measurement units. In the MOSP steering algorithm, which is used for optimal communication interface layout, the shortest path exploration algorithm is used to select the shortest pathways from phasor measurement units' buses to the central control bus. As a result, all the evaluated PMU nodes in set S carried by author in [22] may be measured as original nodes. It was also discovered that, of all the possible routes, the nearest phasor measurement units' bus has consistently the shortest route to the central control bus.

There are two ways of acquiring transmission path information from the phasor measurement units' bus to central control bus, according to [22]; single objective shortest path and multi objective shortest path. The shortest path from each single objective shortest path bus to the central control bus can be obtained directly using a shortest path algorithm. Meanwhile, in a multi objective shortest path algorithm direct

calculation, overlapped ways are used to obtain transmission path information for all phasor measurement units' buses. Therefore, all phasor measurement units' buses, except for the closest one to the CB, can be determined using the multi objective shortest path algorithm, owing to the overlapped ways that can be used as a method for information transmission with reduced optical fiber ground wire coverage.

III. Methodology

Typically, a wide area monitoring system is made up of three components: management, measurement, and communication. It is necessary to build these infrastructures adequately for the optimal operation of the power system. To achieve a satisfactory level of system observability, the measurement and communication infrastructures in a wide area network are planned independently from a management perspective. There have been numerous procedures taken in order to develop the 14-bus test system for this study. In the first step, integer linear programming was chosen to determine the optimal configuration of measuring mechanisms [22]. The dynamic multi-objective shortest path was used in the next step to come up with the best communication infrastructure option. The fault cases were assigned to monitor the performance of the multi-objective shortest path, in which 19 fault cases were simulated. For each case, the distance calculated using the multi-objective shortest route technique was measured and compared.

The flow chart involved in implementing the proposed technique is depicted in Fig. 2. The initial work involves modelling in PSCAD software to develop the IEEE 14-bus network as the test system. The PMU's position was then determined using integer linear programming (ILP). According to [11], the placement of PMUs is solved using binary integer programming in MATLAB. However, due to research limitations, this study did not develop the algorithm for PMU placement and instead relied on [22] for PMU placement using ILP on the IEEE-14 bus system.

To solve the optimum phasor measurement unit location problem, three PMU will be installed in buses 2, 6, and 9. In order to observe the performance of the multiobjective shortest path algorithm, all of the section was assigned by fault. The location of fault is listed in Table I as follows.

The following step is to assign different 19 fault cases at IEEE-14 bus system. To test the performance of the multi-objective shortest route method, each fault was allocated to various places. Fig. 3 shows the location of every fault case that was allocated in the IEEE-14 bus system. Each fault case was simulated using MATLAB software. For each fault condition, the algorithm was modified to evaluate and choose the shortest path. The result will show the distance between each fault condition and other buses. Every distance recorded by each fault case to connect to the other bus using the shortest path was then determined. Data from the 19 failure cases were recorded, and the graphs were plotted to acquire more information regarding the multi-objective shortest route algorithm's performance.

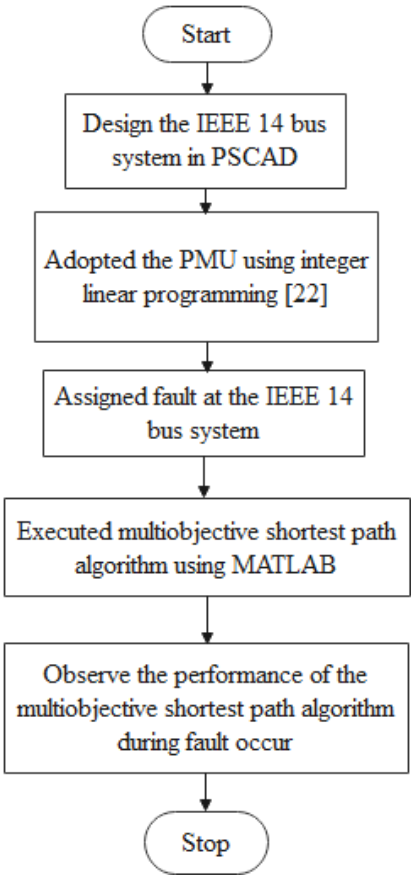


Fig. 2. Flowchart.

TABLE I. List of Fault Location

No. Fault	location	Fault	Location
1	Bus 9 to 10	11	Bus 7 to 9
2	Bus 10 to 11	12	Bus 4 to 9
3	Bus 6 to 11	13	Bus 5 to 6
4	Bus 6 to 13	14	Bus 1 to 5
5	Bus 6 to 12	15	Bus 1 to 2
6	Bus 12 to 13	16	Bus 5 to 2
7	Bus 13 to 14	17	Bus 2 to 4
8	Bus 9 to 10	18	Bus 2 to 3
9	Bus 4 to 9	19	Bus 3 to 4
10	Bus 4 to 7		

IV. Result and Discussion

In the IEEE-14 bus system, 19 faults were assigned at different locations and tested for data collection based on the utilisation of multi objective shortest route algorithm. Each fault was then simulated using a multi-objective shortest path method in MATLAB to determine the shortest path length. All of the data in this section was collected by simulation, and was evaluated by plotting graphs for further discussion.

A. Fault Simulation on Section

The IEEE-14 bus system was used to simulate 19 fault scenarios. Each type of fault was simulated between buses and different locations. Table II shows the details of the fault simulation point. Based on the table, the fault point was simulated in each section (between bus to bus). For example, fault 1 was simulated between buses 9 and 10, with a distance of 18.9 kilometers between them. All the fault points were simulated in the middle of the section.

B. Path Analysis

Path analysis is discussed in this section with different fault simulation points considered. Every fault point will be analyzed, with detailed results and explanations. For each fault case, the result will be compared to the normal path and the new path. Before the fault occurs, the multi-objective shortest path algorithm calculates the normal path distance. Meanwhile, during the fault, the multi-objective shortest path algorithm calculates the new path distance. There were six paths that connect the buses for each fault case.

Table III shows the route taken to connect to other buses, as well as the distance used to calculate fault 1. The first fault occurs in the middle of the section between buses 9 and 10. If the fault occurs at the path connecting the phasor measurement unit and the buses, the path should be changed. The path distance is referred to as the main route in normal conditions (without fault), as shown in Table III.

TABLE II. FAULT POINT

No.	Fault Location	Distance between bus (Km)	Distance of Fault (Km)
1	Bus 9 to 10	18.9	9.45
2	Bus 10 to 11	42.9	21.45
3	Bus 6 to 11	44.4	22.2
4	Bus 6 to 13	29.1	14.55
5	Bus 6 to 12	57.1	28.55
6	Bus 12 to 13	44.7	22.35
7	Bus 13 to 14	77.8	38.9
8	Bus 9 to 14	60.4	30.2
9	Bus 4 to 9	124	62
10	Bus 4 to 7	46.7	23.35
11	Bus 7 to 9	24.6	12.3
12	Bus 4 to 5	9.4	4.7
13	Bus 5 to 6	56.3	28.15
14	Bus 1 to 5	24.9	24.9
15	Bus 1 to 2	13.2	6.6
16	Bus 2 to 5	38.9	19.45
17	Bus 2 to 4	39.4	19.7
18	Bus 2 to 3	44.2	22.1
19	Bus 3 to 4	38.2	19.1

TABLE III. PATH DISTANCE FOR FAULT 1

Route	**Main route**	**Normal path (Km)**	**Route after fault**	**New path (Km)**
1	5 to 2	38.9	5 to 2	38.9
2	5 to 6	56.3	5 to 6	56.3
3	5 to 4 to 7 to 9	80.7	5 to 4 to 7 to 9	80.7
4	6 to 5 to 2	95.2	6 to 5 to 2	95.2
5	6 to 11 to 10 to 9	160.2	6 to 5 to 4 to 7 to 9	137
6	2 to 4 to 7 to 9	110.7	2 to 4 to 7 to 9	110.7

Table III is made up of five columns. The number of the route appears first, followed by the main route (normal condition of the route). The distance for the normal path is represented in the third column. The fourth column represents the route after a fault, and the last column represents the new path's distance. According to Table III, the distance between bus 5 and bus 2 for route 1 is 39.8 kilometers. When fault 1 occurred, no change was observed in the route condition and distance. Change in route number 5, which is the main route path, can be seen starting from bus 6, continuing to bus 11, bus 10, and bus 9. The simulated result for the main path was 106.2 kilometers. Based on the observation, the new path distance was recorded at 137 kilometers. This value differs from the normal path by 30.8 kilometers. It means that the multi-objective method for determining the new path of the distance has a big impact on path distance effectiveness. Route 5 demonstrated a different distance between the main route and the route after the fault occurs, as shown in Fig. 3. The normal distance for Route 5 was 106.2 kilometers, but after the fault, the distance increased to 137 kilometers. This is due to the fact that the fault 1 occurs between buses 10 and 11. The algorithm must calculate the shortest path to connect buses 6 and 9 since the main route used is between bus 10 and 11. The algorithm looked at every possible path and chose the shortest one. The path for route 5 was made up of buses 6, 5, 4, 7, and 9. Since the fault does not interrupt the path, the other route is unaffected.

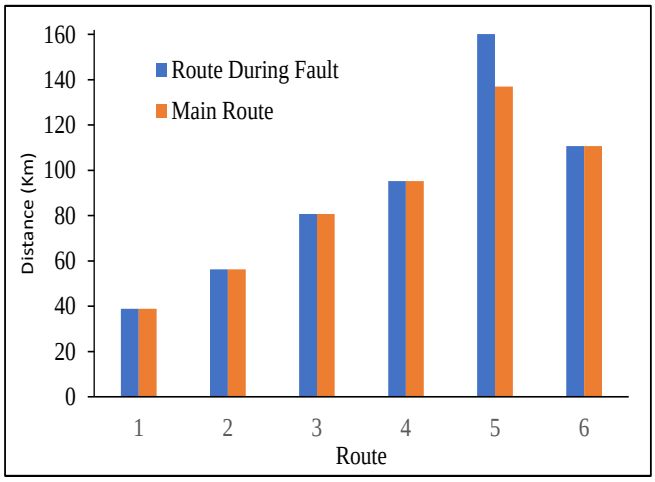


Fig. 3. Distance against Route for Fault 1.

C. Distance Comparison between every Fault for each Route

Distance comparison is discussed in this section for each fault case on each route. An example of a distance comparison between each route's faults is presented in Table IV. The condition of the phasor measurement unit at bus 2 and the center at bus 5 is represented by Route 1.

When fault 1 occurs between buses 9 and 10, it resulted in a path distance of 38.9 kilometers. Except for the distance for fault case 16, all of the distances for route 1 are the same. Since the fault occurs between bus 2 to bus 5, the distance for fault case 16 is 48.8 kilometers. As a result, the path from bus 5 to bus 2 cannot be used, and the multi-objective shortest path algorithm must find a new shortest path. Other faults were unaffected because the path that connects bus 5 and bus 2 was not interrupted.

TABLE IV. DISTANCE FOR ROUTE 1

No. of fault	Distance	No. of fault	Distance
1	38.9 Km	11	38.9 Km
2	38.9 Km	12	38.9 Km
3	38.9 Km	13	38.9 Km
4	38.9 Km	14	38.9 Km
5	38.9 Km	15	38.9 Km
6	38.9 Km	16	48.8 Km
7	38.9 Km	17	38.9 Km
8	38.9 Km	18	38.9 Km
9	38.9 Km	19	38.9 Km
10	38.9 Km		

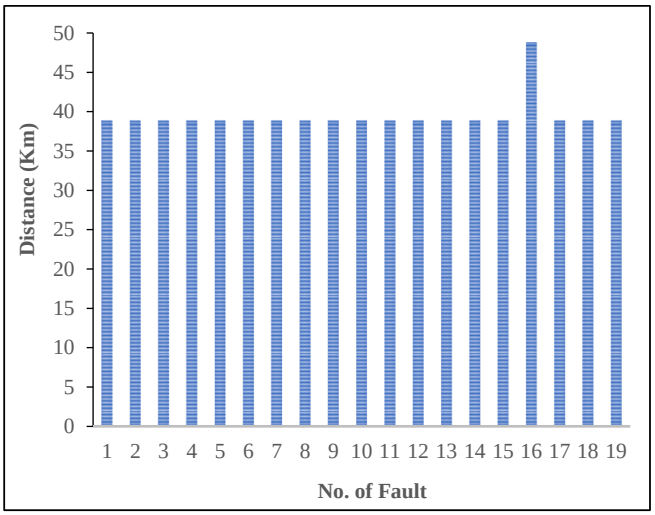


Fig. 4. Distance against Fault for Route 1.

Fig. 4 depicts the distance between bus 5 and bus 2 during a fault occurrence. On the IEEE-14 bus system, 19 different faults were assigned. Except for fault number 16, every fault case takes the same path to connect bus 5 to bus 2 and the path distance recorded is 38.9 kilometers. As the fault happens between buses 5 and 2, the algorithm calculates various paths for fault 16. Although bus 2 is another available option of path, the algorithm needs to choose different path since the path between bus 5 and 2 cannot be used. Therefore, the algorithm then assigns another path, which is between bus 5 to 4.

V. CONCLUSION

This work presents a new method based on multi-objective shortest path algorithm for determining a new path distance of connected buses when faults occur at various locations within a test system. The IEEE-14 bus system was tested with nineteen different types of faults to investigate the performance of the proposed algorithm during the fault events. The path distances were measured and the shortest path for every fault case was then determined. Based on the results, it was found that the shortest path of the buses connected to each other is significantly influenced by the location of the fault. In the event of a fault, the proposed algorithm will choose a new path based on the shortest distance, excluding the old route, which is already affected by the fault. Based on the obtained results, the fault occurrences have significantly affected the main route consisting of routes 6, 11, 10 and 9. It can be concluded that the multiobjective shortest path algorithm is capable of estimating the possibility of new path distance during a fault in the IEEE 14-bus system.

ACKNOWLEDGEMENT

The authors thanks to the Airlangga University, Indonesia and 2021 SATU Joint Research Scheme (JRS) no UM67.

REFERENCES

[1] Venkatasubramanian, Vaithianathan, et al. "Wide-area monitoring and control algorithms for large power systems using synchrophasors." *2009 IEEE/PES Power Systems Conference and Exposition*. IEEE, 2009.

[2] Gore, Rahul, and Mallikarjun Kande. "Analysis of wide area monitoring system architectures." *2015 IEEE International Conference on Industrial Technology (ICIT)*. IEEE, 2015.

[3] Bose, Anjan. "Smart transmission grid applications and their supporting infrastructure." *IEEE transactions on Smart Grid* 1.1 (2010): 11-19.

[4] Chen, Jian, and Ali Abur. "Placement of PMUs to enable bad data detection in state estimation." *IEEE Transactions on Power Systems* 21.4 (2006): 1608-1615.

[5] Nuqui, Reynaldo F., and Arun G. Phadke. "Phasor measurement unit placement techniques for complete and incomplete observability." *IEEE Transactions on Power Delivery* 20.4 (2005): 2381-2388.

[6] Xu, Bei, and Ali Abur. "Observability analysis and measurement placement for systems with PMUs." *IEEE PES Power Systems Conference and Exposition, 2004.*. IEEE, 2004.

[7] Almunif, Anas, and Lingling Fan. "Mixed integer linear programming and nonlinear programming for optimal PMU placement." *2017 North American Power Symposium (NAPS)*. IEEE, 2017.

[8] Li, Yikui, Jie Li, and Lei Wu. "A Novel Integer Linear Programming Based Optimal PMU Placement Model." *2018 North American Power Symposium (NAPS)*. IEEE, 2018.

[9] Ahmed, Muhammad Musadiq, and Kashif Imran. "An Optimal PMU Placement Against N-1 Contingency of PMU Using Integer Linear Programming Approach." *2019 International Conference on Applied and Engineering Mathematics (ICAEM)*. IEEE, 2019.

[10] Reddy, K. SivaRama Krishna, et al. "Implementation of Integer Linear Programming and Exhaustive Search algorithms for optimal PMU placement under various conditions." *2015 IEEE Power, Communication and Information Technology Conference (PCITC)*. IEEE, 2015.

[11] Gou, B. (2008). Optimal placement of PMUs by integer linear programming. IEEE Transactions on power systems, 23(3), 1525-1526.

[12] Chakrabarti, S., & Kyriakides, E. (2008). Optimal placement of phasor measurement units for power system observability. IEEE Transactions on power systems, 23(3), 1433-1440.

[13] Bashian, Amir, et al. "Optimal design of a wide area measurement system using hybrid wireless sensors and phasor measurement units." *Electronics* 8.10 (2019): 1085.

[14] Singh, Satyendra P., and Shiv P. Singh. "Optimal cost wide area measurement system incorporating communication infrastructure." *IET Generation, Transmission & Distribution* 11.11 (2017): 2814-2821.

[15] James, J. Q., et al. "A unified framework for wide area measurement system planning." *International Journal of Electrical Power & Energy Systems* 96 (2018): 43-51.

[16] Martin, Kenneth E. "Synchrophasor measurements under the IEEE standard C37. 118.1-2011 with amendment C37. 118.1 a." *IEEE Transactions on Power Delivery* 30.3 (2015): 1514-1522.

[17] De La Ree, Jaime, et al. "Synchronized phasor measurement applications in power systems." *IEEE Transactions on smart grid* 1.1 (2010): 20-27.

[18] Phadke, Arun G., and James S. Thorp. *Synchronized phasor measurements and their applications*. Vol. 1. New York: Springer, 2008.

[19] Balu, Neal J., Mark G. Lauby, and P. Kundur. "Power system stability and control." *Electrical Power Research Institute, McGraw-Hill Professional* (1994).

[20] Rashidi, Farzan, et al. "Optimal placement of PMUs with limited number of channels for complete topological observability of power systems under various contingencies." *International Journal of Electrical Power & Energy Systems* 67 (2015): 125-137.

[21] Abiri, Ebrahim, Farzan Rashidi, and Taher Niknam. "An optimal PMU placement method for power system observability under various contingencies." *International transactions on electrical energy systems* 25.4 (2015): 589-606.

[22] Ghasemkhani, A., Monsef, H., Rahimi-Kian, A., & Anvari-Moghaddam, A. (2017). Optimal design of a wide area measurement system for improvement of power network monitoring using a dynamic multiobjective shortest path algorithm. IEEE Systems Journal, 11(4), 2303-2314.

An Enhanced Feature Acquisition for Sentiment Analysis of English and Hausa Tweets

Amina Imam Abubakar[1]*, Abubakar Roko[2], Aminu Muhammad Bui[3], Ibrahim Saidu[4]
Department of Computer Science, Usmanu Danfodiyo University, Sokoto, Nigeria[1, 2, 3]
Department of ICT, Usmanu Danfodiyo University, Sokoto, Nigeria[4]

***Abstract*—Due to the continuous and rapid growth of social media, opinionated contents are actively created by users in different languages about various products, services, events, and political parties. The automated classification of these contents prompted the need for multilingual sentiment analysis researches. However, the majority of research efforts are devoted to English and Arabic, English and German, English and French languages, while a great share of information is available in other languages such as Hausa. This paper proposes multilingual sentiment analysis of English and Hausa tweets using an Enhanced Feature Acquisition Method (EFAM). The method uses machine learning approach to integrate two newly defined Hausa features (Hausa Lexical Feature and Hausa Sentiment Intensifiers) and English feature to measure classification performance and to synthesize a more accurate sentiment classification procedure. The approach has been evaluated using several experiments with different classifiers in both monolingual and multilingual datasets. The experimental results reveal the effectiveness of the approach in enhancing feature integration for multilingual sentiment analysis. Similarly, by using features drawn from multiple languages, we can construct machine learning classifiers with an average precision of over 65%.**

Keywords*—Multilingual sentiment analysis; sentiment analysis; social media; machine learning

I. Introduction

Social media have turned the web into a vast source of information that is generated by users about all kinds of topics. Twitter is considered one of the most popular and commonly used social media platform [1] where users communicate with each other, share their opinions and express their emotions (sentiments) in the form of convenient short blogs using limited words [2]. Due to the large volume of information, automated approaches that allow users to effectively interact with opinionated content [3] on the internet have been developed [4]. Such approaches form the field of sentiment analysis. Sentiment analysis represents the process of automatically extracting the sentiment orientation or polarity of an opinion on a specific object [5].

The majority of current sentiment analysis systems address a single language, usually English [4] [6]-[13] and analyzing sentiment in a single language increases the risks of missing essential information in texts written in other languages. However, Twitter users express their opinions in different languages such as Arabic, Spanish, German, French, and Hausa. This prompted the need for sentiment analysis systems that discover sentiment from a Twitter document made up of English and one other language. Such systems are called multilingual sentiment analysis systems. These systems are motivated in building sentiment analysis approaches for different languages [14]. While research on multilingual sentiment analysis has been done in several languages e.g English and Arabic [2], English, German, French and Portuguese [15], Italian, Spanish, French and German [16], Hindi, Telgu, and Tamil [17], none has been extended to Hausa language despite the popularity of the language as one of the most spoken language in Africa [18] and therefore, receive little attention in Natural Language Processing (NLP) task. Similarly, lack of NLP application for a language can deny its speakers the potential benefits of NLP technology and information access.

In this paper, multilingual sentiment analysis of English and Hausa tweets using an Enhanced Feature Acquisition Method (EFAM) is proposed. The method uses machine learning approach to integrate English feature and Hausa features to measure classification performance and to synthesize a more accurate sentiment classification procedure.

The main contribution of this study is the development of two newly defined Hausa features; Hausa Lexical Feature (HLF) and Hausa Sentiment Intensifiers (HSI). These features will determine if the frequency of Hausa words and Hausa intensifiers has any effect on a particular sentiment in a multilingual context.

The paper is organized as follows. Section 2 describes the related works, Section 3 discusses the proposed methodology for multilingual sentiment analysis, Section 4 describes the experiment, results and discussion, and Section 5 gives the conclusion and future work.

II. Related Work

Much research have been put into developing approaches for multilingual sentiment analysis of Tweets. These approaches are aimed at creating Twitter sentiment classification models using multiple languages.

The author in [19] proposed the use of emotion tokens for multilingual Twitter messages for English and non-English languages. The polarities of the tokens are labelled automatically based on their popular co-occurrences of emotions. Using a graph propagation algorithm, they construct a graph whose vertices are regular words and emotion tokens while the weight of edges gives a measure of co-occurrence. The comparative evaluations indicate that the emotion tokens are independent of the tweet for both English and non-English Twitter messages and achieve a better performance than the

*Corresponding Author

traditional semantic-based approach [20]. However, the propagation process assigns large positive scores for the majority of the tokens, and that negative scores do not contain many emotion tokens, resulting in a low recall rate on negative scores, especially for the English language.

The author in [15] examined the characteristics and feasibility of a language-independent, semi-supervised sentiment classification approach for tweets and use emoticons as noisy labels to generate training data from a completely raw set of tweets. Class probabilities for the polarities are calculated using logarithmic probabilities. The approach was evaluated in four different languages (English, German, French and Portuguese) that were manually annotated. They used a method similar to [21] to assign noisy polarity class labels to tweets based on the existence of positive or negative emotions. The evaluation performance for each of the 4 languages shows that the approach is less fit to classify some languages because of their structural differences. Therefore, unique impacts of different languages are needed for a proper classification approach.

The author in [16] presented a method to create a sentiment analysis system for tweets in English using tweets from SemEval 2013 [22] as a training and testing dataset. Using the Google machine translation system, the tweets were translated to four other languages; Italian, Spanish, French and German and are manually corrected to create gold standards for each target language. The result shows that the use of all the languages together improves the overall sentiment classification of sentiment in the data. While their system is found effective in the multilingual classification aspect, it, however, cannot eliminate the problem of translation errors due to differences in a language context.

The author in [14] analyzed a large set of manually labelled tweets to train sentiment classifiers in 13 European languages. The performance of these classifiers and the quality of human labelling are performed with the construction of automated classification models. The classification models depend much more on the quality and size of training data than on the type of model trained. While the performance of these models indicates that humans perceived the sentiment classes as ordered, it is, however, limited by the quality of the labelled data used.

The author in [17] proposed a sentiment analysis system of a very famous Indian movie *Baahubali2* using Twitter comments and posts. The authors use Hindi, Telgu, and Tamil languages which are converted to English language using Google translator. A classification algorithm was implemented for all the language datasets and processes each word in a tweet and store the score (positive, very positive, negative, very negative, and neutral) into Hadoop distributed file system. The proposed method effectively demonstrates the relation of positive, very positive, and neutral tweets which are strongly correlated with each other. However, the positive and very positive parts of the tweets are heavily influenced by the noises present in the dataset.

The author in [2] proposed a Vector Space Model (VSM) approach in handling tweets in both Arabic and English languages with different processing techniques applied. This approach is based on using the Term Frequency Inverse Document Frequency (TF-IDF) to generate the feature vector for the classification process. Experiments were performed on five datasets; two in Arabic and three in English and the performance of seven classification algorithms were analyzed. The experimental results reveal the effectiveness of the approach with a higher classification accuracy when applied to the English dataset than Arabic. However, extracting Arabic feature vectors from Arabic WordNet will have served as an additional feature for the Arabic dataset and thus, add classification performance.

Thoughtfully learning the literature, there is no existing work on multilingual sentiment analysis of Hausa language. This study is the first contribution on NLP for Hausa language.

III. Proposed Method for Multilingual Sentiment Analysis

This section describes the research workflow illustrated in Fig. 1 which comprises the dataset used (Twitter multilingual corpus and HWN), pre-processing methods, feature engineering, classification methods, and evaluation. The components are elucidated as follows.

A. Dataset Description

This study makes use of two resources for multilingual sentiment classification: Hausa WordNet lexical resource and Twitter multilingual corpus.

1) Hausa wordnet lexical resource: Hausa is a language spoken by more than 25 million people representing the original Hausa population [23]. The language stretches across the northern states of Nigeria, southern Niger and Hausa communities in Sudan. It is also spoken as a first language by scattered settlements throughout West Africa and as a second language by millions of non-Hausas in northern Nigeria and the northern parts of Benin, Togo, and Ghana [24]. However, despite the popularity of the language, there are no sufficient tools and resources for various NLP applications, hence, the development of Hausa WordNet (HWN). HWN [25] is a lexical resource for the Hausa language which extracts knowledge from a conventional Hausa dictionary and adopts a substructure of English and Hindi WordNets. It groups words based on different categories, introduces pronunciation, and uses close class categories to address the problem of missing pronunciation and coverage from existing WordNets.

2) Twitter multilingual corpus: The corpus for the study comprises Twitter pre-election data collected from a multilingual community (Nigeria). The dataset was collected using tweepy streaming API, preprocessed, and manually annotated by selected human annotators via a web-based interface. The corpus consists of 12,334 tweets which are both monolingual and multilingual. The monolingual tweets comprises: pure English language tweets and pure Hausa language tweets while the multilingual tweets comprise of the combination of English and Hausa tweets as shown in Table I.

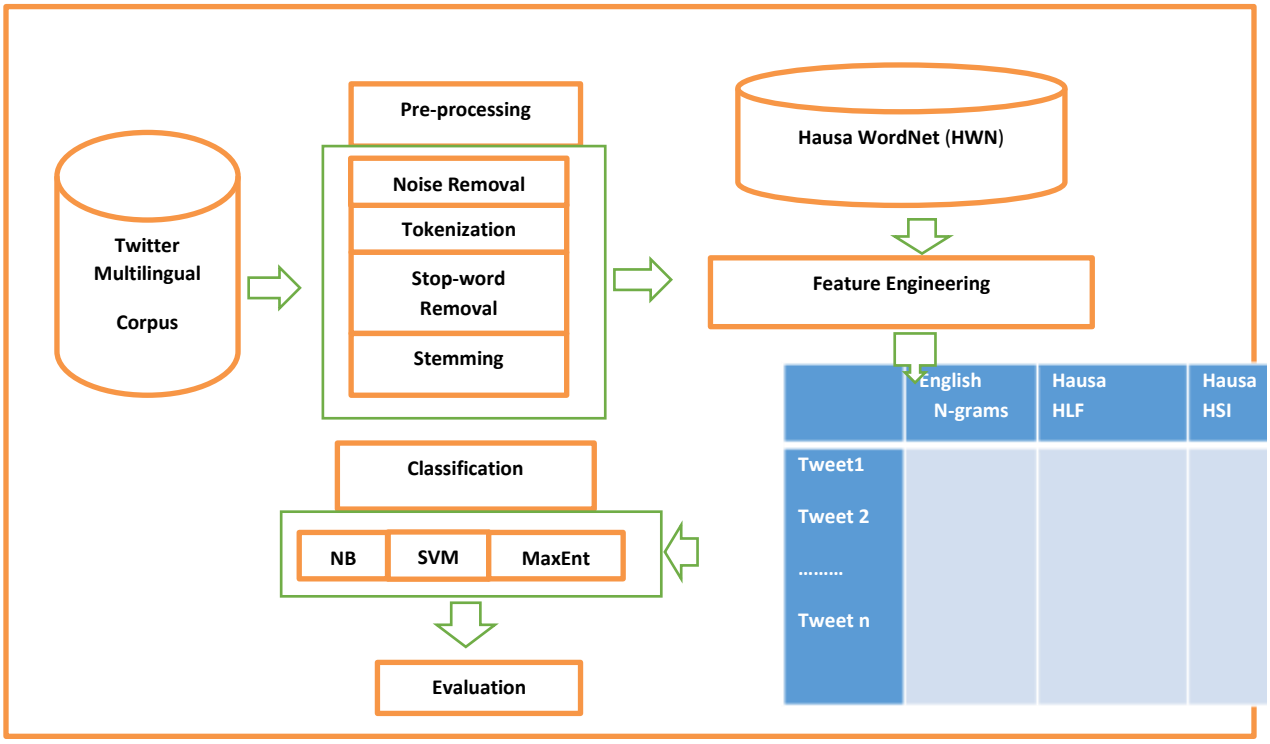
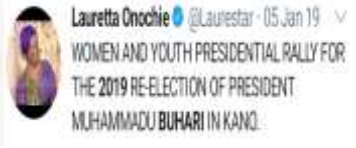


Fig. 1. Multilingual Sentiment Classification Workflow.

TABLE I. EXAMPLE OF MONOLINGUAL AND MULTILINGUAL TWEETS

Tweets	Language Classification
	Pure English (monolingual)
	Pure Hausa (monolingual)
	English and Hausa (multilingual)

B. Pre-Processing

The pre-processing task is an important step in sentiment classification. It is used to remove irrelevant parts from the data, as well as to transform the text to facilitate its analysis. The pre-processing step consists of four steps: noise removal, tokenization, stop word removal and stemming.

1) Noise removal: Removing noise in the text is one of the most essential pre-processing steps though it is highly domain-dependent. For example, Twitter contains a lot of noisy text such as a username (@username), retweets (RT), hashtags (#tag), mentions (@mentions), URL (web pages, web sites), emoticons (icons for smileys) and special characters ($^&*~ etc.). However, the streaming algorithm collected only the tweets without usernames, similarly, the dataset has no emoticons present. As a result, the noise removal process consists of removing all URLs, hashtags, retweets, and special characters from both pure English, pure Hausa, and Multilingual dataset. However, due to the user's constant informal practice of using social media, the only Hausa special character present is " ' " such as in "*ta'adda"*, " *'yanci"* and *"jama'a"*. Therefore, this character together with English special characters is removed.

2) Tokenization: Tokenization is the process of breaking a stream of text up into words, phrases, symbols or other meaningful elements called tokens. The list of tokens becomes an input for further processing such as sentiment classification. This process is very important in sentiment analysis of social media text because sentiment information can be sparsely and unusually represented [26]. The study implements word-level tokenization for all the three datasets. Each tweet will now be tokenized and split into words where each word needs to be captured and subjected to further pre-processing like stop word removal and stemming.

3) Stop-word removal: The removal of stop words in sentiment classification is necessary as the general idea of a text is retained in the absence of these words and also adds quality to the model. For the pure English dataset, NLTK was used in removing all English language stop-words, frequent words such as *the, and, of, a* and *is* that are present in almost

all tweets were removed. Whereas, in the pure Hausa dataset, stop words such as "*amma", " wannan", "yi", "za", "wata", "kuma", " cikin"*…, etc. were removed and this was done by searching for words in a pre-existing list of Hausa stop-words [1] . Similarly, for the multilingual dataset, the combination of pre-existing English and Hausa stop-words are filtered out from the dataset as they carry less discriminative power in analyzing sentiment in a multilingual context.

4) Stemming: Stemming is important in NLP as compound words are replaced by their morphological root. Hausa language stemmer [27] from the NLTK python library was used to perform the Hausa stemming process, for example, the word '*yan-maza* and *mazaje* share the same root word as *maza*. While for the English dataset, the porter English stemmer from NLTK python library was used to perform the word stemming process, for example, the words "*tester*", "*testing*", and "*tested*" all share the same root-word "*test*". At the end of this preprocessing phase, we obtain a set of a stemmed bag of words that represents the original feature vector.

C. Feature Engineering Approach

The study adopts two approaches for multilingual sentiment classification; the first approach is building a baseline method using the corpus to generate n-gram features. Whereas the second approach is improving the baseline approach using Term Frequency to generate the Hausa feature vectors for the classification process. The main objective in this approach is to determine whether the features introduced from Hausa language can improve sentiment classification accuracy.

1) The baseline approach: This approach uses Twitter's multilingual corpus to develop the English feature i.e. n-gram features as the baseline approach for sentiment classification.

a) N-gram Feature: An n-gram is a contiguous sequence of n words from a given piece of text. Typically, n-grams are the basic features used in supervised sentiment classification. The *n*-gram features used in the study are weighted unigrams, bigrams, and trigrams. The study did not explore higher-order *n*-grams to try and minimize the negative effect of high dimensionality.

Weighted n-gram features are generated from our datasets so as to assign weights to each gram in the feature vector to indicate their importance. Term Frequency-Inverse Document Frequency (TF-IDF) technique was implemented to generate weighted features. This technique is a statistical measure used to evaluate how important a word is to a document or corpus. The importance increases proportionally to the number of times a word appears in a document but is offset by the frequency of the word in the corpus [2].

TF-IDF is composed of two terms: Term Frequency (TF) and Inverse Document Frequency (IDF). TF measures how frequently a term (t) occurs in a document.

$$TF\ (t)=\frac{Number\ of\ term\ (t)\ in\ tweet}{Max\ (\ occurence\ of\ terms\ in\ tweets)} \quad (1)$$

IDF measures how important a term is.

$$IDF(t) = \frac{Total\ Number\ of\ tweets}{Number\ of\ tweets\ with\ term\ (t)} \quad (2)$$

Finally, the TF-IDF measure the product of TF and IDF, as follows

$$TF - IDF\ (t) = Tf\ (t) * IDF\ (t) \quad (3)$$

This means that larger weights are assigned to terms that appear relatively infrequent throughout the corpus, but very frequently in individual documents/tweets.

2) Developing hausa features: Developing newly defined features is an important task in sentiment classification and more generally in text classification. Thus, researchers along this line argue that selecting the right feature determines the overall performance of sentiment classification. To this end, the study developed the following Hausa features.

a) Hausa Lexical Feature (HLF): These are Hausa features generated from the co-occurrence of common words from both Hausa lexical resource (HWN) and multilingual twitter corpus. Therefore any word that occurs in both HWN and the corpus is now term as Hausa word. These common words will help us identify Hausa words from the multilingual tweets. Therefore, the feature vector is generated by finding the term frequency (*TF*) of Hausa words (*hw*) in a tweet and can be represented as follows:

$$TF(hw) = \frac{Number\ of\ hw\ in\ tweet}{Max\ (\ occurence\ of\ hw\ in\ tweets)} \quad (4)$$

$$HLF = TF(hw) \quad (5)$$

This approach will normalize the distribution of Hausa words in the corpus. Some users express themselves in Hausa and English when they are sad, angry or frustrated and some otherwise as shown in the following examples.

Example 1: To hell with this government babu komai sai zalunci.

Example 2: PMB is purely direct, bai iya manufunci ba, part of the reason why I support him.

Example 1 has a frequency of 4 Hausa words (*babu, komai, sai, and zalunci*), similarly, example 2 has 4 Hausa words (*bai, iya, manufunci, ba*). Therefore, this feature will determine if the frequency of Hausa words has any effect on a particular sentiment in a multilingual context.

b) Hausa Sentiment Intensifiers (HSI): These are Hausa words that emphasize or intensify sentiment. The study makes use of a dictionary of Hausa intensifiers developed purposely for this study. These intensifiers are generated from Hausa words and then manually annotated (as either positive, negative or neutral intensifiers) with a substantial inter-annotator agreement (Kappa= 0.8). The annotation exercise was conducted by 3 experts in the field of sentiment analysis, Hausa language, and Linguistics who were also proficient speakers of both English and Hausa languages and also able to

[1] https://github.com/stopwords-iso/stopwords-ha

comprehend social media contents. The resulted annotated words are then compared with their existing meaning from HWN to further verify their intensity. Therefore, this feature vector is generated by finding the term frequency of Hausa Positive Intensifiers *TF* (*hpi*), term frequency of Hausa Negative intensifiers *TF (hngi*), and term frequency of Hausa Neutral Intensifiers *TF (hni)* and normalise by their maximum occurrences in the document (tweets). This approach can be represented as follows:

$$TF(hpi) = \frac{Number\ of\ hpi\ in\ tweet}{Max\ (\ occurence\ of\ hpi\ in\ tweets)} \quad (6)$$

$$TF(hngi) = \frac{Number\ of\ hngi\ in\ tweet}{Max\ (\ occurence\ of\ hngi\ in\ tweets)} \quad (7)$$

$$TF(hni) = \frac{Number\ of\ hni\ in\ tweet}{Max\ (\ occurence\ of\ hpi\ in\ tweets)} \quad (8)$$

Table II shows some examples of Hausa words and their sentiment intensification.

TABLE II. SOME HAUSA SENTIMENT INTENSIFIERS

s/n	Hausa Words	Sentiment Intensity
1	Da kyau	Positive
2	Dodar	Positive
3	MashaAllah	Positive
4	Tayani	Neutral
5	Kajifa	Neutral
6	Anya	Negative
7	Tabdijam	Negative
8	Tirkashi	Negative

These words are clear indicators of sentiment when express in a context as shown in the following examples:

Example 3: PMB for the second tenure, anya kuwa?

Example 4: MashaAllah, the Kano rally was conducted, da kyau.

Example 3 has a frequency of only 1 negative intensifier (anya) while example 4 has a frequency of 2 positive intensifiers (mashaAllah, da kyau). Therefore, HSI feature was implemented to determine whether Hausa intensifiers have any effect on a particular sentiment in a multilingual context.

D. Machine Learning Methods: The Classification Algorithm

The pre-processed datasets were split into training (70%) and testing (30%) data. The training data are processed by the classification algorithms in Scikit-learn machine learning in Python [28]. A Tfidf Vectorizer is implemented on the datasets using sublinear_tf to reduce the bias generated by words that appear frequently. Extracted features from HLF were then appended with the baseline features (N-gram) to the training and validation data. Similarly, extracted features from HSI were appended with HLF and the baseline feature to the training and validation data using Numpy.

The vectors were trained on Naive Bayes (MultinomialNB), Support Vector Machine (SVM) and Maximum Entropy (MaxEnt) classifiers. The SVM regularization parameter (C) is set to 1e5 or 100000 (the larger the C the better the validation). MaxEnt regularization parameter (C) is set to 128 (smaller values specify stronger regularization) and then a prediction was generated, the accuracy of the prediction was then tested using accuracy_score.

These classification algorithms were used due to their simplicity, effectiveness and accurateness in a supervised learning classification process. As for the test data, the classification algorithms corresponding to the built model is applied, and thus classification results are obtained. A brief background about these classifiers is presented.

1) Naïve bayes classifier: Naïve Bayes (NB) is a probabilistic classifier that operates by building statistical models of classes from the training dataset. The study make use of a Naive Bayes model class c to represent a class and x for features calculated individually as shown in the formula.

$$P\ (c\backslash x) = \frac{P\ (x\backslash c)P\ (c)}{P\ (x)} \quad (9)$$

2) Support vector machine: Support Vector Machine (SVM) is highly effective and generally outperforms other classifiers at sentiment classification [26]. SVMs utilize hyperplanes to separate classes and seek a decision hyperplane represented by a support vector that separates the positive and negative training vectors of documents with maximum margin.

3) Maximum entropy: Maximum Entropy (MaxEnt) is a feature-based classifier that work on the idea that the most uniform model that satisfies a given constraint should be preferred [28]. In a two-class scenario, it is the same as using logistic regression to find a distribution over the classes. The model is represented by the following.

$$P\ \left(\frac{c}{d}, \lambda\right) = \frac{\exp(\Sigma_i \lambda_i\ f_i\ (c,d))}{\Sigma_d\ \exp\ (\Sigma_i\ \lambda_i\ f_i(c,d))} \quad (10)$$

The above classification algorithms were used to evaluate the feature set.

N-gram: Classifiers that evaluates only n-gram features.

N-gram + HLF: Classifiers that evaluates n-gram and HLF features.

N-gram + HLF + HSI: Classifier that evaluates n-gram, HLF and HSI features.

IV. EXPERIMENT

In this section, dataset characteristics, system setup, evaluation criteria and the results for evaluating the performance of the proposed approach is discussed.

1) Dataset characteristics: Several experiments were conducted on the Twitter multilingual corpus. The corpus comprises 12,405 tweets and each tweet has sentiment annotations on tweet level by 2 human annotators using sentiment classes positive, negative and neutral. The annotators' classes were aggregated to assign a sentiment to tweet, where tweet *t* has sentiment *S* if 2 annotators marked

the tweet with *S*; otherwise, the sentiment of *t* is conflicted. Similarly, if the sentiment of *t* is conflicted then t will be discarded.

$$t=S \text{ if } S1=S_{2 \text{ else}} \; t=\text{conflicted} \quad (11)$$

$$\text{If } t=\text{conflicted then } t=\text{discard} \quad (12)$$

Therefore, after removing all discarded tweets, the corpus now comprises 12,334 tweets; 4,623 of them considered as positive tweets, 6,531 as negative tweets, and the other 1,180 as neutral opinions. Similarly, the tweets are both monolingual and multilingual; the monolingual tweets comprise 1- Pure English language which comprises 10,900 tweets 2- Pure Hausa language with 244 tweets while the multilingual tweets comprise of the combination of English and Hausa language tweets with 1,190 tweets as shown in Table III.

2) System setup: The experimental setup was implemented with the following tools and environment:

- Windows 10 operating system.
- System specification of 30GB Hard disk space, 6GB RAM, and Intel ® core ™ processor @ 2.40GHz.
- Python programming language using Jupyter notebook is deployed which provides an easy and fast modelling workspace for the experiment.

3) Evaluation criteria: To evaluate the quality and usefulness of the classifiers and to efficiently integrate the feature set to synthesize a more accurate classification procedure, experimental results were sorted into the following: accuracy, precision, recall and F1. The contingency table below illustrates the arrangement of actual and predicted classification in a three-class problem (positive, negative, and neutral).

Table IV reports the counts of True Positives (TP), False Positives (FP), True Negative (TN), and False Negatives (FN) which are defined as follows:

- TP (A): TP is the number of positive tweets correctly classified as positive.
- FP (D + G): FP is the number of negative tweets falsely classified as positive.
- TN (E + I): TN is the number of negative tweets correctly classified as negative.
- FN (B + C): FN is the number of positive tweets falsely classified as negative.

Precision (P) for the three classes, positive, negative, and neutral is determined as follows:

$$P_{(Positive)}=A/(A+D+G) \quad (13)$$

$$P_{(Negative)}=E/(B+E+H) \quad (14)$$

$$P_{(Neutral)}=I/(C+F+I) \quad (15)$$

Recall (R) for the three classes, positive, negative, and neutral is determined as follows:

$$R_{(Positive)}=A/(A+B+C) \quad (16)$$

$$R_{(Negative)}=E/(D+E+F) \quad (17)$$

$$R_{(Neutral)}=I/(G+H+I) \quad (18)$$

The F1 for a class is given by the harmonic mean of the class precision and recall as follows:

$$F1=2TP/(2TP+FP+FN) \quad (19)$$

Similarly, accuracy is the number of correct predictions from all predictions made.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (20)$$

$$Accuracy = \frac{A+(E+I)}{A+(E+I)+(D+G)+(B+C)} \quad (21)$$

4) Experimental result and discussion: The performance of the classification algorithms using the feature set is analyzed. The Tables below represent the performance of the classification algorithms for the three classes evaluated by the metrics: accuracy, precision, recall and F1-score. The results were also classified based on datasets and each dataset was categorized based on the feature set used. However, the pure English dataset was evaluated using only the N-gram feature as there are no Hausa words in the dataset, and instantiating.

Hausa features in a non-Hausa dataset will yield an erroneous classification.

From Tables V, VI and VII above, bold font indicates the best performance on a dataset, and asterisk, *, indicates significant difference from the baseline, Ngram. The pure English dataset uses only the baseline features (Ngram) and achieves a little accuracy of 56% with SVM classifier. For the pure Hausa dataset, the best result is obtained using SVM classifier with an accuracy of 71% and for multilingual dataset the best result is obtained with Naïve Bayes classifier with an accuracy of 68%.

TABLE III. NUMBER OF CLASSES PER LANGUAGE

Classes	English	Hausa	Multilingual	Total
Positive	4,134	143	346	4,623
Negative	5,794	73	664	6,531
Neutral	972	28	180	1,180
Total	10,900	244	1,190	**12,334**

TABLE IV. CONTINGENCY TABLE

Classification	Positive	Negative	Neutral
Positive	**A**	**B**	**C**
Negative	**D**	**E**	**F**
Neutral	**G**	**H**	**I**

TABLE V. RESULTS FROM THE PURE ENGLISH TEST DATASET

Algorithm	Accuracy (%)	Positive (%) P R F1	Negative (%) P R F1	Neutral (%) P R F1
Naïve Bayes Ngram	55	49 82 61	66 59 62	29 02 04
SVM Ngram	56	52 77 62	63 62 63	43 12 19
MaxEnt Ngram	55	52 77 62	62 63 62	31 06 10

TABLE VI. RESULTS FROM THE PURE HAUSA TEST DATASET

Algorithm	Accuracy (%)	Positive (%) P R F1	Negative (%) P R F1	Neutral (%) P R F1
Naïve Bayes Ngram	50	45 80 58	64 24 35	52 50 51
Ngram+HLF	53	56 68 61	39 61 48	75 38 50
Ngram+HLF+HSI	**64***	58 70 64	73 73 73	65 50 57
SVM Ngram	57	59 64 62	60 52 56	54 58 56
Ngram+HLF	60	62 64 63	44 67 53	77 53 63
Ngram+HLF+HSI	**71 ***	74 67 70	65 77 71	72 69 71
MaxEnt Ngram	59	57 64 60	65 52 58	56 62 59
Ngram+HLF	60	58 64 61	48 67 56	77 53 63
Ngram+HLF+HSI	**68***	69 67 68	64 73 68	71 65 68

TABLE VII. RESULTS FROM THE MULTILINGUAL TEST DATASET

Algorithm	Accuracy (%)	Positive (%) P R F1	Negative (%) P R F1	Neutral (%) P R F1
Naïve Bayes Ngram	64	69 56 62	65 62 64	60 77 67
Ngram+HLF	**68***	63 71 66	69 68 69	75 67 71
Ngram+HLF+HSI	61	61 63 62	64 50 56	61 72 66
SVM Ngram	65	72 54 62	60 70 65	66 72 69
Ngram+HLF	**66**	62 67 64	68 64 66	69 68 68
Ngram+HLF+HIS	**66**	66 63 65	62 62 62	66 68 67
MaxEnt Ngram	64	71 55 62	63 66 64	61 74 67
Ngram+HLF	**66**	66 61 63	70 63 66	63 73 68
Ngram+HLF+HSI	65	69 63 66	65 60 63	62 73 67

The use of all feature set (Ngram+HLF+HSI) provides the best classification accuracy on pure Hausa dataset as shown in Fig. 2 while achieves little to no improvement on the multilingual dataset, this can be due to the higher frequency of Hausa words and Hausa intensifiers in Hausa dataset compared to the multilingual dataset. Similarly, the feature set (Ngram+HLF) provides the best classification accuracy and on the multilingual dataset.

The developed Hausa features from the two datasets (Hausa dataset and multilingual dataset) improve the accuracy of the baseline (with the exception of naïve Bayes classifier on multilingual dataset) using the 3 classifiers. Similarly, SVM is the best classification algorithm for all the datasets with 71% as shown in Fig. 3.

Furthermore, since there is no any existing work in Hausa language for direct comparison, the obtained Hausa dataset result is compared with result from Arabic dataset [2] using Arabic Sentiment Tweet Dataset (ASTD), we find the 2 results having equal accuracy of 68% when apply to Maximum Entropy (Logistic regression) while ASTD has a higher accuracy when apply on Naïve as shown in Table VIII. However, the SVM classifier in the proposed approach improves the accuracy to 71% and this can be due to its optimal margin gap between separating hyperplanes, thus, it is more robust in classification approaches.

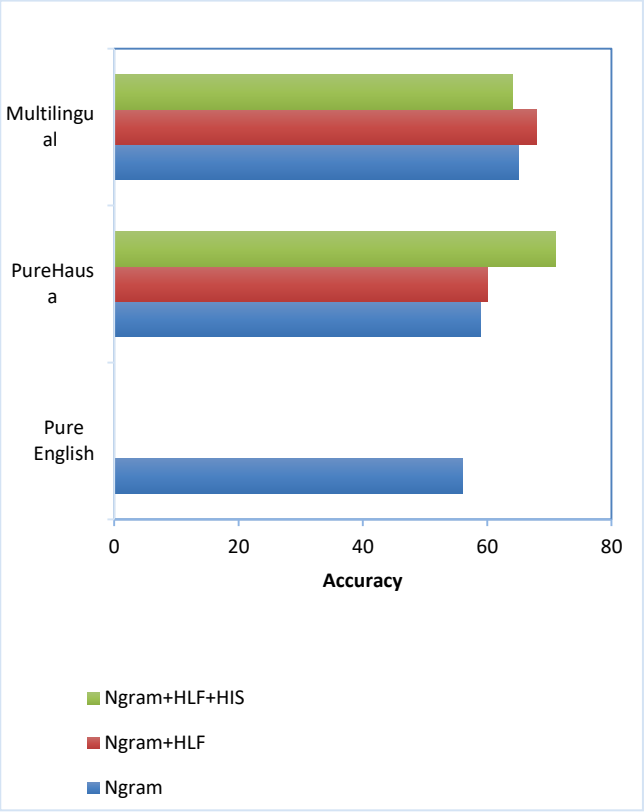


Fig. 2. Feature Set Accuracy Performance on Dataset.

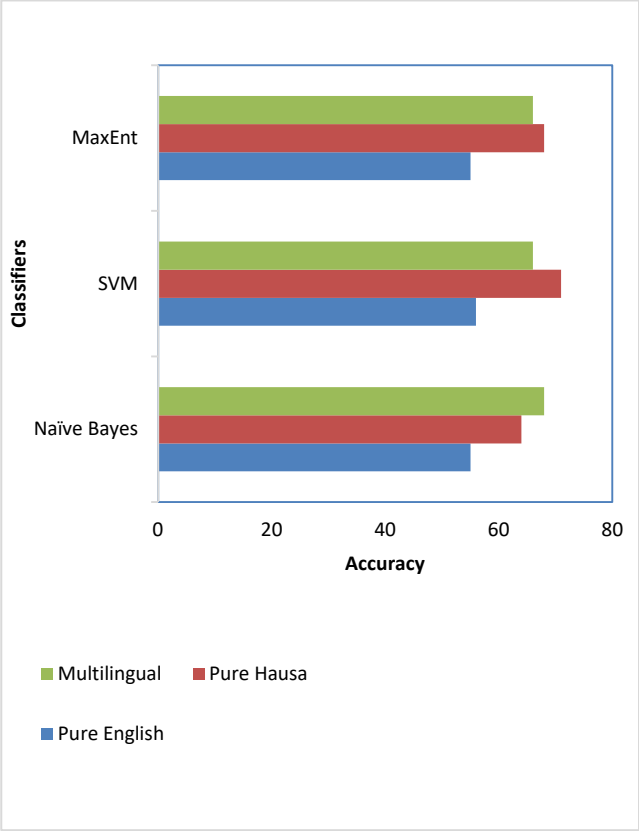


Fig. 3. Dataset Accuracy Performance on Classifiers.

TABLE VIII. ACCURACY COMPARISON USING HAUSA AND ARABIC DATASET

Dataset	Classifier	Accuracy
Proposed Approach		
Hausa	Naïve Bayes	64
	SVM	71
	MaxEnt	68
Elhadad et al., 2019		
Arabic [2]	Naïve Bayes	67
	SVM	NA
	Logistic Regression	68

V. CONCLUSION AND FUTURE WORK

The study proposed multilingual sentiment analysis of English and Hausa tweets using an Enhanced Feature Acquisition Method (EFAM). The method uses feature integration originating from two languages (English and Hausa) into a machine learning approach to multilingual sentiment analysis. We show that an enriched feature set provides effective modelling for sentiment classification of social media text. We achieved better classification performance using an SVM classifier and the use of all feature set (Ngram+HLF+HSI) provides the best classification accuracy on pure Hausa dataset while feature set (Ngram+HLF) provides the best classification accuracy on the multilingual dataset. Similarly, the results demonstrated that each of the newly defined feature set improves sentiment classification performance.

The pitfall of this study is that Term Frequency and Term frequency/Inverse document Frequency serve as a lexical level feature and thus tend to ignores the syntax and semantic of text. For future work, there are many avenues to pursue, including: 1- The use of statistical significant test such as T-test or ANOVA as supplementary technique. 2- Extending the proposed system against various languages other than Hausa or English. 3- Performance can be improve using a deep learning approach to automatically learn high-level features from the dataset by encoding sentiment information using Hausa and English word embedding methods.

REFERENCES

[1] S. L. Lo, E. Cambria, R. Chiong, and D. Cornforth, “Multilingual sentiment analysis : from formal to informal and scarce resource languages,” Artif. Intell. Rev., vol. 48, pp. 499–527, 2017.

[2] K. M. Elhadad, L. Fun, and G. Fayez, “Sentiment Analysis of Arabic and English Tweets,” in Web, Artificial Intelligence and Network, B. L, T. M, X. F, and E. T, Eds. Springer, Cham, 2019, pp. 334–348.

[3] N. Yadav, O. Kudale, S. Gupta, A. Rao, and A. Shitole, “Twitter Sentiment Analysis using Machine learning for Product Evaluation,” in IEEE nternational Conference on Inventive Computation Technologies (ICICT), 2020, pp. 181–185.

[4] A. Bakliwal, F. Jennifer, van der P. Jennifer, O. Ron, T. Lamia, and H. Mark, “Sentiment Analysis of Political Tweets: Towards an Accurate Classifier,” in Proceedings of the Workshop on Language in Social Media, 2013, pp. 49–58.

[5] B. Liu, "Synthesis Lectures on Human Language Technologies," Morgan and Claypool Publishers, 2012.

[6] A. Go, R. Bhayani, and L. Huang, “Twitter Sentiment Classification using Distant Supervision,” Standfort, 2009.

[7] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," Assoc. Comput. Linguist., vol. 37, no. 2, pp. 267–307, 2011.

[8] J. M. Chenlo and D. E. Losada, "A Machine Learning approach for Subjectivity Classification based on Positional and Discourse Features," in Multidiscilinary Information Retrieval, vol. 8201, L. M, K. E, and O. Loizides, F, Eds. Springer, Berlin, 2013, pp. 17–28.

[9] G. Vaitheeswaran and L. Arockiam, "Combining Lexicon and Machine Learning Method to Enhance the Accuracy of Sentiment Analysis on Big Data," Int. J. Comput. Sci. Inf. Technol., vol. 7, no. 1, pp. 306–311, 2016.

[10] M. Zubair, A. Khan, S. Ahmad, M. Qasim, and A. K. Khan, "Lexicon-enhanced sentiment analysis framework using rule-based classification scheme," PLoS One, pp. 1–22, 2017.

[11] A. Gupta, J. Pruthi, and N. Sahu, "Sentiment Analysis of Tweets using Machine Learning Approach," Int. J. Comput. Sci. Mob. Comput., vol. 6, no. 4, pp. 444–458, 2017.

[12] S. Joshi and D. Deepali, "Twitter Sentiment Analysis Systems," Int. J. Comput. Appl., vol. 180, no. 47, pp. 35–39, 2018.

[13] F. Zarisfi, S. Faramarz, and E. Esfandiar, "Solving The Twitter Sentiment Analysis Problem Based on a Machine Learning Based-Approach," J. Evol. Intell., vol. 13, pp. 381–398, 2020.

[14] I. Mozetic, J. Smailovi, and M. Grcar, "Multilingual Twitter Sentiment Classification : The Role of Human Annotators," PLoS One, vol. 11, no. 5, pp. 1–26, 2016.

[15] S. Narr, H. Michael, and S. Albayrak, "Language-Independent Twitter Sentiment Analysis," in In Proceedings of the Knowledge Discovery and Machine Learning., 2012.

[16] A. Balahur and M. Turchi, "Multilingual Sentiment Analysis using Machine Translation," Proc. 3rd Work. Comput. approaches to Subj. Sentim. Anal. Assoc. Comput. Linguist., pp. 52–60, 2012.

[17] N. Suri and T. Verma, "Multilingual Sentiment Analysis on Twitter dataset using Naive Bayes Algorithm," Sch. J. Eng. Technol., vol. 5, no. 9, pp. 473–477, 2017.

[18] A. S. Muhammad, M. M. Aliyu, and S. I. Zimit, "Towards the Development of Hausa Language Corpus," Int. J. Sci. Eng. Res., vol. 10, no. 10, pp. 1598–1604, 2019.

[19] A. Cui, M. Zhang, Y. Liu, and S. Ma, "Emotion Tokens : Bridging the Gap among Multilingual Twitter Sentiment Analysis," in Information Retrieval Technology, no. 7097, M. Saleem, K. Shaalan, F. Oroumchia, A. Shakery and H. Khalalfa, Eds. Springer, Berlin, 2011, pp. 238–249.

[20] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in Proceedings of the Seventh International Conference on Language Resource s and Evaluation (LREC 10), 2010, pp. 2200–2204.

[21] A. Pak and P. Patrick, "Twitter as a corpus for sentiment analysis and opinion mining," in Proceedings of the Seventh International Conference on Language resources and Evaluation (LREC' 10), 2010, pp. 1320–1326.

[22] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and W. Theresa, "SemEval-2013 task 2: Sentiment analysis in twitter," in Proceedings of the International Workshop on Semantic Evaluation, SemEval 13, 2013, pp. 312–320.

[23] B. Comrie and B. Comrie, "Hausa and the Chadic Languages," in The Worlds's Major Languages, Third., Taylor & Francis Group, 2018.

[24] R. M. Newman and P. Newman, "The Hausa Lexicographic Tradition," African Journals Online, vol. 11, pp. 263–286, 2001.

[25] A. Imam, A. Roko, A. Muhammad, and I. Sa'id, "Hausa WordNet : An Electronic Lexical Resource," Saudi J. Eng. Technol., vol. 4, no. 8, pp. 279–285, 2019.

[26] B. A. Muhammad, "Contextual Lexicon-based Sentiment Analysis for Social Media," Robert Gordon University, 2016.

[27] A. Bimba, I. Norisma, K. Norazlina, N. Nur, and L. Valiukas, "Stemming Hausa Text: Using affix-rulesmand reference lookup to stem words in Hausa language," Lang. Resour. Eval., vol. 50, no. 3, pp. 687–703, 2016.

[28] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

Development of Technology for Summarization of Kazakh Text

Talgat Zhabayev, Ualsher Tukeyev
Department of Information Systems
Al-Farabi Kazakh National University
Almaty, Kazakhstan

***Abstract*—This paper presents the solution to the problem of summarizing Kazakh texts. The problem of Kazakh text summarization is considered as a sequence of two tasks: extracting the most important sentences of the text and simplifying the received sentences. The task of extracting the most important sentences of the text is solved using the TF-IDF method and the task of simplifying sentences is solved using the neural network technology "Seq2Seq". Problem of using NMT method for simplification of Kazakh was in absence of Kazakh dataset for training. To solve this problem in this work propose use transfer learning method. The use of transfer learning made it possible to use a ready-made model that was trained on a parallel corpus of Simple English Wikipedia and not create a simplification corpus in Kazakh from scratch. For this, a transfer learning technology for simplifying sentences of the Kazakh language has been developed, based on training a neural model for simplifying sentences in the English language. Main scientific contribution of this work is transfer learning technology for the simplification of Kazakh sentences using the parallel corpus of the English language simplification.**

Keywords*—Summarization; text simplification; low-resource language; seq2seq; transfer learning

I. Introduction

Automatic text summarization is the process of shortening text without losing meaning. It can be the text of one or several documents. Summarization has become widespread in recent years in many application areas. Practical applications - data analytics, automatic creation of headlines or short descriptions, news sites, information aggregators. In these tasks, there is a need for automatic annotation of a huge amount of text data, which is inefficient to perform manually; by automating these actions, you can achieve significant time savings.

There are two types of annotation - abstractive and extractive. Extractive summarization - highlights the most important sentences in a text that most fully describe this text. Abstractive summarization is the reduction of a text by paraphrasing the text into a short form. In this case, the final summarization may contain phrases or sentences that did not occur in the original text.

Text simplification is an area of study in computational linguistics that studies methods and techniques for simplifying textual content [1]. In Natural Language Processing it is used as one of the steps in summarization, text parsing [2], text translation, question-answer systems. Simplification is performed by shortening sentences, combining sentences, transforming sentences, paraphrasing. In this work, annotation is considered for the Kazakh language.

To implement transfer learning, we use the second parallel corpus Kazakh - English. A model trained on a large parent corpus of the English language should give a relatively high-quality result of simplification and be based on the quality of the model and not on the knowledge of the language. The relevance of the study is due to the fact that at present, research in the field of annotating the Kazakh language is focused on extractive summarization and little attention is paid to the abstractive method. The main reason of this situation is the absence of the Kazakh corpora for abstractive summarization and the difficulty of creating it. Applying transfer learning in this work, we use sequentially extractive and abstractive summations to obtain a short version of the text.

The scientific contribution of this work is: 1) in the development of a TF-IDF [3] model for texts of the Kazakh language, using the Kazakh language corpus, processing it to obtain frequencies for TF-IDF; 2) in the development of transfer learning technology for the simplification of Kazakh sentences using the parallel corpus of the English language simplification.

The remaining part of this paper is organized as follows. Section II contains an overview of existing papers on summarization, the use of neural networks for simplification. Section III contains the application of the neural network technology "Seq2Seq" to simplify text. The machine translation of the neural network technology "Seq2Seq" is based on the use of parallel data corpuses. After training, the model is able to generate the simplification of new sentences. Section IV describes the implementation of abstractive summarization and sentence extraction using the TF-IDF method. Section V contains a description of training the model and the results in the form of a table. Section VI concludes the overall study.

II. Related Work

Consider research on abstractive summation. At the moment, to implement this type of summarization, the most common method is using neural networks with "sequence to sequence" architectures. Initially, sequence to sequence neural networks were used in neural machine translation. The architecture of a neural network describes the number, types of layers, the number of neurons in them and how the layers are connected to each other. Seq2seq neural networks are used

together with the element of attention [4]. This type of neural networks consists of a decoder and an encoder, which are recurrent neural networks [5]. The use of sequence to sequence architecture for machine translation has been described in many works [6, 7].

Currently, the most used architecture in the summarization of text is the transformer [8]. The difference between the transformer architecture and seq2seq is parallel, not sequential processing of input sentences. Transformer is the so-called attention-based architecture. Simplification model training is distinguished primarily by the training corpus. A parallel corpus for simplification problems is a corpus, the source part of which is the ordinary language sentences, and the target part is the corresponding simplified language sentences. Thus, simplification is a monolingual task for neural machine translation.

The main corpus for simplification is the Simple English Wikipedia corpus [9]. Training on this corpus forms the basis of most of the papers on text simplification. So in [10] this simplification corpus was used to train the seq2seq simplification model.

The model was trained in the OpenNMT system [11]. It is one of the most popular tools for neural machine translation. There are several implementations – original OpenNMT, OpenNMT-Python, OpenNMT-Tensorflow. Similar papers in the field of simplification are [12,13] where a neural transformer model is also used.

There are many works available on the summarization of texts in low-resource languages. Transfer learning is also increasingly used for low-resource languages. For example, in [14, 15], the use of neural networks for abstractive summation together with the transfer of learning is considered.

In [16], the authors describe the creation of a synthetic set of complete sentences for simplification using a pretrained model. Neural networks are also used for extractive summation. In [17], a summarization corpus is used, where source is a set of ordinary sentences, and the target part of the corpus is the summation of the corresponding set of sentences.

In the [18], the authors used centroids and Word's mover distance for extraction summarization in Kazakh language. Many summarization studies look at TF-IDF and data clustering. Also TF-IDF is used in information extraction [19].

When defining sentences for extractive summarization, it is need to get those sentences that together describe the text as much as possible (and there should be no unnecessary, redundant sentences in summarization) [20]. Work [21] describes a similar implementation of summarization using TF-IDF.

Transfer learning methods find application in the case of low resource languages, such as in [22], where the authors used the general parent model and the child model to translate the Tibetan language. Transfer learning is an area of NLP research that focuses on the problem of retaining knowledge that was obtained by training one model and transferring knowledge to another, similar problem [23, 24].

III. Methodology of Summarization of Kazakh Text

The proposed methodology of summarization of Kazakh text includes two steps:

- Extraction of summarize sentences,
- Simplification of extracted sentences.

Below these two parts detailed are considered.

A. Extraction of Sentences

The TF-IDF metric is used to implement extractive summarization.

There are several options for using TF-IDF: 1) Ranking sentences by the value of TF-IDF or sentence centroids to find the most important sentences in the corpus; 2) search for the most similar sentences by semantic proximity; 3) clustering of sentences by the values of TF-IDF or centroids [25].

In our work, we use the centroid ranking method and clustering.

Below is a step-by-step implementation algorithm:

1) We perform preprocessing, which includes removing punctuation marks, apostrophes, dashes and other uninformative elements, tokenizing the text using the sent_tokenize function in order to get an array of text, where each element is a separate sentence.

2) We get term frequency - it is defined as the ratio of the number of times each unique word appears in the sentence to the number of words in the sentence.

3) We get inverse document frequency - a value that shows the significance or informativeness of a word in a sentence, allowing you to ignore words that appear in most sentences, such as prepositions. It is the logarithm of the ratio of the number of sentences to the number of occurrences of a word.

4) The centroid of each sentence is calculated as the ratio of the sum of TF-IDF values to the total number of unique words in the sentence.

5) We combine all the centroids of the sentences into one array, which contains the sentence number and the centroid value. Then we select several sentences with the largest centroid values.

In Fig. 1 shows a graph of the distribution of centroid values for a text in the Kazakh language, which was obtained as a result of text simplification. On the chart, the X-axis is the ordinal number of the sentence, the Y-axis is the values of the centroids of the corresponding sentence. Centroid - a value from zero to 1. On the diagram, we see how the centroid values of sentences are distributed in the corpus and in which part of the corpus the largest centroid values are.

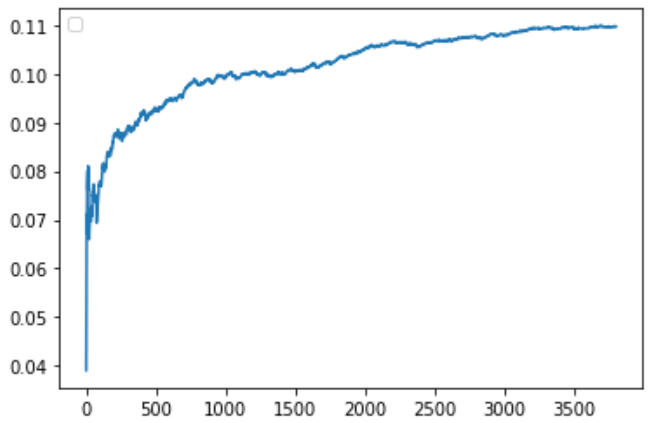


Fig. 1. Distribution of Centroid Values by Sentences of the Kazakh Corpus.

The next step is to analyze the resulting simplified corpus by distributing words into clusters for topic analysis. Table I shows the result of the distribution of the corpus into 6 clusters. For clustering, we use the k-means algorithm [26]. This algorithm allows you to group a set of vectors according to the degree of similarity. In this case, we use centroids as a criterion for the similarity of words. For simplicity, the number of clusters was chosen arbitrarily. As we can see, each cluster contains a set of words that are close in meaning.

TABLE I. WORD CLUSTERING

cluster	sentence
1	эскиздер,байқауына,эскиздерді,жобалау,көзқарас,президент,қызығушылық,қазақстан,иран,бұл
2	республикасының,қазақстан,президенті,президентіне,сауд,сенім,грамоталарын,тапсырды,бар,заңы
3	барлық,қазақстанның,мәселелері,қажет,өсім,ақпарат,салаларында,бар,және,диалог
4	бар,кездесу,жиналыспен,көршілес,қоғамдық,немесе,көптеген,қазақстанда,маңызы,қалада
5	премьер,министр,министрдің,кездесуінде,бұл,министрі,астана,шетелдіктердің,кеңестегі,қала

B. Simplification of Extracted Sentences

In this subsection, we will describe the algorithm for creating a simplification model of the Kazakh text using a method that relates to the transfer learning. The Kazakh language is a language with a small number of parallel corpuses, which makes learning a neural model very difficult. A model for working with the Kazakh language should be trained on a parallel corpus of the Kazakh text. For the simplification of texts in the Kazakh language, there are currently no ready-made simplification parallel corpus. Therefore, to obtain such a corpus, we use the Google translate application with manually edition to translate English parallel simplification corpus to Kazakh parallel simplification corpus.

The proposed methodology of simplification of Kazakh text includes two stages (Fig. 2).

At the first stage, the parent model is trained:

1) First, we define the architecture of the parent model. Before creating the model, it was necessary to choose the architecture of the neural network model that would show the highest score values in the original English corpus. To do this, we train seq2seq and a transformer model on a general corpus (Simple English Wikipedia) and see which architecture has bigger BLEU.

2) The English part of the kaz-eng corpus [27] is translated by the trained model. As a result, we got a simplified text of the English part of the kaz-eng corpus.

Second stage of transfer learning is the training of the child model:

1) The resulting simplified part of the kaz-eng corpus was translated into the Kazakh language, using the public web service of machine translation with the recording of the result in a text file.

2) As a result, a synthetic Kazakh parallel simplification corpus is obtained. The source part of Kazakh parallel simplification corpus is the Kazakh part of the source kaz-eng corpus and the target part is the simplified text of the source English part translated on Kazakh.

3) After that, the training a new neural model on the Kazakh parallel simplification corpus is made.

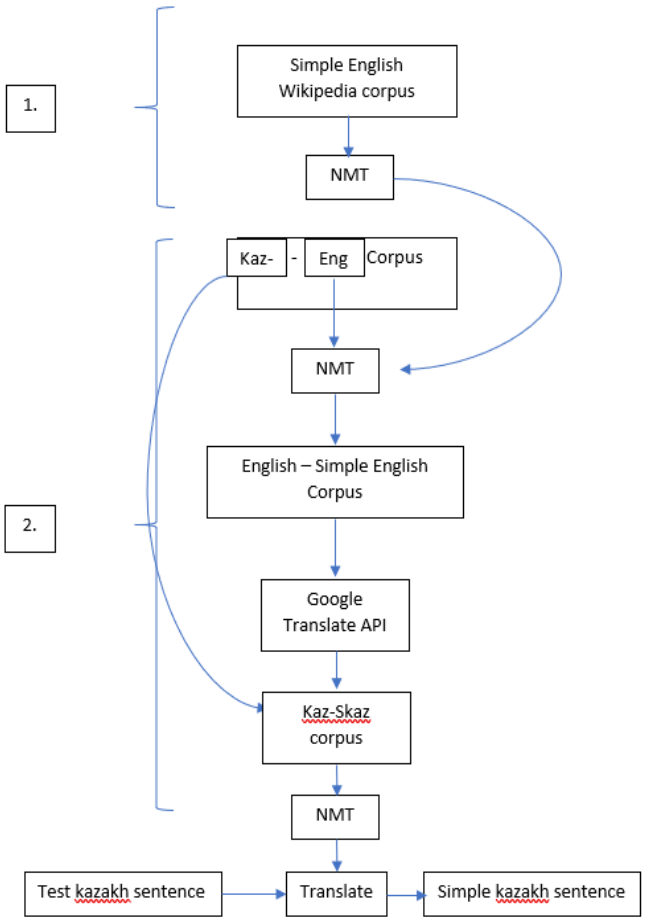


Fig. 2. Transfer Learning Technology for the Simplification of Kazakh Sentences.

The considered method of using the results of model translation to create a simplified part of the corpus belongs to transfer learning methods. These are learning transfer methods that use the generated synthetic data at runtime [28]. A corpus that was created based on the data generated by the model is called synthetic. Also, the creation of a synthetic corpus underlies the method of back-translation [29]. So, in [30] the author uses back-translation to increase the size of the training corpus. In our work, it is possible to use this method to increase the train corpus.

Nevertheless, synthetic data is worse than real data, and when a significant part of the train corpus is synthetic data, the model usually shows worse BLEU results compared to real data [31].

Further, the simplified text in the Kazakh language, allows to define the sentences that convey the essence of the text.

IV. NMT Experiment and Result

This section discusses training the model, obtaining results, and assessing the quality of summarization. Quality assessment is needed to determine how well the model performs, including with text that is very different from the training corpus. To assess the quality of the model, we use the BLEU and SARI metrics.

The BLEU (Bilingual Evaluation Understudy) metric is an assessment of the quality of machine translation from one language to another. The BLEU algorithm compares the number of common words or phrases in predicted sentences with reference sentences. Comparison is performed by counting N-gram matches. The final score for a corpus is the average quality score for all sentences in the corpus [32]. The metric has certain drawbacks in the case of text simplification, since was originally developed for machine translation rather than text simplification [33].

In [34], the SARI (System Output Against References and Input Sentences) metric was presented. This metric can assesses the quality of text simplification based on source, predictions and reference data, correctly taking into account the operations to change the sentence.

When training a model, one of the most important parameters is the number of training epochs and dropout. The Epoch - full cycle through the hull during training, it takes more than one epoch to train the model. Dropout is a technique used in training, which consists in shutting off the outputs of some neurons with a certain probability [35], which avoids overfitting the model. The model works on the basis of a vocabulary that was created during training. The size of the vocabulary affects the performance of the model. Vocabulary size we have set 50 000 words.

The parallel corpus of Simple English Wikipedia contains 284677 lines for training. The model was trained for 20 epochs, until the values of the loss function ceased to decrease significantly. The kaz-eng corpus contains 109 thousand lines, from where 5000 lines are allocated for testing.

To determine the most optimal option, we also applied the model fine-tuning method [36], which refers to inductive transfer learning. It differs from the previous method in the following steps: 1) it is necessary that the model that has been trained on the general corpus is retrained on the domain-specific corpus. To do this, OpenNMT connects the existing model vocabulary to the Kazakh corpus vocabulary; 2) a new savepoint is created in OpenNMT, which uses the new dictionary. The training of the model continues from a new point. Thus, the model trained in English is retrained taking into account the Kazakh language.

Table II shows the scores of BLEU and SARI of neural models depending on the parallel data corpus. These grades are obtained during testing after training the models.

TABLE II. BLEU and SARI Scores for Seq2seq and Transformer Model

№	*Model*	*Simple English Wikipedia BLEU/SARI*	*Kaz-eng BLEU/SARI*	*Kaz-skaz BLEU/SARI*
1	Seq2seq	58.01/52	53/66.18	Not trained
2	Transformer	66.70/66	55/60	7/36
3	Transformer Finetuned	66.70/66	55/60	8/36

Column “Simple English Wikipedia BLEU/SARI” contains the BLEU scores after training the model on the Simple English Wikipedia.

Column “Kaz-eng BLEU/SARI” - BLEU assessment at the stage of translating the English part of the kaz-eng corpus.

Column “Kaz-skaz BLEU/SARI” - BLEU score after training the Kazakh simplication model.

As we can see from the data in Table II for the seq2seq model, the BLEU score has changed from 58 on the Simple English Wikipedia corpus, to 53 on the kaz-eng corpus. The test set is the selected lines from the Simple English Wikipedia train corpus, that is, it is data of a similar subject.The kaz-eng corpus test set for the model was not used for training the model and this corpus was not originally for text simplification.

On the transformer model, the BLEU score is also reduced from 66 to 55. According to the parent data, the model with the transformer architecture has a slight advantage over the seq2seq attention model on the same data. This model is also the parent for fine-tuned Transformer model.

Therefore, the transformer architecture was chosen to create the Kazakh model.

The BLEU score on the resulting Kazakh child model is 7. As we can see from the assessment of the “Transformer Finetuned” model, the assessment increased by 1 and this method of creating the Kazakh model is better.

When translating, the model works with many unfamiliar words, and the meaning of words, depending on the context, may differ. This problem is called domain shift [37]. In other words, a model trained on news data does not work well with data from medicine or another field of science. This is one of the reasons for the low BLEU score on the Kazakh model. Another reason may be an error in training the transformer

model, which affected the quality of the translating, which we will try to fix in the future.

V. Conclusion and Future Work

In this paper, the method for summarizing Kazakh text was considered. Proposed Kazakh text summariszing method based on consequent using of TF-IDF method for extracting summarize sentences and NMT method for simplification of received summarize sentences. Problem of using NMT method for simplification of Kazakh was in absence of Kazakh dataset for training. To solve this problem in our method to propose use transfer learning method. The use of transfer learning made it possible to use a ready-made model that was trained on a parallel corpus of Simple English Wikipedia and not create a simplification corpus in Kazakh from scratch.

In future works, we plane to further improve the model, by increase the volume of the training dataset of the Kazakh corpus. Also we plane investigate using of post-editing NMT technology for increase of Kazakh parallel simplification corpus volume and quality. One of the directions for further research on this area is a method of clustering similar sentences in the train dataset and training a new seq2seq model based on it, as in [38]. Which should improve the performance of the model.

References

[1] Saggion H., "Automatic Text Simplification," Morgan & Claypool Publishers, 2017, p. 2.

[2] Chandrasekar R., Doran C., "Motivations and methods for text simplification," COLING Volume 2: The 16th International Conference on Computational Linguistics, pp. 1041-1042, 1996.

[3] Salton G., Buckley C., "Term-Weighting approaches in Automatic Text Retrieval," Information Processing and Management 24(5), pp. 513-523, 1988.

[4] Bahdanau D., Cho K., Bengio Y., "Neural machine translation by jointly learning to align and translate," 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015.

[5] Hochreiter S., Schimidhuber J., "Long short-term memory," Neural Computation:journal, Vol.9, no.8. - pp.1735-1780, 1997.

[6] Sutskever I.,Vinuals O., Q.V. Le., "Sequence to Sequence Learning with Neural Networks," Advances in Neural Information Processing Systems, pp. 3104-3112, 2014.

[7] Cho K., Van Merrienboer B.. Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," EMNLP 2014 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, Doha, pp. 1724-1734, 2014.

[8] Vasvani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., "Attention is all you need," Advances in Neural Information Processing Systems 30, pp. 5998-6008, 2017.

[9] Hwang W., Hajishirzi H., Ostendorf M., Wu W., "Aligning Sentences from Standard Wikipedia to Simple Wikipedia," Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies, pp.211-217, 2015.

[10] Nisioi S., Stagner S., Ponzetto S.P., "Exploring Neural Text Simplification Models," ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) , Vancouver, V. 2., pp. 85-91, 2017.

[11] Klein G.,Kim Y., Deng Y., Senellart J., Rush A., "OPENNMT:Opensource toolit for neural machine translation," In Proceedings of ACL 2017, System Demonstrations, Vancouver. Association for Computational Linguistics, pp. 67-72, 2017.

[12] Surya S., Mishra A., "Unsupervised text simplification," ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, pp. 2058-2068, 2019.

[13] Maruyama T., Yamamoto K., "Extremely low resource text simplification with pre-trained transformer language model," Proceedings of the 2019 International Conference on Asian Language Processing, IALP 2019, pp. 53-58, 2019.

[14] Chowdhury R.R., Nayeem M.T., Mim T.T.,"Unsupervised abstractive summarization of Bengali text documents," EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, pp. 2612-2619.

[15] Quasmi N.H., Zia H.B., Athar A., Raza A.A.,"SimplifyUR:unsupervised lexical text simplification for urdu," LREC 2020 -12th International Conference on Language Resources and Evaluation, Conference Proceedings, 2020 12th International Conference on Language Resources and Evaluation, LREC 2020, Marseille, 164155, pp. 3484 - 3489, May 2020.

[16] Parida S., Motlicek P., "Abstract text summarization: a low resource challenge," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP), 2019.

[17] Ambrosio A.P., Tonelli S., Turchi M., Negri M., Di Gangi M.A., "Neural text simplification in low-resource conditions using weak supervision," Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation, 2019.

[18] Seitkali, D., Musabayev, R., "Using centroid keywords and word mover's distance for single document extractive summarization," ACM International Conference Proceeding Series, pp. 149-152, 2019.

[19] Hashemzadeh B., Abdolrazzagh-Nezhad M., "Improving keyword extraction in multilingual texts," International Journal of Electrical and Computer Engineering Open AccessVolume 10, Issue 6, pp. 5909 - 5916, December 2020.

[20] Gholipour Ghalandari D., "Revisiting the Centroid-based Method: A Strong Baseline for Multi-Document Summarization," Proceedings of the Workshop on New Frontiers in Summarization, September 2017.

[21] Christian H., Pramodana Agus M., Suhartono D., "Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF)," International Journal of Electrical and Computer Engineering (IJECE) Vol. 10, No. 6, December 2020, pp. 5909~5916.

[22] Zhou M., Secha J., Cai R., "Domain adaptation for tibetan-chinese neural machine translation," ACAI 2020: 2020 3rd International Conference on Algorithms, Computing and Artifical Intelligence December 2020. Article No.: 77 , pp. 1-5, December 2020.

[23] West J., Ventura D., Warnick S., "Spring Research Presentation: A Theoretical Foundation for Inductive Transfer," Brigham Young University, College of Physical and Mathematical Sciences. Archived from the original on 2007-08-01. Retrieved 2007-08-05, 2007.

[24] Malte, Aditya and Pratik Ratadiya, "Evolution of transfer learning in natural language processing," arXiv:1910.07370, 2019.

[25] Radev D., Hongyan J., Stys M., Tam D.. 2004. "Centroid-based summarization of multiple documents," Information Processing and Management 40(6), pp. 919-938, 2004.

[26] Pelleg, D., Moore A., "Accelerating exact k -means algorithms with geometric reasoning," Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '99. San Diego: ACM Press: pp.277–281, 1999.

[27] https://github.com/NLP-KazNU/kaz-parallel-corpora_collect_and_clean.

[28] Pan S.J., Yang Q., "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering (Volume: 22, Issue: 10), pp. 1345-1359, Oct. 2010.

[29] Sennrich R., Haddow B., Birch A., "Edinburgh Neural Machine Translation Systems for WMT 16,"In Proceedings of the First Conference on Machine Translation, pp. 371-376, Berlon, 2016.

[30] Qiang, Jipeng, "Improving Neural Text Simplification Model with Simplified Corpora," arXiv:1810.04428 , 2018.

[31] Wu L., Wang Y., Xia Y., Tao Q., Lai J., Liu T.Y., "Exploiting monolingual data at scale for neural machine translation," Conference

on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, pp. 4207-4216, 2019.

[32] Papineni K., Roukos S., Ward T.,Wei-Jing Zhu, ”BLEU: a method for automatic evaluation of machine translation,” Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002, pp. 311-318, 2002.

[33] Sulem E., Abend O., “Bleu is not suitable for the evaluation of text simplification,” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, pp. 738 - 744, 2020.

[34] Xu W., Napoles C., Pavlick E., Chen Q., “Optimizing Statistical Machine Translation for Text Simplification,” Transactions of the Association for Computational Linguistics, Volume 4. pp 401-415, 2016.

[35] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” Journal of Machine Learning Research 15, pp. 1929-1958, 2014.

[36] Chenhui C., Wang R., “A survey of domain adaptation for neural machine translation,” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp.1304-1309, 2018.

[37] Baochen S., Feng J., Saenko K., “Return of frustratingly easy domain adaptation,” 30th AAAI Conference on Artificial Intelligence, AAAI , pp. 2058-2065, 2016.

[38] C. Fan, Yu. Tian, Y. Meng, N. Peng, X. Sun, Fei Wu, Jiwei Li, “Paraphrase Generation as Unsupervised Machine Translation,” arXiv:2109.02950v1, 2021.

An Energy-aware Facilitation Framework for Scalable Social Internet of Vehicles

Abdulwahab Ali Almazroi, Muhammad Ahsan Qureshi
University of Jeddah, College of Computing and Information Technology at Khulais
Department of Information Technology
Jeddah, Saudi Arabia

***Abstract*—The Internet of Things (IoT) has eventually evolved into a more promising service provisioning paradigm, namely, Social Internet of Things (SIoT). Social Internet of Vehicles (SIoV) symbolizes a multitude of components from the existing Vehicular Ad-Hoc Networks (VANETs) such as OBUs, RSUs, and cloud devices that necessitate energy for proper functioning. It is speculated that the connected devices will surpass the 40 billion mark in the year 2022 in which the devices related to ITS will constitute a significant part. Therefore, the ever-increasing number of components increases the communication hopping that results in the immense escalation of energy consumption. However, the energy consumption at the object level increases due to individual communication, storage, and processing capabilities. The existing research in SIoV is focused on providing state-of-the-art services and applications; however, a significant goal of energy efficiency is largely ignored. Therefore, extensive research needs to be performed to come up with an energy-efficient framework for a scalable SIoV system to meet the future requirements of ITS. Consequently, this study proposed, simulated, and evaluated an energy-aware efficient deployment of RSUs scheme. The proposed scheme is based on network energy, data acquisition energy, and data processing energy. To achieve efficiency in terms of energy, traveling salesman problem with ant colony optimization algorithm are utilized. The experiments are performed in an urban scenario with different numbers of RSUs. The outcomes of the experiments exhibited promising results in energy gain and energy consumption having implications for society and consumers at large.**

Keywords*—Social Internet of Vehicles (SIoV); energy optimization; Travel Sales Person (TSP) problem; Ant Colony Optimization (ACO)

I. Introduction

VANETs [1] deploy vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) communication to facilitate a wide range of applications Such as safety, navigation, routing, emergency healthcare, and infotainment. These applications are developed using VANETs due to the ability of VANETs to handle different topologies and management of continuously changing network densities. The effective utilization of VANETs is essential due to encouraging possible applications that includes both V2I and V2V communications. Both V2I and V2V communications are established in urban, suburban and highway environments with a wide range of specific topological features [2]. However, the advancements in network technology and the use of the internet allows the researchers to conceptualize the Internet of Vehicles (IoV) to develop applications that were not possible in traditional VANETs due to lack of all-time connectivity and unavailability of the internet [3]. IoV allows each network object to connect to the internet; hence, making it possible for potentially all OBUs, RSUs, and network users (passenger and drivers etc.) to share information resulting in a saleable communication environment.

Over time, IOV evolved in Social Internet of Vehicles (SIoV) in which the network objects can share information about common interests such as road conditions, traffic information, available parking spaces and resource sharing [4]. The network objects in SIoV are not limited to RSUs and OBUs, as the network may include users pedestrians, drivers, and passengers. Therefore, the vehicles in SIoV maintain social relations among them and other network objects depending upon the mutual interests such as real-time traffic conditions on a particular road segment. Due to highly dynamic nature of SIoV, the network objects join and leave the networks very fast and the connections are based on the same travel route and similar configurations etc. Therefore, the objects in the SIoV are equipped with communication, storage and processing capabilities, hence, making them smart objects.

The dynamic nature of the network, real-time information sharing, and the smart nature of the network objects in SIoV causes an immense generation of data that needs to be communicated, processed and stored [5]. Therefore, the energy consumption by individual network objects increases due to individual communication, storage, and processing capabilities [6]. Furthermore, the decentralized nature of the network does not impose any restriction on redundant data generation, dissemination, and storage that eventually results in wastage of energy and resources [7]. The existing research in the SIoV does not focus on energy-efficiency framework for SIoV systems considerably.

The technological advancements in recent years have led to increased carbon dioxide emissions. In the domain of IoT, efficient utilization of energy is needed to enable a greener IoT environment. The major focus for a greener IoT environment is on the efficient deployment of the fixed network objects to reduce the energy cost [18]. Although, the enabling technologies for greener IoT environment such as green sensing networks and green tags exists, however, there is a lack of a framework that considers all the possibilities to optimally utilize power, especially in SIoV domain. The green IoT itself is considered as the enabling technology for scalable and sustainable development and has multiple benefits such as

environmental protection and customer satisfaction [19]. Power consumption in IoT systems is modeled at multiple levels such as measuring network energy, estimating data acquisition energy and modeling data processing energy [20]. Furthermore, multiple wireless energy harvesting techniques are available for empowering the self-sustained network with potentially an unlimited supply of power [21]. Therefore, the ground is all set to propose a novel energy-efficient framework for SIoV with the help of enabling technologies, energy prediction, and measurement models.

Therefore, the aim of this work is to propose an energy-efficient framework to efficiently utilize the fixed network objects to reduce the energy cost for SIoV. To achieve the aim of the study, the following research objectives are proposed:

Objective 1: To study state-of-the-art in energy efficiency and SIoV.

Objective 2: To propose an energy efficiency framework suitable for SIoV systems.

Objective 3: To evaluate the proposed framework in terms of optimal utilization of energy.

Objective 4: To compare the proposed framework with existing solution(s).

The rest of the paper is organized as follows: Section 2 presents state-of-the-art studies related to SIoV and energy efficiency. This is then followed by Section 3, which proposes an energy-efficient framework for SIoV. The experimental setup and results are presented in Section 3 and 4, respectively. The last section concludes the study, discusses future directions, and limitations of the study.

II. Related Work

Traditional IoV frameworks provide safety, infotainment and traffic efficiency related information that is shared among other vehicles and infrastructure [8]. Some of the existing studies also deal with social networks [7, 9]. Firstly, a few studies related to vehicular communication and SIoV are presented to show the immense amount of data that is generated and disseminated using the applications and services. Secondly, the major work related to energy efficiency in the domain of IoV and SIoV is described in this section. The current focus of energy efficiency is in three dimensions, which are, Internet, tags, and sensors. Fig. 1 shows the enabling technologies for green IoT.

The traffic congestion and road accidents are caused by the number of ever-increasing vehicles on the road. SIoV has the potential to resolve these issues by providing services and applications ranging from infotainment to security. According to authors in [9], SIoV consists of six components: tNote message, OBUs, RSUs, Home Base Units (HBU), tNote cloud and User Interface. A framework for social information sharing among vehicles is provided, however, redundant data generation is not handled which is the main cause of excessive energy consumption in the SIoV domain. An attempt [17] was made to reduce the overall power consumption of IoT system by deploying unmanned aerial vehicles (UAVs) that serve as base stations. The UAVs were used to collect data from IoT devices for reliable uplink communications. The study also presents a framework for the optimal placement of UAVs. However, this framework is not suitable for SIoV implementation because the nature of the vehicular environment varies significantly in terms of vehicular speed and road topology.

Fig. 1. Enabling Technologies for Green IoT [18].

Another model for VANETs exists in the literature which is referred to as VANET-cloud [10] that provides multiple services, applications and digital platforms to users at a relatively low cost [11]. Further, this model claims to improve safety by exchanging traffic data to vehicles and other network objects for the appropriate action in multiple traffic situations. Moreover, this model offers revenues to the drivers by sharing their onboard computing resources with other network objects. Modern autonomous vehicles may become the future of urban transportation for a relatively safer experience [12]. These autonomous vehicles are equipped with sensors that generate large amounts of data. The road is instrumented with components, RFID tags, and embedded microcontrollers. Vehicular cloud provides a communication and computing environment in order to provide services for autonomous vehicles. Hence the amount of data is constantly generated and processed.

Social vehicular network (SVN) technique consists of Vehicular Cloud (VC), RSU, and Internet Cloud (IC) components [8, 13]. Vehicular cloud establishes a connection with RSU. VC communicates with IC via RSU. However, no attention is given to minimizing the number of messages exchanged among the network objects. A Smart-Eye technique for a safer driving experience is presented in [14]. The information is shared in multiple formats ranging from text, video and coordinates. However, this technique lacks security and privacy along with the redundant data generation and dissemination causing the wastage of energy. Further, the existing studies advocate the use of fog computing to overcome the deficiency of central data processing related to traditional cloud computing by offloading the selected tasks on the edge of the network [15, 16]. Hence, it reduces the communication cost; thereby resulting in the consequent reduction in the communication energy usage. However, in the IoV domain, the decentralized nature of fog computing may cause redundant

data to be generated and processed eventually. Therefore, without a framework for energy efficiency, the mere use of fog computing cannot guarantee the optimal use of power and energy in communication and processing.

III. Proposed Framework

An energy-aware facilitation framework for SIoV is proposed in the current study to efficiently utilize the fixed network objects to reduce the energy cost. Specifically, the proposed framework offers energy-aware efficient deployment of RSUs based upon three energy consumption types: (i) network energy, (ii) data acquisition energy, and (iii) data processing energy. The proposed framework is depicted in Fig. 2. The input to the system is the network traffic and the output is energy-aware deployment of RSUs. In the framework, the resource request is received to RSUs from network traffic. After that, network energy is measured. Subsequently, data acquisition energy is estimated and data processing energy is modeled. Based on these three types of energies, intelligent energy-aware deployment of RSUs is proposed resulting in lower energy consumption, thus dropping the overall energy cost.

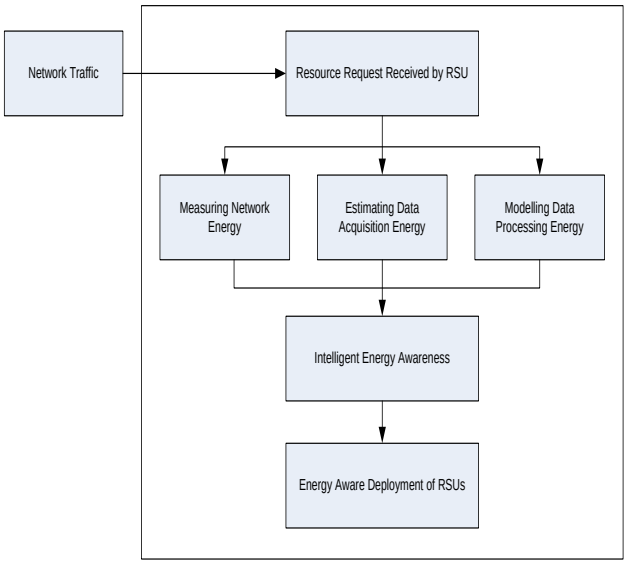


Fig. 2. Proposed Energy Facilitation Framework for SIoV.

The intelligent energy awareness in this framework can be obtained by using state-of-the-art artificial intelligence and machine learning techniques. One of the candidates in this regard is the solution of the classical traveling salesperson (TSP) problem based on the ant colony optimization algorithm (ACO) [22-25]. In the first step, a path with optimal energy consumption is calculated using the shortest Hamiltonian cycle. The second step consists of the ant colony optimization application. The two steps are explained below:

Given a collection of RSUs, the objective is to find the shortest Hamiltonian cycle connecting all the RSUs in a given set. So that, the acknowledgment should also be received. This is because the required resource might be present at any node which is connected to a particular RSU. This can be mapped as the classic traveling saleperson problem. The collection of the RSUs can be mapped onto a collection of cities in TSP. The path between RSUs having optimal energy consumption can be regarded as the shortest Hamiltonian cycle. The mathematical model of TSP in terms of SIoV is presented below:

Consider a weighted graph G=(R, C) where R is the set of n RSUs and C represents the connection linking the RSUs.

Each connection (i,j) ∈ C is associated with E_{ij} that represents the energy cost between RSUs i,j.

However, SIoV is a highly dynamic environment, therefore, the energy cost $E_{ij}$is bound to be changed. So,

$E(T)=\{ E_{ij} (T)\}n*n$

Here, E() is the energy cost and T is the period in which there is no change in the energy cost.

T=t/f , here t is the time and f encodes the frequency of the occurrence of the change.

In the current study, the famous ant colony optimization algorithm is used for optimization purpose. In the field of computer science, this algorithm is used to find the optimal path through a graph by using the probabilistic technique. The ACO can be applied to optimization problems to find the best path on a weighted graph. In SIOV terms, RSUs can be regarded as the nodes of a weighted graph whereas the edges between the nodes encode the energy consumption. Initially, ants are placed at RSUs randomly; the probability with which an ant "*k*" chases the next RSUs is given by the following Equation:

$$\rho_{i\,j}^{k}(t) = \frac{[\tau_{i\,j}(t)]^{\alpha} . [\eta_{i\,j}]^{\beta}}{\sum_{l \in j_i^k} [\tau_{i\,l}(t)]^{\alpha} . [\eta_{i\,l}]^{\beta}}$$

In the above equation τ_{ij} is the pheromone value deposited for a transition from one state to another, and η_{ij} is the reciprocal of E_{ij} (T) depicting the visibility, the parameter α is related with pheromone and β is associated with visibility information.

The pheromone value τ_{ij} is updated using the following equation:

$$\tau_{x\,y} \leftarrow (1 - \rho)\tau_{x\,y} + \sum_{k}^{m} \Delta\tau_{x\,y}^{k}$$

IV. Experimental Setup

Simulation of urban mobility (SUMO) tool is utilized in the work for traffic simulation as it is widely used for traffic traces generation and simulating road conditions in different scenarios. The focus of the study is urban scenario instead of highway scenario. A 9 Km2 area is selected for modeling urban traffic as shown in Fig. 3. Energy consumption values are randomly assigned to each RSU dynamically in the selected area based on their processing requirements. The Ant Colony optimization is implemented using Python 3.9. The experimental setup is depicted in Fig. 4. For the experimental purpose, the number of RSUs varies in the selected area, that is, five RSUs are incremented in every experiment, and therefore, the experimental results are attained with 5, 10, 20 and 25 RSUs.

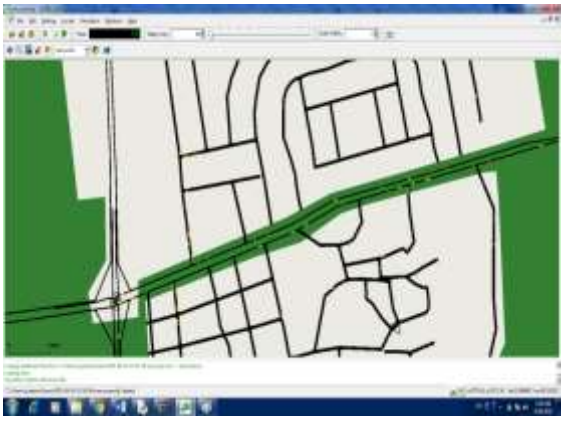


Fig. 3. Urban Scenario as Represented in SUMO.

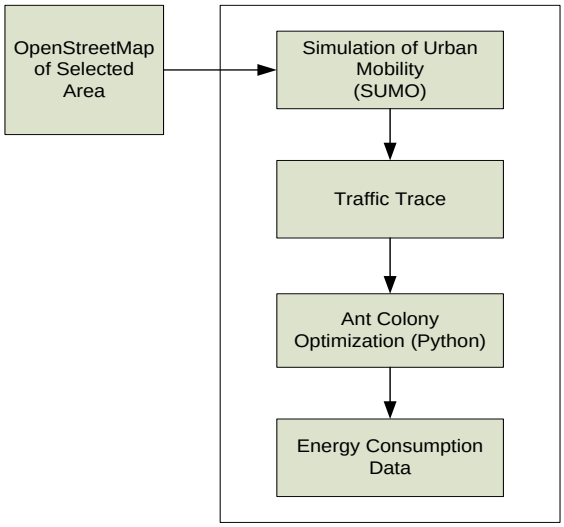


Fig. 4. Experimental Setup.

V. Experimental Results

This section presents the results of the experiments conducted by using the proposed framework. The energy consumption is calculated for different numbers of RSUS. Specifically, the energy consumed is measured with 5, 10, 15, 20, and 25 RSUs as depicted in Fig. 5. The experimental results support that the proposed framework saves energy in all experiments with different number of RSUs. The energy consumption with 5, 10, 15, 20, 25 RSUs is less with ant colony optimization as seen in Fig. 5. Therefore, it is concluded that optimization techniques can results in less energy consumption, thus, minimizing the overall cost.

With the increased number of RSUs (and hence communicating vehicles), the relative energy consumption is reduced. However, other optimization techniques may also be used to further validate the results of this study.

Fig. 6 exhibited the energy gain by conducting the experiments with energy-aware facilitation framework for SIoV. The considerable energy gain is accomplished with different numbers of RSUs. The highest energy gain with a value of 7 is achieved with 25 RSUs which is remarkable. Therefore, it can be inferred that the energy gain can be increased by adding more RSUs. However, it is also interesting to note that the second highest gain in energy (a value of 5) is a result of five RSUs. Therefore, we can say that the best energy gain is achieved with the deployment of minimum and maximum RSUs. On the other hand, a considerable energy gain of 3, 4, and 4 values is observed with 10, 15, and 20 RSUs, respectively. Therefore, the results of the study confirmed the energy saving by deploying the proposed energy-aware framework.

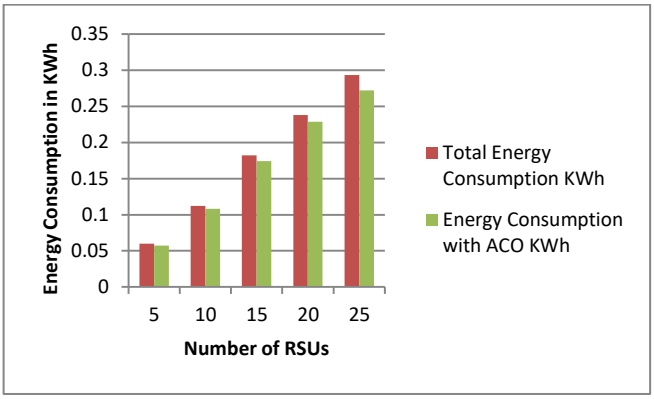


Fig. 5. Energy Consumed with different Number of RSUs.

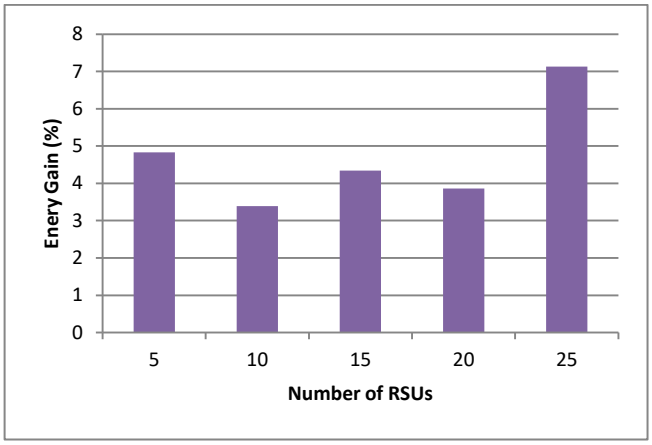


Fig. 6. Energy Gain.

VI. Research Implications

The expected outcomes of the study have implications for the scientific community, society and consumers. The implications of expected results in scientific community perceptive are many folds. First, the expected results can be utilized for the scientific community to design and develop new applications, standards and protocols based upon the recommendations of the proposed energy-efficient framework. Second, the new applications, standards and protocols in SIoV domains will be evaluated in terms of the amount of energy they consume. Third, the decision of whether to adopt a new application for commercial use may depend upon the evaluation from the proposed framework causing only energy-efficient applications to be available commercially resulting in less energy consumed in SIoV. Fourthly, redundant information generation, dissemination, and storage will be reduced, thus, saving cost as well as energy. In commercial use, energy saving and cost reduction play an important role in escalating the revenue of an organization. The main focus of

any business organization is to maximize revenue. Therefore, the proposed framework may have a huge impact on business organizations that want to accelerate their profit by reducing cost and energy consumption. Lastly, due to global energy crises, many countries, like Saudi Arabia will not rely on fossil fuel and currently looking for energy-efficient solutions. Therefore, the energy-efficiency is required as it has a huge impact on general public as well as the corporate section. In short, the greener technological advancements can be advantageous for both consumers and corporate and participating in the betterment of society at large.

VII. Conclusion, Future Work and Limitations

The innovation in network technology and IoV resulted in the conceptualization of SIoV. There are multiple objects including, RSUs, OBUS, vehicles, drivers, and passengers that have a social connection with each other for sharing information related to traffic condition and density, parking spaces and other resources in SIoV. These objects have communication, storage, and processing competencies to maintain social connections. The smart nature of network objects, real-time information sharing, and dynamic nature of network, redundant data generation resulted in the production of enormous data that require energy and resources for processing, storage, and communication. Therefore, the current work provides an energy-efficient framework to reduce the energy cost in SIoV. The proposed framework is based on network energy, data acquisition energy, and data modeling energy. The classical travel salesperson problem and ant colony optimization algorithm are applied for calculation of path requiring less energy from source to destination. The proposed framework is simulated in the urban setting within a 9 KM2 area. The experiments conducted with 5, 10, 15, 20, and 25 RSUs supporting significantly the energy gain achieved through the proposed method. Further, the total energy consumption is less by applying the proposed framework as highlighted by the experimental results. The study has some limitations. First, the proposed solution is evaluated on small scale by developing a real-world testbed application. Second, only simulation results will be presented. Future directions of this study include real-world evaluation of the framework with a higher number of RSUs.

Acknowledgment

This work was funded by the Deanship of Scientific Research (DSR), University of Jeddah, Jeddah, under grant No. ((UJ-02-066-DR). The authors, therefore, acknowledge with thanks DSR technical and financial support.

References

[1] M. A. Qureshi, et al., "A survey on obstacle modeling patterns in radio propagation models for vehicular ad hoc networks," Arabian Journal for Science and Engineering, vol. 40, pp. 1385-1407, 2015.

[2] M. A. Qureshi, et al., "A lightweight radio propagation model for vehicular communication in road tunnels," PloS one, vol. 11, p. e0152727, 2016.

[3] O. Kaiwartya, et al., "Internet of vehicles: Motivation, layered architecture, network model, challenges, and future aspects," IEEE access, vol. 4, pp. 5356-5373, 2016.

[4] T. A. Butt, et al., "Social Internet of Vehicles: Architecture and enabling technologies," Computers & Electrical Engineering, vol. 69, pp. 68-84, 2018.

[5] L. Maglaras, et al., "Social internet of vehicles for smart cities," Journal of Sensor and Actuator Networks, vol. 5, p. 3, 2016.

[6] G. Xiong, et al., "Cyber-physical-social system in intelligent transportation," IEEE/CAA Journal of Automatica Sinica, vol. 2, pp. 320-333, 2015.

[7] Z. Zhou, et al., "Social big-data-based content dissemination in Internet of vehicles," IEEE Transactions on Industrial Informatics, vol. 14, pp. 768-777, 2018.

[8] D. Kwak, et al., "Seeing is believing: Sharing real-time visual traffic information via vehicular clouds," IEEE access, vol. 4, pp. 3617-3631, 2016.

[9] K. M. Alam, et al., "Toward social internet of vehicles: Concept, architecture, and applications," IEEE access, vol. 3, pp. 343-357, 2015.

[10] S. Bitam, et al., "VANET-cloud: a generic cloud computing model for vehicular Ad Hoc networks," IEEE Wireless Communications, vol. 22, pp. 96-102, 2015.

[11] I. Ahmad, et al., "The role of vehicular cloud computing in road traffic management: a survey," in International Conference on Future Intelligent Vehicular Technologies, 2016, pp. 123-131.

[12] E.-K. Lee, et al., "Internet of Vehicles: From intelligent grid to autonomous cars and vehicular fogs," International Journal of Distributed Sensor Networks, vol. 12, p. 1550147716665500, 2016.

[13] M. Aloqaily, et al., "An auction-driven multi-objective provisioning framework in a vehicular cloud," in 2015 IEEE Globecom Workshops (GC Wkshps), 2015, pp. 1-6.

[14] D. Singh and M. Singh, "Internet of vehicles for smart and safe driving," in 2015 International Conference on Connected Vehicles and Expo (ICCVE), 2015, pp. 328-329.

[15] W. Zhang, et al., "Cooperative fog computing for dealing with big data in the internet of vehicles: Architecture and hierarchical resource management," IEEE Communications Magazine, vol. 55, pp. 60-67, 2017.

[16] K. Zhang, et al., "Mobile edge computing and networking for green and low-latency Internet of Things," IEEE Communications Magazine, vol. 56, pp. 39-45, 2018.

[17] M. Mozaffari, et al., "Mobile unmanned aerial vehicles (UAVs) for energy-efficient internet of things communications," IEEE Transactions on Wireless Communications, vol. 16, pp. 7574-7589, 2017.

[18] F. K. Shaikh, et al., "Enabling technologies for green internet of things," IEEE Systems Journal, vol. 11, pp. 983-994, 2017.

[19] M. Maksimovic, "Greening the future: Green Internet of Things (G-IoT) as a key technological enabler of sustainable development," in Internet of things and big data analytics toward next-generation intelligence, ed: Springer, 2018, pp. 283-313.

[20] B. Martinez, et al., "The power of models: Modeling power consumption for IoT devices," IEEE Sensors Journal, vol. 15, pp. 5777-5789, 2015.

[21] P. Kamalinejad, et al., "Wireless energy harvesting for the Internet of Things," IEEE Communications Magazine, vol. 53, pp. 102-108, 2015.

[22] Aadil, F., Ahsan, W., Rehman, Z.U., Shah, P.A., Rho, S. and Mehmood, I., 2018. Clustering algorithm for internet of vehicles (IoV) based on dragonfly optimizer (CAVDO). The Journal of Supercomputing, 74(9), pp.4542-4567.

[23] Ebadinezhad, S., Dereboylu, Z. and Ever, E., 2019. Clustering-based modified ant colony optimizer for internet of vehicles (CACOIOV). Sustainability, 11(9), p.2624.

[24] Zhang, J., Liu, B., Zhang, W. and Jiang, D., 2020, August. An IoV Route Planning Service Based on LEO Constellation Satellites. In International Conference on Simulation Tools and Techniques (pp. 194-203). Springer, Cham.

[25] Kumar, S., Solanki, V.K., Choudhary, S.K., Selamat, A. and González Crespo, R., 2020. Comparative Study on Ant Colony Optimization (ACO) and K-Means Clustering Approaches for Jobs Scheduling and Energy Optimization Model in Internet of Things (IoT). International Journal of Interactive Multimedia & Artificial Intelligence, 6(1).

The Role of Ontologies through the Lifecycle of Virtual Reality based Training (VRT) Development Process: A Review Study

Youcef Benferdia[1], Mohammad Nazir Ahmad[2], Mushawiahti Mustafa[3], Mohd Amran Md Ali[4]
Institute IR4.0, Universiti Kebangsaan Malaysia, Bangi, Malaysia[1, 2, 4]
Department of Ophthalmology, Faculty of Medicine Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia[3]

***Abstract*—The size of learning content continually challenges education and training providers. A recent advanced technology called Virtual Reality (VR) has emerged as a promising choice to facilitate knowledge acquisition and skill transfer in a variety of sectors. The main challenge in this technology is the increasing costs, time, effort, and resources needed for designing Virtual Reality based Training (VRT) applications as educational content. To fill such gaps, ontology approach was introduced to support VR development. Therefore, this review has the objective of investigating on how ontologies have been applied throughout the life cycle of a VR development process. Accordingly, articles from the year 2015 onwards have been explored. Findings show that VR developers do not incorporate ontology in all phases of the lifecycle of VR methodology, but only cover some phases like creation and implementation. Creating novel solutions without a complete methodology results in a long development process and an ineffective product. This could consequently raise high dangers in real life, especially when VRT is for fields containing trivial details that are vital for saving lives such as healthcare. This research thus presents a proposal of methodological guidance on designing VR applications with the use of an ontology approach throughout all the life cycles of VR construction.**

Keywords*—Virtual reality; ontology; methodology; training and learning

I. INTRODUCTION

In today's world, people have been surrounded by sophisticated technology that forces them to live on the edge of a technological revolution. Consequently, it will fundamentally change their life in terms of the way they live, study, and communicate with one another. This revolution has created new services and products in order to make their life easy. Nowadays many services can be remotely provided including ordering food, grab car, booking a flight, and training.

The VR, for example, is one of important enabling technologies for 4IR. It has increasingly attracted many researchers in several application domains such as social media, education, culture heritage, entertainment, training and healthcare. This technology "provides an immersive multimedia 3D simulation of real life, supports interactivity with the created environment and enables sensorial experiences" [1]. The immersive environment can be similar to the real world or it can be fantastical, creating an experience not possible in ordinary physical reality. VR has become a prevalent application in a variety of domains such as entertainment, tourism, e-commerce, education, and training. Recently it is most commonly used for the training and learning environment. VR offers important benefits by providing flexible and efficient training processes, which notably result in cost reduction and the removing of any risk associated with training in real environments.

Even though VR applications is not truly a new technology, its development is still in the earlier stages. The successful rate of its adoption in education for example, is not well reported in the literature [2]. Some reasons for that may be the cost [3], [4]; lack of understanding and capturing training scenarios and learning contents in an explicit manner, and so on [4], [5]. Besides that, VR is very knowledge-intensive. For example, capturing knowledge on the "know-how" among domain experts, conducting different training scenarios, where the absence of any details can lead to serious problems. Additionally, [3] the development process for the training and learning discipline is a tedious task and needs a long development life cycle, due to the much knowledge-intensive task and complex area required to be dealt with. To simplify a VR implementation, there is a requirement to have a modelling tool.

Ontology, in this context, is introduced to design a standardized conceptualization model at a high level of abstract and expressiveness with the purpose of offering a shared and common understanding of the domain. This tool can capture and represent training scenarios and the activities taking place in a Virtual Environment (VE) in an expressive manner. Unfortunately, in spite of the fact that ontology is developed for VRT in a variant of domains such as healthcare as reviewed in [6], presently, the existing VRT development methodologies do not pay much attention to having knowledge-based models explicitly, throughout the VR development life cycle. Designing ontology-based Unified Foundation Ontology (UFO), for example, is a significant step which models a large domain world that consequently makes it highly reusable across different domains. The underlying logic for adopting and generating foundation ontology, according to [7], is mainly to have a minimal collection of particular and generic concepts including key terms, proprieties, potential axioms and relationships. These concepts play a role as common knowledge, describing the real world, which significantly facilitates the extending and reusing of tasks that essentially promise the adoption across all domains. For example,

ophthalmology domain consists of many diseases such as cataract, glaucoma and so on. Let us assume that the ontology for cataract surgery has already been designed. If any ontology engineer wishes to design an ontology for glaucoma disease, they need to only use the existing one and extend it with the process of glaucoma surgery, because the domain knowledge is already there. Hence, as it is obvious that designing ontology for cataract domain provides many possibilities to reuse it several times in order to make a huge ontology that can cover the whole ophthalmology domain. Therefore, selecting the right upper ontology and its language is highly recommended. This would help promote the effectiveness of reusing an existing ontology, which would assist in drastically avoiding any redundant efforts and time spent for building up new virtual training scenarios.

Heaving in hand, guideline and user-friendly tools for ontological engineers, VR developers and domain experts throughout all phases of the VR development life cycle, this becomes critical to benefit the reduction of the needed time, costs, and efforts. The consequently leads to promoting VR adoption. This guideline is proposed to guarantee the systematic rigor in the designing and evaluation of VR application. Therefore, this paper attempts to answer the following question:

Q1. How is ontology used through the lifecycle of a VR development methodology?

The remnant of this paper is arranged as follows: Section II provides a brief definition of the ontology and describes the significant roles of ontologies for VRT. In Section III, research methods for selecting the primer studies are discussed. The results of this review are presented in Section IV. In Section V, a critical discussion on the findings is presented, Finally, Section VI presents the paper's conclusion and the new proposed guidance methodology.

II. Literature

A classic definition of ontology is "an explicit specification of a conceptualization" [8]. The conceptualization is an abstract simplified view of some selected part of the world (portion of a reality), containing concepts and their relationships between them. It exists on a community's mind as shared knowledge. The community members in this case are doctors, youths, VR experts and other stakeholders. An explicit specification of a conceptualization is an ontology, which is a conceptual model (knowledge-based model) and has been identified as a prominent tool to represent shared knowledge explicitly [9] [10]. Feilmayr & Wöß [11] recently defined ontology as "A formal, explicit specification of a shared conceptualization that is characterized by high semantic expressiveness required for increased complexity". This means that ontology is an abstract of a selected part of the world, which is unambiguously represented using a formal language. This representation should have high semantic expressiveness, should be shared among a variant of the stakeholders, and restricted to a definite domain of interest.

The training scenario and the learning content are the core part in any VRT application. The knowledge for this comes from domain experts in a particular area. Capturing high semantic and expressive training scenarios, on know-what and know-how, are crucial for an ideal VRT domain. Thus, in order to capture and represent this knowledge in sequence and in a coherent manner, ontologies were applied to a number of projects of VRT in a variety of fields. For example, in the Smart Home Simulator (SHS) project by Baldassini et al. [12], ontology was used to provide elder people a system, enabling them to follow a healthy lifestyle. It was applied to manage all heterogeneous data (e.g., devices, users, and environments). In order to ensure that users were following the suitable activities at home, a reasoning process was also enabled to query the desired data. BKOnto [13] was developed to support a virtual exhibition system, which was built based on biographical history. The aim of ontology here was to assist virtual presentation by offering structure descriptions and definitions that explicitly present the historical materials, places, and events. This ontology behaved as a storyline that enabled users to easily navigate a semantic web with the help of VR technology. Walczak and his colleagues [14] proposed a new approach in developing a VR training scenario for electrical operators with a help of ontology. They utilized two technologies including semantic web technology and VR Scenario Editor (VRSEd) application in order to facilitate knowledge representation. Finally, Liang and his team [15] attempted to develop a semantic framework to design collaborative animation for project art as shadow puppetry. This approach helped to minimize the intensive efforts and a long process for designing VRT application. It also assists to enhance the reusability of animation properties.

III. Method

In order to reach the objective of this research, an inclusive review method of published papers up to the year 2021 was conducted. Some key features of the SLR methodology were applied to support this study. The main objective of this review is to investigate how the proposed ontologies for VRT in general areas have been utilized to support the phases of VRT development methodology. At the first stage, therefore, a set of keywords were firstly applied in electronic bases in order to retrieve the first collection of papers. Accordingly, a variety of electronic databases were used including Science Direct, Springer, ACM, Web of Science, Emerald, Taylor & Francis, IEEE Xplore Digital Library, Wiley Online, and Google Scholar. As a result, the first set of papers were retrieved based on titles. During the Step 2, duplicated studies were removed using either Mendeley software or manually. In Step 3, keywords and abstracts were reviewed and papers were excluded when the inclusion criteria were not met. Abstracts with insufficient data were left to the next step. The complete text of extracted studies were analyzed in Step 4 using inclusion and exclusion criteria. These criteria were applied to extract the most relevant papers. Thus, the criteria were reviewed to include:

- Published studies between January 2015 and July 2021 were included.
- The articles included were related to ontology using VR for training and learning.

- Studies were excluded if they had been written in a language other than English, OR.
- Their designed ontology was focused in representing 3D content such as appearance, and geometry logic.

IV. RESULT

After the premier studies were selected, the articles were ready for the analysis and synthesis task. The next sections give summary descriptions of the proposed ontologies and their roles in the VR development process.

A. The Designed Ontologies in different Domains

In the following sections, a short description of each ontology in various sectors is provided along with an indication of the most significant design components such as methodology for developing ontology, tools, language, and so on (see Table I).

TABLE I. THE DESIGN COMPONENT OF ONTOLOGY AND ITS NATURE OF IMPLEMENTATION

Reference	Ontology Name	Area	Design Component				Evaluation
			Type of ontology	Language	Tool	Methodology	
Walczak et al. [14]	Ontology for VRSEd project	Industry	Domain ontology	RDF, RDFS and OWL	NM	NM	Yes
Tielman et al. [16]	Ontology for virtual coach	Healhacre	Domain ontology	Class diagram	NM	NM	Yes
Heyse et al. [17]	Ontology for VR Exposure Therapy (VRET)	Healhacre	Upper and domain ontology	Use case diagram, DL	NM	Co-design method	NM
Antoniou et al. [18]	ENTICE ontology	Healhacre	Domain ontology	RDF	NM	NM	NM
Dris et al. [19]	IVE ontology	Industry	Domain ontology	OWL	Protégé	Noy & McGuinness	Yes
Vincent et al. [21]	Inoovas ontology	Industry	Domain ontology	UML	NM	NeOn Methodology	Yes
Filho et al. [22]	Ontology for operator training simula-tor scenarios	Industry	Domain ontology	OWL-DL	Protégé	MCCA	Yes
Elenius et al. [23]	SAVE	Industry	Domain ontology	Flora2	Sunflower	NM	Yes
Liang et al. [15]	Ontology for virtual shadow play performance	Art	Domain ontology	OWL SWRL	Protégé	NM	Yes
Yeh and Huang [13]	BKOnto	Art	Domain Ontology	OWL	NM	NM	NM
Dragoni et al. [25]	PRESTO Ontology	Healthcare	Upper ontology	OWL	NM	NM	Yes
Baldassini et al. [29]	Ontology for SHS project	Healthcare	Domain Ontology	RDF, OWL	NM	NM	Yes

*NM: Not Mentioned

1) Ontology for VRSEd project: Walczak et al. [14], in their work proposed a new method in designing a VR training scenario for electrical operators with the help of semantic web technology. The latter technology enables knowledge representation. Both semantic modeling approach and the user-friendly VRSEd application were implemented as an expansion to Microsoft Excel. Domain experts were enabled to build training scenarios utilizing domain concepts defined by ontologies. RDF, RDFS and OWL standards were used to implement the scenario ontology. However, the tool and methodology are not mentioned in this paper. The new method was implemented and demonstrated as a desktop application for developing VR scenarios, which was further evaluated by domain experts.

2) Ontology for virtual coach: In Tielman et al. [16]'s work, an ontology-based question system was built in order to support a virtual coach. The latter technology was used to provide self-therapy for post- traumatic stress disorder patients, which enables patients to follow therapy at their own home. The vital side of this therapy is on how to assist patients in recollecting their traumatic memories. Ontology, therefore, was applied to support a dialogue system in virtual coach, where it was utilized to capture and represent knowledge and meaning of the real domain (see Fig. 1). In this paper, the ontology is presented using a class diagram, whereas methodology and tool are not mentioned. The ontology based system was evaluated using a within-subject experiment in order to confirm the performance of the ontology in helping patients to recollect their lost memories.

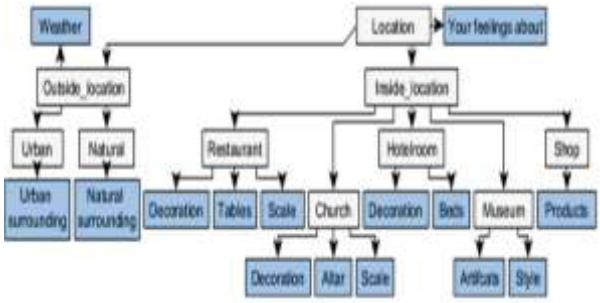


Fig. 1. Ontology of Holiday Moment Locations [16].

3) Ontology for VR exposure therapy (VRET): In this project [17], semantic ontology was designed with the purpose to model the necessary knowledge (e.g., concepts and relation) in a way to represent the domain of anxiety therapy in VE. The aim of the designed ontology was also to provide semantic reasoning in order to deduce essential knowledge from low-level data in VRET. This can be archived by using Description logic (DL) language, which permits to formulate such rules. The use case diagram was utilized to represent the proposed ontology which contains three layers of ontologies including foundation, domain, and application ontologies. The co-design method was applied to design the project's ontology, while a tool was missed in this study. The integration of the ontology inside VE and its evaluation was left to further work.

4) ENTICE ontology: The aim of immersive educational technology including Augmented, Virtual and Mixed Reality (VR/AR/MR communally XR), is to facilities skills acquisition and knowledge retention in the healthcare field. Designing XR immersive educational content is considered as the core challenge in terms of cost, effort, time, and resources required for developing. Therefore, Antoniou et al. [18] proposed an approach including the ENTICE ontology to enhance the content development and to facilitate the XR development process such as digital asset discoverability and reusability through visual authoring tools. The medical ontology terms were represented by using RDF. The integration of ontology into the XR environment and evaluation were planned as further research. The tool and methodology were not indicated in this project.

5) Interactive virtual environment ontology: Dris and his colleagues [19] in their work tried to propose an ontology that can help improve the use of Building Information Modeling (BIM) models as a Virtual Interactive Environment (VIE) generator. BIM is considered an approach that helps minimize the time spent in designing VE as a model for providing realistic 3D VE in the construction sector. In order to design this ontology, authors reused IFC (Industry Foundation Classes) ontologies [20] as the first step. The role of IFC ontology is to discover any possible incident of each fault that can be performed by modifying, adding, or removing objects inside the VE. Noy & McGuinness's methodology, Protégé software, and OWL were the ontology engineering components used to build this ontology (see Fig. 2). Three sheets of questionnaires were designed to evaluate the ontology's effectiveness. The first was conducted prior to training so as to classify the trainees in terms of knowledge and technology. The second was done after training to evaluate knowledge acquisition. Lastly, a month later, the same questionnaire was provided to them again.

6) Inoovas ontology: The Inoovas ontology was designed by Vincent et al. [21]. Its aim is to solve the problem of how all resources including people (mechanical or electrical designer, IT maintenance), heterogeneous software, and tools (VR, AR) can work together, when they remotely join in the procedure with an effective data exchange method in an Augmented and Virtual Reality (AVR) environment. Inoovas ontology represents the knowledge base of the company which contains three important parts. Fig. 3 presents the Real Thing that describes the physical parts of the system, data exchange with the system, and other classes of managing requirements on the system. Twin Thing represents the 3D model of the system. Lastly, Real and Twin Thing ontologies concerns with defining AR classes. Vincent et al. [21] used UML to represent the classes involved in the ontology. This paper used NeOn methodology, where the editor tool was missed in this paper. Inoovas was evaluated by developing an application named MProd. This application is grounded on the Inoovas ontology concepts and properties.

Fig. 2. Ontology OpenBIM based IVE Ontology- Extraction from Protégé [19].

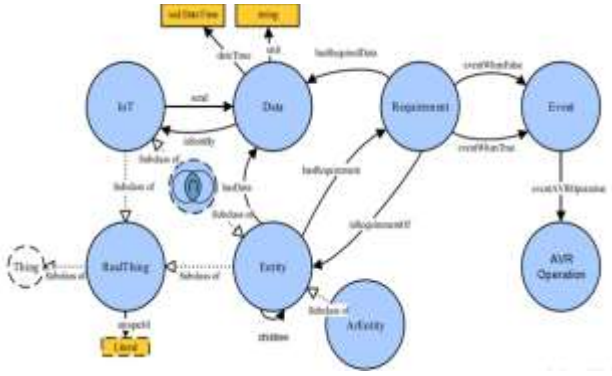


Fig. 3. Inoovas Ontology [21].

7) Ontology for operator training simulator scenarios project: This ontology is about designing error and training scenarios for electrical power system operators. The scenarios were developed according to accident reports (consisting of human error scenarios). According to this study, the construction of simulator training scenarios had posed various challenges. Dealing with multidisciplinary team of experts could cause several obstacles including producing implicit training scenarios and sharing a lack of knowledge among team members. Filho et al. [22] attempted to build an ontology that facilitates the development process of training scenarios and enhances common language sharing among stakeholders. The OWL-DL language was used to design ontologies, with the help of the Protégé tool. Incident Scenario Conceptual Model (MCCA) was the applied method to develop the ontology of this study. The designed ontology consists of two ontologies: training and error scenario ontology. A case study was the validation pathway to evaluate the ontology-based correctness and completeness of the terms.

8) SAVE ontology: Elenius et al. [23] designed a framework called Semantically Enabled Automated Assessment in Virtual Environments (SAVE). This ontology tries to provide an automated approach by using the semantic method. This helps describe or facilitate the action, event and rules including disassembling and assembling a Rifle. SAVE ontology reused Sunflower, which is an integrated development environment for ontologies and rules. Sunflower has a set of libraries and tools based on the Flora 2 language, which is a fully expressive language. Its root is based on OWL in descriptive languages. SAVE uses four components, namely, an ontology of components (physical objects), rules for creating components (and their sub-components), an ontology of actions, and rules for performing actions on components. However, the adoption methodology in this project again was not declared. In the evaluation part, all ontology models were tested by Subject Matter Experts (SMEs).

9) Ontology for virtual shadow play performance: In the VR domain, designing interactive animation is still a challenge and is labour-intensive. The reason for this is that during the development process many functional requirements need to be handled including massive data assets management, graphics, physics, etc. The purpose of Liang and his team [15]'s work was to design a semantic framework to develop collaborative animation for classical shadow play art (shadow puppetry). In the same way, it enables prompting reusability of animation properties. As a result, the development process was facilitated and extended. Two specific ontologies were built. The first one is Hand- and Gesture-Based Interaction Ontology (HGBIO) (see Fig. 4), and the second one is Digital Chinese Shadow Puppetry Assets Ontology (DCSPAO).

Having OWL enables integration of SWRL rules, which can be represented by utilizing SPARQL queries. The feasibility verification of ontology was performed using user experience tests of the ontology. At first, more descriptions of the operation of the system and 15 minutes of training were delivered to seven users. Then they were separated into two groups in order to conduct a qualitative test. The first three users tested ontology-based assets retrieval, while the other four users who were young children, examined the interaction comfort. As a result, both groups provided positive feedback regarding retrieval of material from animation resources, freedom of movement, ease of use, and naturalness.

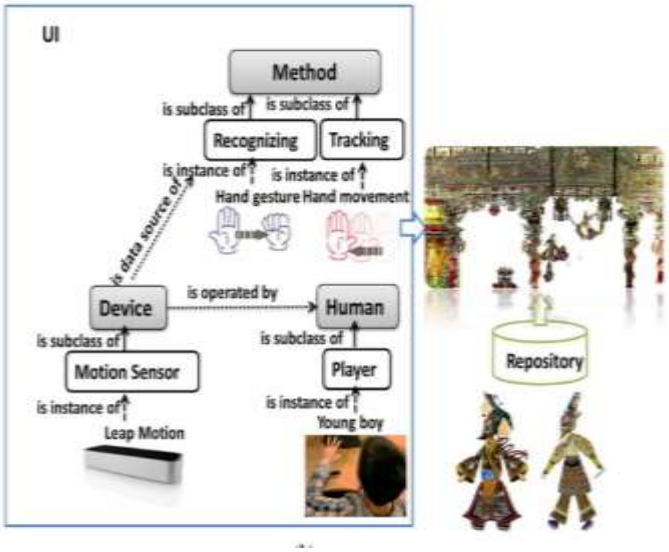


Fig. 4. Hand- and Gesture-based Interaction Ontology [15].

10)BKOnto ontology: Yeh and Huang [13] developed a virtual exhibition system based ontology knowledge. This ontology was designed based on biographical history, which is called BKOnto. The ontology's aim is to provide basic knowledge to assist virtual presentation. This ontology behaves as a storyline while assisting to provide structure definitions that systematically present the historical materials and events. BkOnto used the OWL mark-up language to describe cognitive knowledge bias for biographical historical material. This was further transformed into a VR exhibition space, which enabled users to easily navigate semantic structure in the form of a 3D space. This form can be reused in museums as a virtual exhibition that facilitates to manage several historical materials of semantic structure through the internet and provides visual experiences of a temporal event for internet users. The tool, methodology, and validated design of BkOnto were not indicated in this study.

11)PRESTO ontology: The PRESTO ontology was designed by Dragoni et al. [24] for the PRESTO (Plausible Representation of Emergency Scenarios for Training Operations) project. This project tried to describe the behaviour of an artificial agent into VE. Thus, the purpose of this ontology is to facilitate the development of a VR scenario and a character behavioral model. It enhances the source code's reusability, whereas the VR developers are plugged to a variety of source coding and underlying 3D-libraries. OWL language and lightweight ontologies were used to enhance semantics and provide explicit descriptions of existing

scenarios in a VR environment. DOLCE, as top level ontology, was applied to select the entities of a VR scenario. However, tools and methodology were not clarified in this study. For validation purposes, modellers, and developers were interviewed to evaluate the effectiveness and usefulness of the designed ontology-based system.

12)Ontology for smart home simulator project: Baldassini and his colleagues [12] presented an ontology for the Smart Home Simulator project (SHS). The main challenge was the ability to provide elder people a system that could enable them to follow an active and healthy lifestyle. Basically, the ontology was used to manage all heterogeneous data regarding users, surrounding devices, and environments. The reasoning tools in this conceptual model enable a query process that provides the desired data to ensure that users follow the proper activities. The designed ontology relies on three ontologies that are based on several languages. These languages are RDF and OWL. The ontology components, just like tools and methodology, were not clarified. Task-based evaluation was applied to check the usability and ergonomics of the system. A number of healthy subjects (from 25 to 30 years old) were used. The aim of this kind of evaluation is to test whether the intended tasks have been achieved or not.

13)VEULMoR ontology: Designing VR applications for upper limb motor rehabilitation is a difficult task. Designers are required to master various aspects including stroke-survivor, characteristic motor rehabilitation, interaction devices, and so on. Therefore, Ramírez-Fernández et al. [25] designed the VEULMoR ontology. The proposed ontology helped capture domain expert knowledge and presented it into the ontology. This approach shortened the time and facilitated the development of VR applications. The VEULMoR ontology was designed with the help of the Protégé editor, Methontology methodology and the use of the OWL language. The evaluation was implemented with therapists and patients in terms of patient safety and the administration of therapy.

B. Role of Ontology through the Life cycle of VRT

TABLE II shows how ontology is applied in each phase of the life cycle of a VR development methodology. The phases are adopted from the work reported in Polcar et al. [26]. This methodology is only chosen because it covers all the phases of the life cycle of developing VR; as such, this entirely enables an illustration of the role of ontology. The description related to each phase is provided as follows:

- Assignment Phase: In this phase, the domain experts and ontology engineers work together in order to design Ontology Requirement Specifications Documents (ORSD) that include defining the goal and the scope of designing the ontology, the intended end-users, and develop competency questions. The answering of competency questions and other requirements in ORSD assists in extracting the overall concepts of the VR application in terms of requirements and wishes, which helps to design a robust training scenario with high fidelity.
- Analysis Phase: This phase cares about the output of the previous phase, including the extracted concepts and other collected knowledge form relevant documents and observations. This knowledge should be analyzed to get the foundation upon which the overall concepts of the VR application are built. In the same way, all objects with similar or same functions and appearance should be classified. The result of this phase can include common and shared knowledge, glossaries of terms, list of actions and objects, and a story board of the detailed scenario.
- Creation Phase: In this stage, the ontology model is completely designed to be used through the rest of the VR's phases. Ontology becomes the major guiding force which assists the VR developer in selecting the right objects, properties, and level of detail. This phase is mainly concerned about constructing the assets of VE including scripts, texts, graphics, animations, sounds, and hardware. Some assets can be created according to the ontology model as well.
- Testing Phase: Coordinated validation and verification are conducted by SME and IT experts to check the connection between the first prototype of VR to the ontology model in terms of 3D graphical objects, tasks, terms, concepts and so on.
- Implementation Phases: The transformation of the ontology from a visual model (e.g., conceptual model) to implementation mode (e.g., OWL, Java) helps VR development in terms of providing decision making, exchanging of data, retrieving information and so on. The finished product (the integration of ontology within VR) will be verified by domain experts.
- Operation Phase: This is the end point where the ontology model is used as a reference to compare the intended objectives with the observed results. In this stage, VR designers and SMEs make a systematic verification of the ontology and 3D by inviting end users to evaluate the utility of the artefact, for maintenance purposes, in order to suggest future extending and verification.

It is obvious that the role of ontology is messed or misused throughout the phases of the lifecycle of the VR development methodology. A majority of the ontology engineers focus only in some phases like analysis, creation, and implementation. To conclude from above table, VR methodology does not incorporate ontology in all phases of its lifecycle.

Regarding the misuse of ontology, there should be a call for more attention to address this matter and give further guidance on how designers should apply ontology throughout all phases of VRT development. In the next section we will try to address these challenges and fill the gaps.

TABLE II. THE ROLE OF ONTOLOGY THROUGHOUT THE LIFECYCLE OF A VR DEVELOPMENT METHODOLOGY

Authors	Phases of VR					
	Assignment	**Analysis**	**Creation**	**Testing**	**Implementation**	**Operation**
Antoniou et al. [18]	Create a brainstorming deliverable	-To describe the related terms - To use Knowledge Organization System (SKOS)	- Using RDF to design ontology - Using ontology as a reference to develop some assets for VR	NM	NM	NM
Walczak et al. [14]	NM	NM	Using RDF, RDFS, and OWL to design ontology	NM	- Integrating ontology inside VR application	Verified by domain experts
Hyse et al. [17]	- To collect the essential knowledge - To define the scope of ontology - To create a list of competency questions	- To determine the most relevant knowledge by conducting workshops			- Integrating ontology inside VR application	NM
Teilman et al [16]	NM	NM	- Using class diagram to design ontology	NM	- Implemented in VR system	Ontology based system was evaluated by 24 healthy participants
Dris et al. [19]	To determine the domain and the scope of the ontology	Conducting acquisition of knowledge to select the main terms from documents, standards and the existing risk hunting courses	- Using OWL to design ontology - Using ontology as a reference to develop some assets for VR	NM	- Integrating ontology inside VR to improve interoperability	The utility of VR with ontology was evaluated by trainees
Vincent et al. [21]	NM	Getting concepts from experts and guidance procedures	- Using UML to design ontology - Using ontology to create some assets for VR	NM	- Integrating ontology inside VR to exchange data, in an interoperable way	Application based evaluation was conducted to evaluate the ontology effectiveness
Filho et al. [22]	NM	Building terms from literature	- Using OWL to design ontology - Using ontology to create some assets for VR	NM	- Integrating ontology inside VR to exchange data, in an interoperable way	Ontology within VR system was evaluated through a case study
Elenius et al. [23]	NM	Selecting terms and concepts from articles	- Using Flora code to design ontology - Using ontology as a reference to develop some assets for VR	NM	- Integrating ontology inside VR to provide reasoning	All ontology models were tested by SMEs
Liang et al. [15]	NM	Selecting terms and concepts from traditional Chinese shadow	- Using OWL to design ontology - Using ontology as a reference to develop some assets for VR	NM	- Integrating ontology inside VR to support ontology-based retrieval, which improves searching performance	The feasibility verification of ontology was performed using user experiences test
Yeh et al. [13]	NM	NM	- Using OWL to design ontology - Using ontology as a reference to develop some assets for VR	NM	- Integrating ontology to support VR presentations	NM
Dragoni et al. [24]	NM	Defining terms and concepts by using expert help	- Using OWL to design ontology - Using ontology as a reference to develop some assets for VR	NM	- Using ontology to describe the agent behavioral script - Testing the utility of VR with ontology	Modelers, and developers were interviewed to evaluate the effectiveness of the designed ontology-based system.
Baldassini et al. [12]	NM	Using use cases to select terms and concepts	- Using RDF and OWL to design ontology - Using ontology as a reference to develop some assets for VR	NM	- Implementing ontology to retrieve desired data about the domain	A number of healthy subjects (from 25 to 30 years old) were used to test the ontology-based system
Ramírez-Fernández et al. [25]	NM	Selecting terms and concepts from a contextual study and SLR	- Using classes and OWL to design ontology - Using ontology as a reference to develop some assets for VR	NM	- To use ontology for facilitating VE development	The evaluation was implemented with therapists and patients

*NM: Not Mentioned

*NM: Not Mentioned

V. DISCUSSION

The main objective of the study is to explore how ontology has been used to support VR development. Several ontologies have been proposed in the way to help the VR designing process. However, this review indicates that there is a general misuse on how to use ontology throughout all the stages of the methodology for VRT development. According to TABLE II, it is clear that most of the ontology engineers did not pay more attention to the assignment phase. Only a few studies had discussed about the scope and the objectives of ontology. However, they missed out on the inclusion of the ontology requirement specifications, which is as an agreement between the ontology engineer and the domain expert [27]. This approach enables the ontology engineer to include and exclude the most important concepts. Additionally, the involvement of domain experts is an essential part for knowledge acquisition; any lack of key experts can highly result in a partial model [11]. Domain experts should therefore be consulted in the earlier stages in order to avoid generating poor models which lack expressiveness, truthfulness, and details.

In the creation phase, most of the designed ontologies were represented using a tag or code. This approach breaks the Gruber's design criteria for ontology that suggests that the conceptualization should be represented at the knowledge level, which is free from any specific symbol-level encoding [8]. This kind of approach provides a good interaction among domain experts, the ontology team, and VR developers. This consequently conducts a better verification and validation process, since the ontology graphical model is frequently used to test the comprehensiveness of designed assets including 3D objects, scenarios, scripts and so on.

On the other hand, according to TABLE II, the test phase is totally missed. This evaluation is considered as an ex ant evaluation, which refers to evaluation of the prototype before the implementation stage. This is to avoid any kind of risk and effort before the design goes through construction [28]. Here the domain experts and VR developers verify whether the 3D graphical including scripts, scenarios and text, expressively reflect what is presented in ontology model. Thus, ignoring this phase can definitely lead to negative consequences such as conducting unnecessary redesigning or remodeling.

It can be seen from the above literature that ontology can play a significant role throughout all the stages of the methodology for VR development. It is, however, important to note the limitations of clearness on how ontology can go through all the stages. It may occur because, on one hand there is obvious missing or usage of immature methodology for designing ontology. On the other hand, it is the result of the lack of specific methodology of designing 3D modeling and VRT [29], [30], [31].

Over and above the latter shortages, it is difficult to guarantee the explicitness and truthfulness of training scenarios provided by VRT in the mentioned fields, because the ontology was missed or misused to be incorporated in all design phases of VR methodology. This could consequently raise high dangers in real life, especially when VRT is for areas involving trivial details that are important for saving lives. These sectors may include emergency response, healthcare, industry, army and so on.

VI. CONCLUSION

Despite the great decision on selecting ontology as a tool to support VR development solutions, based on the review, there is still a problem of implementing this tool throughout all the phases of life cycle of VR development process in various domains. The major barrier preventing VRT from being fully adopted is that most solutions are immediate, designed only for the current perspective purpose, without applying an effective methodology that could facilitate the construction process to be faster, cost effective, and create expressive training scenarios with minimal mistakes.

As mentioned above, it is not clear how ontology plays a role throughout the life cycle of VRT design. TABLE III briefly shows the connection between phases of VR and stages of ontology development methodology. It provides an idea about how ontology can play a significant role to facilitate VR construction. Therefore, further research can be conducted in this way to systematically define the right methodology of designing VRT that can easily cooperate with the role of ontology.

This paper proposes a novel guideline to design VRT applications. This approach, provided in this research, aids on making VR implementations faster, enables reduction of the required time, and effectively creates semantic learning content and tarring scenario for safe VRT. Consequently, it significantly improves user outcomes and promotes the use of VR in training.

TABLE III. CONNECTION BETWEEN METHODOLOGY FOR VR DEVELOPMENT AND ROLE OF ONTOLOGY

Phases of VR Methodology	The Role of Ontology
Assignment	1) The SMEs and ontology engineers work together in this phase. They make what is called ontology requirements specification that serves as an agreement between SME and ontology engineering. These ontology requirements include: - Define the goal and scope of designing ontology to be used in VR development. - Define the intended end-users. - Competency questions.
Analysis	2) The answers to competency questions from SME and other relevant knowledge from (e.g., documents, observation) are analyzed to get the foundation upon which the overall concept of the VR application is based in terms of requirements and wishes. Actions include: - Glossaries of terms - List actions and objects derived from the extracted knowledge. 3) The identified terms are not completely valuable as they exist in the current domain. Thus, they should be adjusted and reengineered. Some actions comprise: - Check and compare terms - Identify shared knowledge - Classify all objects – objects with similar or same functions and appearance. - List all activities and interactions in another list. - Define the states of objects. - Assign actions to the objects. - Draft the story board of the detailed scenario
Creation	4) Designing the ontology becomes compulsory to guide designers through the rest of the VR's phases. - Construct ontology model using visual modeling languages such as UFO and OntoUML 5) This artifact serves as the major guidance to assist the VR developer to select the right objects, properties and level of detail that he or she needs to represent by using multimedia modeling.
Testing	6) Ontology is used as a tool to facilitate communication among team members - Coordinated validation and verification are conducted by SME and IT experts to check the designed VR. - Verify the level of connection between the designed VR to ontology model in terms of 3D graphical object, tasks, terms, concepts and so on.
Implementation	7) Transforming ontology from visual model to implementation helps VR development. Actions include: - Provide reasoning process (e.g., decision making, student evaluation, retrieve information). - Enhance interoperability. 8) Conducting a test on the finished product.
Operation	9) Use the artifact as a reference to compare the intended objectives with the observed result: - The end users evaluate the utility of the artefact, for maintenance purposes, in order check any problems or difficulties. - A meeting between VR designers and SMEs is held to discuss experiences learned from this project and to determine future expansions of the ontology.

ACKNOWLEDGMENT

This research is supported by Transdisciplinary Research Grant Scheme (TRGS), Ministry of Higher Education (MOHE) and Universiti Kebangsaan Malaysia (UKM), Vot. No: TRGS/1/2020/UKM/02/6/2. We highly appreciate the enormous support received for this research project.

REFERENCES

[1] D. Bogusevschi, C. H. Muntean, and G.-M. Muntean, "Teaching and learning physics using 3D virtual learning environment: A case study of combined virtual reality and virtual laboratory in secondary school," J. Comput. Math. Sci. Teach., vol. 39, no. 1, pp. 5–18, 2020, [Online]. Available: https://ezp.waldenulibrary.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edo&AN=142858119&site=eds-live&scope=site.

[2] C. W. Chang, S. C. Yeh, M. Li, and E. Yao, "The introduction of a novel virtual reality training system for gynecology learning and its user experience research," IEEE Access, vol. 7, pp. 43637–43653, 2019, doi: 10.1109/ACCESS.2019.2905143.

[3] M. Cook, Z. Lischer-Katz, N. Hall, J. Hardesty, J. Johnson, R. McDonald, and T. Carlisle, "Challenges and strategies for educational virtual reality: results of an expert-led forum on 3D/VR technologies across academic institutions," Inf. Technol. Libr., vol. 38, no. 4, pp. 25–48, 2019, doi: 10.6017/ital.v38i4.11075.

[4] Y. Benferdia, M. N. Ahmad, M. Mustapha, H. Baharin, and M. Y. Bajuri, "Critical success factors for virtual reality-based training in ophthalmology domain," J. Heal. Med. Informatics, vol. 9, no. 3, pp. 1–14, 2018, doi: 10.4172/2157-7420.1000318.

[5] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, "A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda," Comput. Educ., vol. 147, no. 2020, p. 103778, 2020, doi: 10.1016/j.compedu.2019.103778.

[6] U. H. Mohamad, M. N. Ahmad, Y. Benferdia, A. Shapi'i, and M. Y. Bajuri, "An overview of ontologies in virtual reality-based training for healthcare domain," Front. Med., vol. 8, no. July, pp. 1–13, 2021, doi: 10.3389/fmed.2021.698855.

[7] L. Olsina, "Analyzing the usefulness of thingFO as a foundational ontology for sciences," in Proceedings of ASSE'20, 49 JAIIO, 2020, pp. 1–20, doi: 10.13140/RG.2.2.15135.59043/1.

[8] T. R. Gruber, "A translation approach to portable ontology specifications," Knowl. Acquis., vol. 5, no. 2, pp. 199–220, 1993, doi: 10.1006/knac.1993.1008.

[9] D. Dermeval, J. Vilela, I.I. Bittencourt, J. Castro, S. Isotani, P. Brito, and A. Silva, "Applications of ontologies in requirements engineering: A systematic review of the literature," Requir. Eng., vol. 21, no. 4, pp. 405–437, 2016, doi: 10.1007/s00766-015-0222-6.

[10] S. Khantong and M. Ahmad, "An ontology for sharing and managing information in disaster response: In flood response usage scenarios," Data Semant., pp. 1–14, 2020, doi: 10.1007/s13740-019-00110-6.

[11] C. Feilmayr and W. Wöß, "An analysis of ontologies and their success factors for application to business," Data Knowl. Eng., vol. 101, no. 2016, pp. 1–23, 2016, doi: 10.1016/j.datak.2015.11.003.

[12] D. Baldassini, V. Colombo, D. Spoladore, M. Sacco, and S. Arlati, "Customization of domestic environment and physical training supported by virtual reality and semantic technologies: A use-case," in IEEE 3rd International Forum on Research and Technologies for Society and Industry (RTSI), Modena Italy : IEEE, 2017, pp. 1–6.

[13] J. H. Yeh and X. M. Huang, "BkontoVr: A virtual reality exhibition system for biographic ontology-based semantic structure," in Proceedings of the 2018 2nd International Conference on Software and e-Business, S. Yang, and Y. Wang, Eds. Zhuhai,China:ACM, 2018, pp. 69–73, doi: 10.1145/3301761.3301775.

[14] K. Walczak, J. Flotyński, D. Strugała, S. Strykowski, P. Sobociński, A. Gałązkiewicz, F Górski, P Buń, P. Zawadzki, M. Wielgus, and R. Wojciechowski, "Semantic modeling of virtual reality training scenarios," in 17th EuroVR International Conference, EuroVR 2020, 2020, vol. 12499 LNCS, P. Bourdot, V. Interrante, R. Kopper, A.H. Olivier, H. Saito, G. Zachmann, Eds. Valencia, Spain: Springer, 2020, pp. 128–148, doi: 10.1007/978-3-030-62655-6_8.

[15] H. Liang, S. Deng, J. Chang, J. J. Zhang, C. Chen, and R. Tong, "Semantic framework for interactive animation generation and its application in virtual shadow play performance," Virtual Real., vol. 22, no. 2, pp. 149–165, 2018, doi: 10.1007/s10055-018-0333-8.

[16] M. Tielman, M. Van Meggelen, M. A. Neerincx, and W. P. Brinkman, "An ontology-based question system for a virtual coach assisting in trauma recollection," in 15th International Conference, IVA 2015, W.P. Brinkman, J. Broekens, D. Heylen, Eds. Delft, The Netherlands: Springer, 2015, pp. 17–27, doi: 10.1007/978-3-319-21996-7_2.

[17] J. Heyse, F. Ongenae, J. De Letter, A. All, F. De Bakcere, and F. De Turck, "Design of an ontology for decision support in VR exposure therapy," in 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, J. Meyer, L, Eds. Mamykina. Trento, Italy: EAI, 2019, pp. 1–4, doi: 10.4108/eai.20-5-2019.2283493.

[18] P. E. Antoniou, E. Chondrokostas, C. Bratsas, P.-M. Filippidis, and P. D. Bamidis, "A medical ontology informed user experience taxonomy to support co-creative workflows for authoring mixed reality medical education spaces," in 7th International Conference of the Immersive Learning Research Network (iLRN). Eureka, CA, USA: IEEE, 2021, pp. 1–8.

[19] A.-S. Dris, F. Lehericey, V. Gouranton, and B. Arnaldi, "OpenBIM based IVE ontology: An ontological approach to improve interoperability for virtual reality applications," in Advances in Informatics and Computing in Civil and Construction Engineering, I. Mutis, T. Hartmann, Eds. Chicago, United States: Springer, 2019, pp. 129–136, doi: 10.1007/978-3-030-00220-6_16.

[20] J. Beetz, J. Van Leeuwen, and B. De Vries, "IfcOWL: A case of transforming EXPRESS schemas into ontologies," Artif. Intell. Eng. Des. Anal. Manuf. AIEDAM, vol. 23, no. 1, pp. 89–101, 2009, doi: 10.1017/S0890060409000122.

[21] H. Vincent, J. Benoit, S. Xavier, and B. David, "Inoovas - industrial ontology for operation in virtual and augmented scene: The architecture," in Proceedings 2017 4th International Conference on Control, Decision and Information Technologies, CoDIT 2017. Barcelona, Spain: IEEE, 2017, pp. 300–305, doi: 10.1109/CoDIT.2017.8102608.

[22] F. T. Filho, Y. P. C. Aguiar, and M. D. F. Q. Vieira, "Ontology based modelling of operator training simulator scenarios from human error reports," in Proceedings of the 5th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH-2015), M.S. Obaidat, J Kacprzyk, L. Ören, Eds. Setubal, Portugal: SCITEPRESS 2015, pp. 279–288, doi: 10.5220/0005543502790288.

[23] D. Elenius, G. Denker, and M. Kim, "Semantically enhanced virtual learning environments using sunflower," in Research Conference on Metadata and Semantics Research, E. Garoufallou, I.S Coll, A. Stellato, J. Greenberg, Eds. Göttingen, Germany: Springer, 2016, vol. 672, pp. 81–93, doi: 10.1007/978-3-319-49157-8.

[24] M. Dragoni, C. Chidini, P. Busetta, M. Fruet, and M. Pedrotti, "Using ontologies for modeling virtual reality scenarios," in 12th European Semantic Web Conference, F. Gandon, M. Sabou, H. Sack, C. d'Amato, P. Cudré-Mauroux, A. Zimmermann, Eds. Portoroz, Slovenia: Springer, 2015, vol. 9088, pp. 575–590, doi: 10.1007/978-3-319-18818-8.

[25] C. Ramírez-Fernández, E. García-Canseco, A. L. Morán, and J. R. Gómez-Montalvo, "Evaluation results of an ontology-based design model of virtual environments for upper limb motor rehabilitation of stroke patients," in Proceedings of the 3rd 2015 Workshop on ICTs for improving Patients Rehabilitation Research Techniques, H.M Fardoun, P. Gamito, V.M.R Penichet, D.M Alghazzawi, Eds. Lisbon, Portugal: Association for Computing Machinery, 2015, pp. 105–108, doi: 10.3414/ME16-02-0017.

[26] J. Polcar, M. Gregor, P. Horejsi, and P. Kopecek, "Methodology for designing virtual reality applications," in Proceedings of the 26th DAAAM International Symposium. Vienna, Austria: DAAAM International Vienna, 2015, pp. 768–774, doi: 10.2507/26th.daaam.proceedings.107.

[27] M. C. Suárez-Figueroa, A. Gómez-Pérez, and B. Villazón-Terrazas, "How to write and use the ontology requirements specification document," in OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", R. Meersman. T. Dillon, P. Herrero, Eds. Vilamoura, Portugal: Springer, 2009, vol. 5871 LNCS, pp. 966–982, doi: 10.1007/978-3-642-05151-7_16.

[28] J. Venable, J. Pries-Heje, and R. Baskerville, "FEDS: A Framework for evaluation in design science research," Eur. J. Inf. Syst., vol. 25, no. 1, pp. 77–89, 2016, doi: 10.1057/ejis.2014.36.

[29] T. D. Parsons, T. Bowerly, J. G. Buckwalter, and A. A. Rizzo, "A controlled clinical comparison of attention performance in children with ADHD in a virtual reality classroom compared to standard neuropsychological methods," Child Neuropsychol., vol. 13, no. 4, pp. 363–381, 2007, doi: 10.1080/13825580600943473.

[30] P. B. L. Klavdianos, M. Parente, L. M. Brasil, and J. M. Lamas, "ONTO-MAMA: An unified ontology and 3D graphic model of the female breast anatomy," in Proceedings ofthe International Conference on Health Informatics (HEALTHINF-2012). Vilamoura, Algarve: SciTePress, 2012, pp. 106–116, doi: 10.5220/0003796401060116.

[31] J. Dascal, M. Reid, W.W, Ishak, B. Spiegel, J. Recacho, B. Rosen, and I. Danovitch, "Virtual reality and medical inpatients: A systematic review of randomized, controlled trials," Innov. Clin. Neurosci., vol. 14, no. 1–2, pp. 14–21, 2017.

Components and Indicators of Problem-solving Skills in Robot Programming Activities

Chacharin Lertyosbordin, Sorakrich Maneewan, Daruwan Srikaew
Faculty of Industrial Education and Technology
King Mongkut's University of Technology
Thonburi Bangkok, Thailand

***Abstract*—The objective of this research was to study the components and indicators of problem-solving skills in robot programming activities for high school students. This is done by analyzing the second order of confirmatory factor analysis (CFA) based on data from the behavioral assessment with regard to the robot programming activities of 320 students from specialized science schools. The results of the research revealed that the problem-solving skills in robot programming activities had five components and 15 indicators. All the components were tested for consistency using CFA statistics with the support of R-Studio program. The model analysis results were found to be consistent with empirical data with Chi-Square = 98.273, df = 80.000, p-value = 0.081, GFI = 0.961, NFI (TLI) = 0.924, CFI = 0.985, RMSEA = 0.027, RMR = 0.007. This indicates that all the identified components and indicators are involved in problem-solving skills in the robot programming activities of high school students.**

Keywords*—Components; indicators; problem solving; robot programming

I. Introduction

The Organization for Economic Cooperation and Development (OECD) [1] has published the OECD Future of Education and Skills 2030 report, indicating that robotics engineering is the number one task that the world will need. Consequently, inspiring students and encouraging learning with regard to robotics engineering should go along with the critical thinking and problem-solving skills [2] that are components of learning and innovation within the 21st Century Skills [3]. These are widely known in education, and are in line with the views of the World Economic Forum (2016) [4] which has defined critical thinking and problem-solving skills as the core competencies for students in 21st Century. In addition, both of these higher-order thinking skills are essential aspects of fostering skills across all learning and innovation skills groups [5].

In this research we focus on problem-solving skills. By this we mean the capability to use thinking methods based on knowledge and experience, in order to achieve the expected goals. This is done in a step-by-step fashion, by gathering and linking factors and facts [6 - 8]. These skills are considered to be the most important and fundamental skills for learners in the 21st Century [3,4]. Therefore, at present, the education profession is making great efforts to devise a learning management approach that can be used to improve problem-solving skills. Since 2018, the Ministry of Education of Thailand [9] has set the standard for developing thinking skills, calculation skills, analytical thinking skills and systematic problem-solving skills at the high school level. Moreover, the Ministry of Science of Thailand [10] has encouraged the organization of robot programming activities for high school students in order to support high school educational standards throughout the country.

Robot programming is an activity used as part of the learning process for the development of many important 21st Century skills [11 - 14], especially problem-solving, which are proven to respond well to programming [15 – 18]. At the present time, Thailand has no clear standards for assessing problem-solving skills in terms of robotics programming activities for high school students [19]. In addition, the researchers found that the search term "problem solving skill component and indicator on robot programming" does not appear in any Google Scholar databases from 2010 to 2020, nor are there any concrete elements and indicators. Therefore, this research studied the components and indicators of problem-solving skills in programming activities for high school students in Thailand to model prototype guidelines that can be used for student activity design and skill measurement in the future.

II. Research Objective

1) To study the components and indicators of problem-solving skills in robot programming activities for high school students.

2) To design the model hypothesis of components and indicators of problem-solving skills in robot programming activities for high school students.

3) To evaluate the validity of the model hypothesis of components and indicators of problem-solving skills in robot programming activities for high school students by analyzing the second order of confirmatory factor analysis.

III. Theoretical Framework

In conducting this research, the researcher reviewed the literature to synthesize the components and indicators of problem-solving skills in robot programming activities for high school students. This aspect is divided into two main parts: Part I, problem solving skills; Part II, robot programming procedures. The details are as follows:

A. Part I: Problem-solving Skills

Problem-solving skills refer to a capability to use thinking methods based on knowledge and experience to achieve an expected goal. This is done step-by-step, by gathering and linking factors and facts [6 - 8]. Consequently, many scholars have detailed the components of problem-solving skills including Bransford and Stein [20] who defined the problem-solving skill component as "IDEAL" which consists of 1) identify 2) define 3) explore 4) act and 5) look back. This conforms with the work of Foshay and Kirkley [21] who defining the solution components in terms of principles for teaching problem solving to be consistent with 1) identifying the problem 2) defining the problem through thinking about it and sorting out the relevant information 3) exploring solutions through looking at alternatives, brainstorming, and checking out different points of view 4) acting on the strategies 5) looking back and evaluating the effects of the activity. In addition, there are Polya's problem solving techniques [23] that is a generally-accepted problem-solving process in mathematics which consists of four steps to solve a problem. These are as follows: 1) understand the problem 2) devise a plan (translate) 3) carry out the plan (solve) 4) look back (check and interpret). The details with regard to the components of problem-solving skills from many other sources can be synthesized as shown in Table I.

From Table I, which is the result of the synthesis of the components of problem-solving skills gleaned from theories and academic articles, the researcher can conclude that the components of problem-solving skills consist of five elements:

1) Identifying the problem. This refers to explaining the details and the boundaries of the problem, and determining what needs to be solved.

2) Goal setting. This refers to sorting out related information which leads to things which need to be done for the problem to be resolved.

3) Creating a solution. This refers to looking for alternative ways to resolve a particular problem through brainstorming and reviewing to create a problem solution.

4) Acting on the solution. This refers to the implementation of the created solution.

5) Returning to check the results. This refers to the assessment of the results, and evaluating the success in order to ensure the problem has been solved.

B. Part II: Robot Programming Procedures

Robot programming refers to the control of a robot by writing computer programs which can be used to create and instruct the mechanical device using electronic systems to perform a desired task [26 - 29]. Nowadays, many scholars have detailed the procedures of programming. For example, Sharma [30] described the Software Development Life Cycle (SDLC) as a process that programmers use to create productivity outcomes. These consist of 1) requirement analysis 2) planning 3) software design 4) software development 5) testing and 6) deployment. These conform to the five-step programming process outlined in Wikibooks [32]: 1) clarifying / defining the problem 2) designing the program, 3) coding the program, 4) testing the program, and 5) documenting and maintaining. In addition, the School of Computer Science at the University of Birmingham [36] provides a simple four step programming procedure: 1) identify the problem 2) design a solution 3) write the program and 4) check the solution. The detail of the programming procedures from many other sources can be synthesized as shown in Table II.

TABLE I. SYNTHESIZATION OF PROBLEM-SOLVING SKILLS

Problem-solving's Components	IDEAL Aspects [20]	PLATO Learning Aspects [21]	Jonassen [8]	Great Schools Partnership Aspects [22]	POLYA Aspects [23]	Astuti, Suranto & Masykuris [24]	Franestian, Suyanta &Wiyono [25]
Identifying the problem	✓	✓	✓	✓	✓	✓	✓
Goal setting	✓	✓	✓	✓	✓	✓	✓
Creating a solution	✓	✓	✓	✓	✓	✓	✓
Acting on solution	✓	✓	✓	✓	✓	✓	✓
Returning to check results	✓	✓	✓	✓	✓	✓	✓

TABLE II. SYNTHESIZATION OF ROBOT PROGRAMMING PROCEDURES

Robot programing Procedure	Sharma [30]	Valenzuela [31]	Wikibooks [32]	Amjo [33]	Person [34]	Department of Computer Science and Statistics The University of Rhode Island [35]	School of Computer Science University of Birmingham [36]
Identify the Problem	✓	✓	✓	✓	✓	✓	✓
Design a Solution	✓	✓	✓	✓	✓	✓	✓
Code the Program	✓	✓	✓	✓	✓	✓	✓
Test the Program	✓	✓	✓	✓	✓	✓	✓
Implement the Program	✓	✓	✓	✓	✓	✓	✓

From Table II, the researcher can conclude that synthesizing the robot programming procedures consists of five steps:

1) Identifying the problem. This refers to analyzing the problem and determining the Input and Output processes that need to be incorporated in order to solve the problem.

2) Designing a Solution. This refers to arranging the order of the algorithms used by drawing flowcharts or writing pseudo code.

3) Coding the program. This refers to converting the instructions and sequence of methods from the flowchart to computer language.

4) Testing the program. This refers to the validation of the grammar of the computer language and the interpretation of the results for computer operation purposes. In addition, it involves testing for compatibility with hardware, including the Input and Output Processes.

5) Implementing the program. This refers to the results based on the program used. This should be followed up with further improvements.

From studying the problem-solving skill components and the robot programming procedures as shown above, the researcher can create a conceptual framework with regard to problem-solving skill components, and the indicators in robot programming activities. This framework is as shown in Fig. 1.

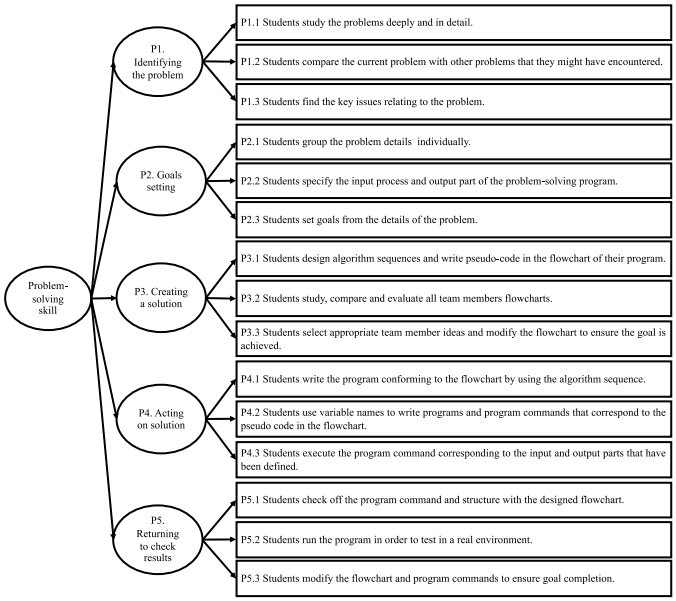


Fig. 1. Conceptual Framework with Regard to Problem-Solving Skill Components and the Indicators in Robot Programming Activities (Model Hypothesis).

IV. Research Methodologgy

A. Population

In this research, the population consisted of students in a group of the Princess Chulabhorn's Science High Schools (one of the Thai governments specialized scientific schools) who had completed a course in "robot programming" in the academic year 2020. This involved three schools in the northern province, three schools in the central province, three schools in the northeastern province, and three schools in the southern province; a total are 12 schools: each of which had 40 students in grade 10, and 40 students in grade 11, a total are 80 students per school. Consequently, the population was 960.

B. Samples

The researcher performed multi-stage sampling, starting with a sampling unit of three schools which was randomized to two schools per province, for a total of eight schools and 640 students. Then, stratified random sampling was used to select 30 students in grade 10 and 30 students in grade 11. As a result, the sample group was 60 students per school, totaling 480 students. The final step was taking a random sample of 40 students per school, and a total net sample of 320 students. This number conformed to the sampling size suggested by Yamane [37] and related to the appropriate sample size for structural equation model (SEM) statistics [38].

C. Research Instrument

In this research, the researcher used behavioral frequency self-assessment in terms of robot programming activities which was validated by seven experts. The assessment design was based on the Likert 5-point Scale to divide the behavior frequency (Level 5 = Always, 4 = Usually, 3 = About half the time, 2 = Seldom and 1 = Never).

D. Research Experts

To validate the research instrument in the form of a behavioral frequency self-assessment instrument in terms of robot programming activities, the researcher worked with seven experts from various fields, whose qualifications were as follows:

1) Two Lecturers in educational evaluation.
2) Two Lecturers in Computer engineering.
3) One Lecturer in Educational technology.
4) One Lecturer in Psychology.
5) One Psychiatrist with at least 5 years of adolescent behavior research experience.

V. PROCEDURES

The research was conducted in five steps as follows:

1) Study of the theories and research related to "Problem-solving skills" and "Robot programming" to identify the components and indicators of problem-solving skills with regard to robotics programming activities to obtain model hypotheses.

2) Creation of behavioral assessment in order to identify problem-solving skills in terms of the robot programming activities. This was based on the components and indicators of problem-solving skills in robot programming activities that followed the researcher's model hypotheses.

3) Evaluation of the validity of the behavior assessment instrument by using the Index of Item-Objective Congruence (IOC) involving seven experts.

4) Collection of behavior assessment data (Sum of the sample made up of 320 students in eight schools.).

5) Analyzing the data using confirmatory factor analysis (CFA) with the support of R-Studio software to examine the construct validity of the problem-solving skill components and indicators with regard to robot programming activities.

VI. RESEARCH RESULTS

1) To create and evaluate the validity of the behavioral frequency assessment instrument, seven experts determined the validity of the questionnaire in robot programming activities for high school students using the Index of Item-Objective Congruence (IOC). The results had been shown in the Table III conclude that the statements in the model hypothesis are acceptable.[39].

2) From Table IV, it can be seen that the 5 components have a standardized factor loading (β) greater than 0.4, and consequently passed the standard statistically criteria [40]. In addition, by examining the goodness of it statistical indicator, the Chi-Square value was 98.27 at 80.00 degrees of freedom (df), with a probability (p-value) of 0.08. In addition, Goodness of Fit Index (GFI), Adjusted Goodness of Fit Index (AGFI), Normed Fit Index (NFI) and Comparative Fit Index (CFI) were 0.96, 0.94, 0.92 and 0.99, respectively. The Standardized Root Mean Square Residual (SRMR) was 0.04 and the Root Mean Square Error of Approximation (RMSEA) was 0.03. It can be concluded that all 15 components and 5 indicators with regard to problem-solving skills in the robot programming activities fit the empirical data, and correspond to the researcher's hypothesis [41].

3) From Table V and Fig. 2, it can be seen that 15 indicators have a standardized factor loading (β) greater than 0.4. Consequently, they achieved the standard statistically criteria [40]. As a result, these 15 indicators could be appropriate for use as element of all five components.

TABLE III. SYNTHETIZATION OF PROBLEM-SOLVING SKILLS EVALUATION OF THE VALIDITY OF THE QUESTIONS IN THE BEHAVIORAL FREQUENCY ASSESSMENT

Robot programming procedure	**Problem – solving skills**		**IOC**	**Meaning**
	Component	**Indicator**		
Identify the problem	P1. Identifying the problem	P1.1 Students study the problems deeply and in detail	1.00	Accept
		P1.2 Students compare the current problem with other problems that they might have encountered	1.00	Accept
		P1.3 Students find the key issues relating to the problem	1.00	Accept
	P2. Goals setting	P2.1 Students group the problem details individually	1.00	Accept
		P2.2 Students specify the input process and output part of the problem-solving program.	1.00	Accept
		P2.3 Students set goals from the details of the problem.	1.00	Accept
Design a solution	P3. Creating a solution	P3.1 Students design algorithm sequences and write pseudo-code in the flowchart of their program.	1.00	Accept
		P3.2 Students study, compare and evaluate all team members flowcharts.	1.00	Accept
		P3.3 Students select appropriate team member ideas and modify the flowchart to ensure the goal is achieved.	1.00	Accept
Code the program	P4. Acting on solution	P4.1 Students write the program conforming to the flowchart by using the algorithm sequence.	1.00	Accept
		P4.2 Students use variable names to write programs and program commands that correspond to the pseudo code in the flowchart.	1.00	Accept
		P4.3 Students execute the program command corresponding to the input and output parts that have been defined.	1.00	Accept
Test the program / Program implementation	P5. Returning to check results	P5.1 Students check off the program command and structure with the designed flowchart.	1.00	Accept
		P5.2 Students run the program in order to test in a real environment.	1.00	Accept
		P5.3 Students modify the flowchart and program commands to ensure goal completion.	1.00	Accept

TABLE IV. FACTOR LOADING OF THE INDICATORS FOR EACH COMPONENT

Components	**βi**	**bi**	**S.E.**	**R^2**
P1	0.88	1.00	-	0.77
P2	0.86	0.94	0.13	0.75
P3	0.94	0.82	0.12	0.88
P4	0.56	0.52	0.08	0.32
P5	0.61	0.46	0.08	0.38
Chi-Square = 98.27, df = 80.00, p-value = 0.08, GFI = 0.96, AGFI = 0.94, NFI = 0.92, CFI = 0.99, SRMR =0.04, RMSEA = 0.03				

TABLE V. FACTOR LOADING OF PROBLEM-SOLVING SKILL COMPONENTS IN ROBOT PROGRAMMING ACTIVITIES

Com.	**P1**			**P2**			**P3**			**P4**			**P5**			R^2
Ind.	βi	bi	S.E.	βi	bi	S.E.	βi	bi	S.E.	βi	bi	S.E.	βi	bi	S.E.	
P1.1	0.67	1.00	-													0.45
P1.2	0.63	0.92	0.10													0.36
P1.3	0.67	0.99	0.11													0.44
P2.1				0.65	1.00	-										0.41
P2.2				0.57	0.86	0.11										0.33
P2.3				0.49	0.79	0.12										0.24
P3.1							0.51	1.00	-							0.26
P3.2							0.59	1.13	0.17							0.35
P3.3							0.65	1.23	0.15							0.42
P4.1										0.71	1.00	-				0.50
P4.2										0.61	1.00	0.10				0.38

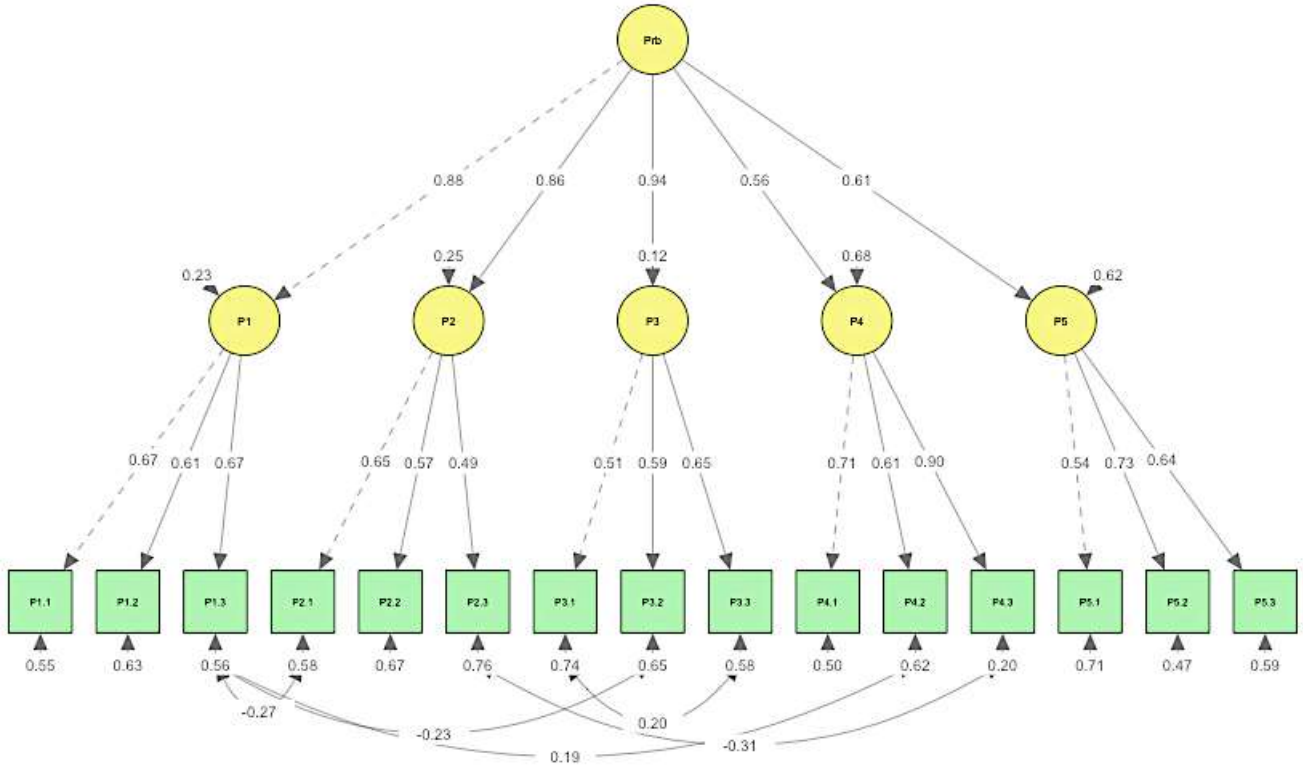


Fig. 2. Factor Loading Diagram of Components and Indicators of Problem-solving Skill in Robot Programming Activities on the Part of High School Students.

VII. Conclusion

From the results, the researcher can summarize that the components and indicators with regard to problem-solving skills in robot programming activities for high school students, are composed of five components and fifteen indicators. The details of the supporting information are as follows:

1) Component P3 (Creating a solution) had the highest factor loading at 0.94, followed by component P1 (Identifying the problem) which had a loading factor of 0.88. The third component was P2 (Goals setting) which had a loading factor of 0.86. In addition, the factor loading of the 3 components was more than 0.7, implying to these 3 components have sufficiency variance effected on problem-solving skill [42]. Moreover, the conclusion confirms to the meaning of the phrase "Problem-solving-approach" as defined by the APA Dictionary of Psychology [43]. That is, it is “The process whereby difficulties, obstacles, or stressful events are addressed through the use of coping strategies.” Accordance to Jonassen [8], programming activities could be classified as one solution for design problem-solving that focuses on analysis and planning. This also corresponds to Chandrasekaran [44], who states that the key to problem-solving is the critical thinking step, in order for the student to understand the problem, and determine the structure and sequence of work to fit the problem.

2) The factor loading of component P4 (Acting on the solution) and component P5 (Returning to check results) were 0.56 and 0.61, respectively. This means that both of these components are important when it comes to problem-solving skills and is statistically acceptable [40]. Similarly, McFadden [45] demonstrated that programming must follow the plan strictly in order to achieve its goals. In addition, the continuous review and development of the results can increase problem-solving fluency. In line with Taheri, Sasaki and Ngetha [46], we can conclude that problem-solving skills can be built from the programming process, while repeating trials are also a part of the development of problem-solving ability. Moreover, Glenda and Westhuizen [47], also suggest that programming can demonstrate logical reasoning, numerical reasoning and language reasoning, all of which are essential components of problem-solving skills.

3) Indicator P1.1 (Study the problems deeply and in detail), P1.2 (Compare the current problem with other problems that they might have encountered), and P1.3 (Find the key issues relating to the problem) have factor loadings of 0.67, 0.63 and 0.67, respectively. These three indicators are statistically acceptable, which means that all the indicators under component P1 (Identifying the problem) are elementary problem-solving skills [40]. According to Kember [48], identifying the problem means the perception of the problem, and the development of a clear problem framework from digesting information. This leads to effective and accurate problem identification. This in line with Staniewicz [49] who said that problem identification is the first step in coming up with an engineering solution by gathering previously used similar concepts to those of the present problem, then identifying what should be improved.

4) Indicator P2.1 (Group the problem details individually), P2.2 (Specifying the input process and the output of the problem-solving program), and P2.3 (Set goals from the details of the problem) have factor loadings of 0.65, 0.67 and 0.49, respectively. These three indictors are statistically acceptable, which means that all the indicators under component P2 (Goals setting) are elementary problem-solving skills [40]. This conforms with Wharton Executive Education [50], which indicates that a good solution should not stick by a lot of problem information, but should be manage with factor-and-effect relationships to set a new definitive scope for the problem. Moreover, Markman [51] explained that problem resolution targeting involves classifying a large number of existing problems and looking for a correlation in terms of the information to achieve inventive and precise solutions.

5) Indicator P3.1 (Designing algorithm sequences and write pseudo-code in the flowchart of their program), P3.2 (Study, compare and evaluate all team members’ flowcharts), and P3.3 (Select appropriate team members’ ideas and modify the flowchart to achieve the goal.) have factor loadings of 0.51, 0.59 and 0.65, respectively. The fact that these three indictors are statistically acceptable, means that all the indicators under component P3 (Creating a solution) are elementary problem-solving skills [40]. According to Crews & Ziegler [52], writing flowcharts with regard to algorithmic sequencing and pseudo-coding was a process of great importance with regard to engineering solutions. Additionally, Bryant [53] encouraged the use of the Critique, Explore, Compare, and Adapt framework in the process of designing, in that it can give results that are both comprehensive and well-informed.

6) Indicator P4.1 (Write the program conforming to the flowchart by using the algorithm sequence), P4.2 (Use variable names to write programs and program commands that correspond to the pseudo code in the flowchart.), and P4.3 (Execute the program command corresponding to the input and output parts that have been defined) have factor loadings of 0.71, 0.61 and 0.90, respectively. That these three indictors are statistically acceptable means that all the indicators under component P4 (Acting on a solution) are elementary of problem-solving skills [40]. According to Whipp, Tenkanen and Heikinheimo [54], computer programming is a step-by-step aspect of the problem-solving process. Naming each variable and choosing the correct command will support the solution of the problem more easily. It also conforms with Bilotta and Pantano [55], who explained that programming must consider the number and type of sensors which are part of the input, as well as the number and size of motors that are part of the output, to achieve the robot mission satisfactorily.

7) Indicator P5.1 (Check off the program command and structure with the designed flowchart), P5.2 (Run the program by testing in a real environment), and P5.3 (Modify the

flowchart and program commands to achieve the goal completely) have factor loadings of 0.54, 0.73 and 0.64, respectively. That these three indictors are statistically acceptable means that all the indicators under component P5 (Returning to check results) are elementary of problem-solving skills [40]. This conforms with Young, Sharlin and Igarashi [56], who explained that robot control programming needs validation results in terms of the robot's performance in a real environment. The program developer will systematically collect data with regard to the robot's performance in order to improve the robot operation in order to complete the mission. In addition, when solving a robot foot walking problem, Maicon, Aramizo, Houman and Mohammad [57] found that using a robot errors recovery programming technique involving equation generation could reduce the errors in the next operation, and solve the problem more quickly.

VIII. Discussion

A. Integration of the Emerging Model with the Literature

The components and indicators of problem-solving skills were evaluated by confirmatory factor analysis (CFA). It was related to the result of synthesization of problem-solving skills by literature review [20-25]. Based on theoretical studies and research articles related to problem-solving skills and robot programming, the researchers were able to summarize the five components of problem-solving skills: 1) identifying the problem, 2) setting goals, 3) creating a solution, 4) acting on the solution and 5) checking the results. In addition, the researchers concluded that the robot programming procedure consists of five steps: 1) identifying the problem, 2) designing a solution, 3) coding the program, 4) testing the program, and 5) implementing the program.

B. Reflection on Methodology and Limitations

Because problem solving skills are explicitly a latent variable, the researchers had to transform latent variables into observable variables by using the behavioral statement in the problem-solving process. In addition, the researchers apply the statement with samples only found in science/research institutions. This point is because data for accurate CFA processing in this research is required only from those expected to have the required skills in robot programming.

C. Suggestions for Future Research

1) This research uses the confirmatory factor analysis (CFA) technique to study the structural validity of the components and indicators. In the next step, creating a standardized skill measurement instrument means that information can be tested for content validity and the accuracy of the questions checked before applying them to each context.

2) In the case of constructivism learning theory, which is popular in education systems dealing with such aspects as problem-based learning (PBL), project-based learning (PjBL) or challenge-based learning (CBL) [58], all the focus is on developing higher order thinking skills. Therefore, the researcher suggests that the five components and fifteen indicators identified in this study could be used in the design of learning and evaluation processes in modern education.

Acknowledgment

This research article was supported form "Petchra Pra Jom Klao Ph.D" Research Scholarship from King Mongkut's University of Technology Thonburi.

References

[1] OECD. (2020). Learning Compass 2030 - OECD Future of Education and Skills 2030. Oecd.org. Retrieved 5 January 2020, from https://www.oecd.org/education/2030-project/teaching-and-learning/learning/learning-compass-2030/.

[2] OECD. (2020). Future shocks and shifts: challenges for the global workforce and skills development. Retrieved 5 January 2020, from https://www.oecd.org/education/2030-project/about/documents/Future-Shocks-and-Shifts-Challenges-for-the-Global-Workforce-and-Skills-Development.pdf.

[3] Partner for 21st Century Skills, "Framework for 21st Century Learning. (2009). Framework for 21st Century Learning. Teacherrambo.com. Retrieved 5 January 2020, from https://www.teacherrambo.com/file.php/1/21st_century_skills.pdf.

[4] Economic Forum Executive. (2018). The Future of Jobs Report 2018. World Economic Forum. Retrieved 5 January 2020, from https://www.weforum.org/reports/the-future-of-jobs-report-2018.

[5] Canter, A. (2004). A Problem-Solving Model for Improving Student Achievement. Casponline.org. Retrieved 7 January 2020, from https://www.casponline.org/pdfs/pdfs/rti0004.pdf.

[6] Bloom, B. (1984). Taxonomy of educational objectives. Longman.

[7] TENOPYR, M., Guilford, J., & HOEPFNER, R. (1966). A factor analysis of symbolic memory abilities: studies of aptitudes of high-level personnel. By M. L. Tenopyr, J. P. Guilford and R. Hoepfner. University of Southern California, Psychological Laboratory.

[8] Jonassen, D. (2011). Learning to solve problems. Routledge.

[9] Institute for the Promotion of teaching Science and Technology. (2018). Summary of curriculum and indicators Information Technology and Communication Curriculum 2008 and Technology (Computational Science) Curriculum Improvement 2018. Retrieved 9 January 2020, fromhttp://oho.ipst.ac.th/download/mediaBook/cs-ict.pdf.

[10] National Science and Technology Development Agency. (2020). Space Flying Robot Programming Challenge 2020 (SRPC2020). NSTDA. Retrieved 9 January 2020, from https://www.nstda.or.th/jaxa-thailand/krpc2020.

[11] Alimisis, D. (2013). Educational robotics: Open questions and new challenges. Themes In Science & Technology Education, 6(1), 63-71. Retrieved 10 January 2020, from https://www.researchgate.net/publication/284043695_Educational_robotics_Open_questions_and_new_challenges.

[12] Kanbul, S., & Uzunboylu, H. (2017). Importance of Coding Education and Robotic Applications For Achieving 21st-Century Skills in North Cyprus. International Journal Of Emerging Technologies In Learning (Ijet), 12(01), 130. https://doi.org/10.3991/ijet.v12i01.6097.

[13] Demertzi, E., Voukelatos, N., Papagerasimou, Y., & Drigas, A. (2018). Online Learning Facilities to Support Coding and Robotics Courses for Youth. International Journal Of Engineering Pedagogy (Ijep), 8(3), 69. https://doi.org/10.3991/ijep.v8i3.8044.

[14] Tuomi, P., Multisilta, J., Saarikoski, P., & Suominen, J. (2017). Coding skills as a success factor for a society. Education And Information Technologies, 23(1), 419-434. https://doi.org/10.1007/s10639-017-9611-4.

[15] Grover, S., & Basu, S. (2017). Measuring Student Learning in Introductory Block-Based Programming. Proceedings Of The 2017 ACM SIGCSE Technical Symposium On Computer Science Education. https://doi.org/10.1145/3017680.3017723.

[16] Romero, M., Lepage, A., & Lille, B. (2017). Computational thinking development through creative programming in higher education.

International Journal Of Educational Technology In Higher Education, 14(1). https://doi.org/10.1186/s41239-017-0080-z.

[17] Silapachote, P., & Srisuphab, A. (2017). Engineering Courses on Computational Thinking Through Solving Problems in Artificial Intelligence. International Journal Of Engineering Pedagogy (Ijep), 7(3), 34. https://doi.org/10.3991/ijep.v7i3.6951.

[18] Topalli, D., & Cagiltay, N. (2018). Improving programming skills in engineering education through problem-based game projects with Scratch. Computers & Education, 120, 64-74. https://doi.org/10.1016/j.compedu.2018.01.011.

[19] Lertyosbordin, C., Maneewan, S., Yampinij, S., & Thamwipat, K. (2019). Scoring Rubric of Problem-Solving on Computing Science Learning. International Education Studies, 12(8), 26. https://doi.org/10.5539/ies.v12n8p26.

[20] Bransford, J., & Stein, B. (1993). THE IDEAL PROBLEM SOLVER A Guide for Improving Thinking, Learning, and Creativity. Tntech.edu. Retrieved 13 February 2020, from https://www.tntech.edu/cat/pdf/useful_links/idealproblemsolver.pdf.

[21] Foshay, R., & Kirkley, J. (2003). Principles for Teaching Problem Solving. Vcell.ndsu.nodak.edu. Retrieved 13 February 2020, from http://vcell.ndsu.nodak.edu/~ganesh/seminar/2003_Foshay_PLATO%20Learning%20Inc._Tech%20Paper%20%234_Principles%20for%20Teaching%20Problem-Solving.pdf.

[22] Great schools partnership. (2016). Performance Indicators for Problem Solving. Great schools partnership. Retrieved 14 February 2020, from https://www.greatschoolspartnership.org/wp-content/uploads/2017/01/PDFTaskModelforProblemSolvingNov22-2016.pdf.

[23] Nurkaeti, N. (2021). Polya's Strategy: An Analysis of Mathematical Problem Solving Difficulty in 5th Grade Elementary School. Retrieved 2 May 2021, from. http://doi.org/10.17509/eh.v10i2.10868.

[24] Astuti, F., Suranto, S., & Masykuri, M. (2019). Augmented Reality for teaching science: Students' problem solving skill, motivation, and learning outcomes. Jurnal Pendidikan Biologi Indonesia, 5(2). https://doi.org/10.22219/jpbi.v5i2.8455.

[25] Franestian, I., Suyanta, & Wiyono, A. (2020). Analysis problem solving skills of student in Junior High School. Journal Of Physics: Conference Series, 1440, 012089. https://doi.org/10.1088/1742-6596/1440/1/012089.

[26] Fernández, E. (2015). Learning ROS for robotics programming. Packt Publishing.

[27] Owen-Hill, A. (2018). Python vs C++ vs C# vs MATLAB: Which Robot Language is Best. robodk. Retrieved 10 March 2020, from https://robodk.com/blog/robot-programming-language/.

[28] Plant Automation-technology. (2020). Different Types of Robot Programming Languages," Plant Automation-technology. Plantautomation-technology.com. Retrieved 12 March 2020, from https://www.plantautomation-technology.com/articles/different-types-of-robot-programming-languages.

[29] The University of Sheffield. (2020). Programming skills for robotics. FutureLearn. Retrieved 13 March 2020, from https://www.futurelearn.com/courses/robotic-future/0/steps/29368.

[30] Sharma, M. (2017). A Study of SDLC to Develop Well Engineered Software. International Journal Of Advanced Research In Computer Science, 8(3), 502-523. https://doi.org/https://doi.org/10.26483/ijarcs.v8i3.3045.

[31] Valenzuela, J. (2018). Computer programming in 4 steps | ISTE. Iste.org. Retrieved 10 June 2020, from https://www.iste.org/explore/Computer-Science/Computer-programming-in-4-steps.

[32] Wikibooks. (2021). The Computer Revolution/Programming/Five Steps of Programming - Wikibooks, open books for an open world. En.wikibooks.org. Retrieved 11 March 2020, from https://en.wikibooks.org/wiki/The_Computer_Revolution/Programming/Five_Steps_of_Programming.

[33] Amjo. (2018). Six steps in the programming process | Dotnetlanguages. Dotnetlanguages.net. Retrieved 10 March 2020, from https://www.dotnetlanguages.net/six-steps-in-the-programming-process/.

[34] Person, M. (2020). Programming Development Process: 4 Steps to Better Programming STEP 1: Define and Analyze the Problem. Academia.edu. Retrieved 13 May 2020, from https://www.academia.edu/11050507/Programming_Development_Process_4_Steps_to_Better_Programming_STEP_1_Define_ and_Analyze_the_Problem.

[35] Department of Computer Science and Statistics. Computer Programming. The University of Rhode Island. Retrieved 10 May 2020, from https://homepage.cs.uri.edu/faculty/wolfe/book/Readings/Reading13.htm.

[36] School of Computer Science. The Programming Process. University of Birmingham. Retrieved 10 March 2020, from https://www.cs.bham.ac.uk/~rxb/java/intro/2programming.html.

[37] Yamane, T. (1974). Statistics ; an introductory analysis. Harper & Row.

[38] Wolf, E., Harrington, K., Clark, S., & Miller, M. (2013). Sample Size Requirements for Structural Equation Models. Educational And Psychological Measurement, 73(6), 913-934. https://doi.org/10.1177/0013164413495237.

[39] Turner, R., & Carlson, L. (2003). Indexes of Item-Objective Congruence for Multidimensional Items. International Journal Of Testing, 3(2), 163-171. https://doi.org/http://doi.org/10.1207/S15327574IJT0302_5

[40] Hair, J. (2010). Multivariate data analysis. Pearson Education.

[41] Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural Equation Modeling: Guidelines for Determining Model Fit. Electronic Journal On Business Research Methods, 6(1), 53-60. Retrieved 2 May 2021, from https://www.researchgate.net/publication/254742561_Structural_Equation_Modeling_Guidelines_for_Determining_Model_Fit.

[42] Statistics Solutions. (2020). Factor Analysis - Statistics Solutions. Statistics Solutions. Retrieved 15 June 2020, from https://www.statisticssolutions.com/factor-analysis-sem-factor-analysis.

[43] APA. (2020). APA Dictionary of Psychology. APA Dictionary. Retrieved 5 June 2020, from https://dictionary.apa.org/problem-solving-approach.

[44] Chandrasekaran, B. (1990). Design Problem Solving: A Task Analysis. AI Magazine, 11(4), 59-71. https://doi.org/https://doi.org/10.1609/aimag.v11i4.857.

[45] McFadden, C., McFadden, C., Wendorf, M., Engineering, I., & Bergan, B. (2020). How to Think like a Programmer When Problem Solving. Interestingengineering.com. Retrieved 13 May 2020, from https://interestingengineering.com/how-to-think-like-a-programmer-when-problem-solving.

[46] Taheri, S., Sasaki, M., & Ngetha, H. (2015). Evaluating the effectiveness of problem solving techniques and tools in programming. 2015 Science And Information Conference (SAI). https://doi.org/10.1109/sai.2015.7237253.

[47] Barlow-Jones, G., & van der Westhuizen, D. (2017). Problem Solving as a Predictor of Programming Performance. Communications In Computer And Information Science, 209-216. https://doi.org/10.1007/978-3-319-69670-6_14.

[48] Kember, J. (2018). Problem Identification - FastBridge. FastBridge. Retrieved 11 June 2020, from https://www.fastbridge.org/2018/01/problem-identification/.

[49] Staniewicz, S. (2013). Problem Identification in Engineering Design | Electrical and Computer Engineering Design Handbook. Sites.tufts.edu. Retrieved 15 June 2020, from https://sites.tufts.edu/eesenior designhandbook/2013/problem-identification-in-engineering-design/.

[50] Wharton Executive Education. (2015). How to Identify the Real Problem - Nano Tools for Leaders. Wharton Executive Education. Retrieved 14 June 2020, from https://executiveeducation.wharton.upenn.edu/thought-leadership/wharton-at-work/2015/06/identify-the-real-problem/.

[51] Markman, A. (2017). How You Define the Problem Determines Whether You Solve It. Harvard Business Review. Retrieved 18 June 2020, from https://hbr.org/2017/06/how-you-define-the-problem-determines-whether-you-solve-it.

[52] Crews, T., & Ziegler, U. The flowchart interpreter for introductory programming courses. FIE '98. 28Th Annual Frontiers In Education Conference. Moving From 'Teacher-Centered' To 'Learner-Centered'

Education. Conference Proceedings (Cat. No.98CH36214). https://doi.org/10.1109/fie.1998.736854.

[53] Bryant, D. (2021). Critique, Explore, Compare, and Adapt (CECA): A New Model for Command Decision Making. DTIC. Retrieved 18 June 2020, from https://apps.dtic.mil/sti/citations/ADA605875.

[54] Whipp, D., Tenkanen, H., & Heikinheimo, V. (2020). Good coding practices - Selecting "good" variable names. Geo-python.github. Retrieved 10 July 2020, from https://geo-python.github.io/site/notebooks/L1/gcp-1-variable-naming.html.

[55] Bilotta, E., & Pantano, P. (2021). Some problems of Programming in Robotics (pp. 209-220). Cozenza, Italy. Retrieved 2 May 2021, from https://www.ppig.org/files/2000-PPIG-12th-bilotta.pdf.

[56] Young, J., Sharlin, E., & Igarashi, T. (2013). Teaching Robots Style: Designing and Evaluating Style-by-Demonstration for Interactive Robotic Locomotion. Human–Computer Interaction, 28(5), 379-416. https://doi.org/10.1080/07370024.2012.697046.

[57] Ficanha, E., Ribeiro, G., Dallali, H., & Rastgaar, M. (2016). Design and Preliminary Evaluation of a Two DOFs Cable-Driven Ankle–Foot Prosthesis with Active Dorsiflexion–Plantarflexion and Inversion–Eversion. Frontiers In Bioengineering And Biotechnology, 4. https://doi.org/10.3389/fbioe.2016.00036.

[58] Lynce, M. (2017). What is the Difference Between Problem, Project, and Challenge Based Learning? - The Edvocate. The Edvocate. Retrieved 17 July 2020, from https://www.theedadvocate.org/difference-problem-project-challenge-based-learning/.

A Hybrid Ensemble Word Embedding based Classification Model for Multi-document Summarization Process on Large Multi-domain Document Sets

S Anjali Devi, S Sivakumar
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur, Andhra Pradesh, India -522502

***Abstract*—Contextual text feature extraction and classification play a vital role in the multi-document summarization process. Natural language processing (NLP) is one of the essential text mining tools which is used to preprocess and analyze the large document sets. Most of the conventional single document feature extraction measures are independent of contextual relationships among the different contextual feature sets for the document categorization process. Also, these conventional word embedding models such as TF-ID, ITF-ID and Glove are difficult to integrate into the multi-domain feature extraction and classification process due to a high misclassification rate and large candidate sets. To address these concerns, an advanced multi-document summarization framework was developed and tested on number of large training datasets. In this work, a hybrid multi-domain glove word embedding model, multi-document clustering and classification model were implemented to improve the multi-document summarization process for multi-domain document sets. Experimental results prove that the proposed multi-document summarization approach has improved efficiency in terms of accuracy, precision, recall, F-score and run time (ms) than the existing models.**

Keywords*—Word embedding models; text classification; multi-document summarization; contextual feature similarity; natural language processing

I. INTRODUCTION

Machine learning (ML) has become a key approach to problem solving and data predictions. Machine learning allows a classifier to learn a set of rules, or the criterion of decision, from a set of labelled data that an expert has annotated. This approach enables better scaling and reduced time when classifying topic domain data as compared to a system that relies only on manual input. Most of the research was done on binary classifiers in the field of machine learning based on the classification of multi-domain document data. In many fields, the purpose of using machine learning for pattern mining has an important role in decision-making systems. A set of input documents is split into two or more classes in the text classification (TC) process [1], with each document belonging to one or more classes depending on its contents. Document clustering [2] is the method of categorizing text documents into a hierarchical cluster or category, so that the documents are identical in the same cluster, whereas the documents in the other clusters are different. It is one of the vitals of text mining processes. In particular, text mining has gained significant significance and involves various tasks, such as the development of granular taxonomies, document summarization, etc., to produce knowledge of higher quality from text. The supervised strategy is utilized to solve the problem if we have a predetermined class or classes. A prediction-based model is a decision tree. It is distinguished by a tree-like system of rules and is mostly used to solve classification problems. The decision tree is built using data from training. With this strategy, a tree is built to represent the categorization problem. The majority of previous works [3] used single-document summarization. Approaches based on sentence extraction from documents are used in single-document summarization. Most single-document summarization systems employ a simple method for summary generation, which consists of extracting the first sentence from each paragraph and placing them in the same order as they were written. Later on, the presence of multiple sources delivering the same information causes problems for news providers' end users, who must read the same material repeatedly. As a result, recent work [4] has centred on multi-document summarization. To combine information held in distinct documents for multi-document summarization, valuable procedures are necessary. This usually means that some operations, such as key matches, matching terms, sentence position, and sentence length, must be performed below the sentence level. As a summary [5], multi-document summarization may successfully handle the concerns by generating shorter summaries including the important points of the original documents using criteria for decreasing redundancy and maximising variety in the selected articles. Before reordering the phrases into the document's original sequence, most extractive summary optimisation algorithms score them based on their value. Without access to the real summary analysis mechanism, it is not always possible to build partial rank lists of sentences using only the original document and the summary. The two major types of text summarization are abstraction and extraction [6]. The sentence with the highest score among the other sentences is chosen

during the document extraction process. Whereas, abstraction entails employing linguistic techniques to create something new, which may or may not be present in the source, and substituting it for the summary without altering its original meaning. The entire collection is searched for important objects in the extractive summarization task, with no changes to the objects themselves. Conciseness, accuracy, and objectivity are three qualities of a good summarizer. The goal of this paper's proposed methodology [7] is to create an extractive text summarizer that can generate variable-length summaries. According to [8], the summary frequently includes sentences that are not closely related to one another. This can be handled by generating the sentence set with a sufficient threshold. As a result, one of our issues is deciding on a sufficient threshold. The order of the sentences in the summary is the next problem. Another challenge with news summarization systems is how to handle huge feature sets, as the complexity of weight adjustment increases exponentially as the number of features increases. As a result, higher-performance systems with more useful features are required. Among the three types of summarization systems, extractive summarization is perhaps the most investigated. Although the phrase is most commonly used to refer to sentence extraction and reordering, numerous extractive approaches also focus on sub-sentence extraction. An extractive system can be topic-based, centrality-based, or a combination of the two. The relevance of particular words or phrases is prioritized in topic-based systems. Although specialized machine learning techniques such as neural networks (NN) and support vector machines (SVM) are used in many fields to classify data into one or more classes, traditional models must be improved on large datasets with high dimensionality. Some demonstrations of supervised learning include Linear Regression, Logistic Regressions, Decision Trees, and SVM. These are some demonstrations of supervised learning. Classification [9] can be defined as the procedure of classifying objects of interest into different previously defined categories or classes.

Recently, extractive single document summaries have been generated using machine learning methods. Nave-bayes, Hidden Markov Model (HMM), and log-linear models are some of the methodologies that fall within machine learning approaches. Automatic text summarization using artificial intelligence and neural networks has been the subject of a few studies. Given a set of features, the Hidden Markov model [HMM] estimates [10] the posterior probability that each phrase is a summary sentence or not. This model has fewer assumptions of independence than the naive Bayesian approach. The number of terms in the sentence, as well as the likelihood of the terms given the baseline of terms (Baseline term probability) [11] and the document terms (Document term probability).

Wrapper techniques use a black box for a single learner to evaluate the function subsets on the basis of their predictive effectiveness. The embedded techniques select the features in the integrated phase and are generally particular to one individual instance. PSO and neural action provide a possible optimization solution [12]. Each particle accelerates during each iteration towards the best global location discovered by the representative points. Scalability is inefficient at identifying the globally optimal solution. Dynamic goals and connectivity are taken as tasks rather than restrictions. The Multi-Objective Data Relations (MODP) approach is used to resolve all existing problems in order to improve anomaly relations. Further work can be undertaken in the future in order to significantly reduce the normalized root-mean-square error. Recently, ensemble learning models have become popular and widely accepted for high-dimensional and imbalanced datasets. Most of the traditional ensemble classification models are processed with limited feature space and small data size. As the size of the feature space increases, traditional ensemble classifiers select a predefined number of features for classification. The main objective of the ensemble learning models based on feature selection is to classify high-dimensional features on high-dimensional datasets with high computational efficiency and a high true positive rate [13]. Severe problems such as performance and scalability may result from learning classification models with all their high-dimensional features. Many textual content classifiers [14] have been proposed in the literature, including those that use machine learning techniques, probabilistic models, and so on. Decision trees, nave-bayes, rule induction, neural networks, nearest [15-17], and, most recently, guide vector machines are some of the techniques used.

The main contributions in this paper are:

1) Proposed a hybrid multi-domain glove optimization model on the large document sets.

2) Proposed a multi-document clustering method for the document summarization process.

3) Implemented a hybrid multi-document Bayesian approach based document summarization process on large document sets.

The main sections presented in this paper are:

Section II describes the overall literature work of the word embedding models and multi-document summarization. In the section III, a hybrid word embedding measures are proposed in order to classify the multi-domain features for the multi-document summarization process. Also, a hybrid multi-document cluster based classification model is proposed in the section 3. In the section IV, experimental results and its discussion are discussed. Finally, in the section V, conclusion of the work is presented.

II. Related Work

Wu et.al, proposed key extraction by combining multidimensional information, and they named their proposed system MIKE. They used two datasets from the ACM world wide web to form the ACM knowledge discovery and data mining. They compared their results to the TF-IDF and TextRank algorithms to assess their performance [18]. LAKE is a key phrase based summarizer system that extracts relevant key phrases from documents using statistical analysis. In terms of text summarization methods, neural networks outperform other traditional methods in terms of extractive methods for handling semantics and redundancy, but fall short in terms of coherence when compared to abstractive methods. There are various approaches to abstractive summarization,

including linguistic-based approaches, semantic graph-based approaches, and hybrid extractive/abstractive approaches [19]. Syntactic representations and tree structures are used in linguistic-based approaches, but semantic meanings are not abstracted. As discussed in a previous study, semantic graph-based approaches focus on semantic role labelling to determine the abstraction of input to core meaning to form graphs to filter out redundancy, followed by a text generator to build summaries. Extractive methods are used in hybrid approaches to obtain an output summary that is fed into a text generator to build non-key words and phrases to improve sentence coherence and readability.

SUMMARIST [20] is a key phrase summarizer used to find the boundaries of extraction using a rank-based abstraction approach. The FEMsum summarizer is used to create summaries using a graph-representation and to identify the relationships between the candidate sentences, as well as a syntactic and semantic representation of the phrases. The data structure required for recognizingtopics in document sets and creating various forms of summaries is built by using a fuzzy co-reference cluster graph technique [21]. The intra- and inter-document co-reference chain families generated by a co-reference method under various (fuzzy) clustering criteria are given as input to this algorithm. In other words, each cluster assigns a topic to each document: some themes appear in all documents (common topic), while others appear in only a subset or a single document (contrastive/distinctive topic). In [22], a set of distance functions for assessing structural similarity between online documents is analyzed. They analysed different Tag Frequency Distribution Analysis (TFDA), parametric functions, and edit distance between documents as three distinct ways of defining similarity. [23] proposed a label discovery technique that uses a hierarchical structure to express the relationship between text data in online documents collected from the web. Their programme correctly classifies web pages by discovering similar labels that describe the same type of content. [24] utilised a model that combined documents from various taxonomies. For the classification challenge, their model used the Nave Bayes algorithm. Content-based classifiers are clearly used by some research tools, such as NewsDude, to select valuable articles and to remove articles that appear to be excessively repetitious of previously read articles. [25] proposed employing a support vector machine (SVM) classifier to identify web pages based on both text and context features. They tested their online classification methods using the WebKB dataset, and the results demonstrate that using context features, particularly hyperlinks, can greatly enhance classification performance.

Conventional statistical methods have been included in many models. The main drawback of conventional statistical methods is the rigidity of dynamic situations and therefore the difficulty of optimal modelling. Most of the traditional ensemble classification models are processed with limited feature space and small data size. As the size of the feature space increases, traditional ensemble classifiers select a predefined number of features for classification. [26] proposed a novel discretization approach to continuous attributes for decision tree learning. The main issue with traditional decision tree models is that each attribute is assumed to be either nominal or categorical. To overcome this issue, a dynamic discretization model on the continuous label is applied to each attribute during the tree construction process. Traditional decision tree models such as CART and C4.5 use discretization methods in the preprocessing phase along with noise removal methods. But, the main limitation of this model is that the data should be of a continuous type and it doesn't support mixed types.

Feature selection is a process that selects an optimal feature sub-set based on a particular requirement. The measuring feature subsets are specified in the criterion. The criterion will be selected according to the purposes for which the feature is selected. For example, an optimal subset can be a minimum subset. It can provide the best predictive accuracy estimate in a sub-set. In some circumstances [27], a subset with the specified number that meets the criterion can be found in view of the number of features. Rough Sets Attribute Reduction (RSAR) is a filter-based tool for feature reduction used to extract data and maintain information while reducing the amount of knowledge involved. Analysis of Rough Sets is performed on the basis only of the data provided, and no external parameters are required to operate [28]. This makes use of the data granularity structure. It does, however, continue to assume the model that there is some information available with every item in the discourse universe that truly and accurately reflects the real world. The ideal criterion for the selection of Rough Sets is to find the shortest or minimum reductions while obtaining high grades for the selected features. The redundancy of a feature or feature subset is determined. A feature is declared relevant if the decision feature is predictive, otherwise it is irrelevant. A Principal Component Analysis (PCA) approach to a reduction in dimension is achieved by building main components that are linear combinations of the original predictor or the explaining variables. The PCA approach is based on the supposition that large variance in characteristics provides useful information, and, in contrast, small variance is considered less useful. Ortholy-linear combinations have been designed to maximize features in the linear combination of explicative variables. There are two basic stages of Fuzzy ELM (F-ELM), [29] called preparation and prediction. P. Verma and H. Om [30] proposed the Correlation-based Feature Selection (CFS) method. The correlation between the attribute and the class is calculated by this approach, with the hypothesis that an optimal collection of features should be strongly correlated with the class but not correlated with other features. This is to ensure that redundancies and feature numbers [31-34] (explaining the pattern with as few features as possible but still maintaining high performance) are reduced. Artificial intelligence is a notion that today has a lot of excitement around it. They trained the decision tree using a rotated feature space. Hence, they proposed the rotation forest algorithm. In this method,[35-37]samples from the main datasets are obtained. These samples form a new subset which is fed into a new feature space.

III. Proposed Model

Initially, document sets are taken as input for text pre-processing. In the preprocessing phase, each document is preprocessing using the Stanford NLP library. This library is

used to perform various operations such as document tokenization, stemming and stop word removal on different domain fields. After performing the data pre-processing operations on the large document sets, word embedding model is used to optimize the document to word vectors. In this work, a hybrid multi-domain word embedding model is proposed in order to optimize the word embedding key words on large document-sets.

Proposed multi-domain glove optimization model is designed to find the main and its contextual key features on large document sets. Multi-document contextual features are extracted using the main words of the glove model. A boosting contextual similarity is computed based of the main words, contextual words, string hash similarity and multi-document contextual similarity features to filter essential top k voted features in the document sets. In the next step, a multi-document clustering approach is developed on the filtered top k-contextual voted features for the multi-document summarization process. In the multi-document clustering process, an efficient KNN distance measure is used to compute the nearest clusters by using the structural similarity between the main and contextual scores. Each document and its key features are labelled with the cluster class for the multi-document summarization process. In the proposed multi-document summarization, a hybrid Bayesian probability based classification approach is developed to find the multi-document summarization process as shown in Fig.1.

A. Multi-Document Glove Optimization Word Embedding Model

In the multi-document glove optimization model, each pre-processed document is given as input to compute word co-occurrence matric. Let X_{ij} represents the word occurrence matrix in order to compute main word and contextual word on large document set. W_i and W_j represent main and contextual word vectors of W_c.

b_i, b_j: $main$ *word bias vector and context word bias vector*
$\theta = Min\{b_i, b_j\}.D((w_i.b_i),(w_j,b_j))$

1) Defining multi-document summarization cost function and its constraints using the word, main vectors and its biases:

$$C = CostFunction = tan^{-1}(\eta) * b_i w_i^T w_j + b_j w_i^T w_j + exp(\eta) * max\{b_m, b_c\} - \frac{cos(X_{ij}}{\sqrt[3]{||w_i|| * ||w_j||}}$$

$multi-document$ weights are defined as

$$\eta = multi_{weight} = f(X_{ij}) = \begin{cases} (\frac{\sqrt{X_{ij}}}{X_{max}})^{\alpha} & \text{if } X_{ij} < X_{MAX} \\ \sqrt{X_{ij}}^{\alpha} & \text{otherwise} \end{cases}$$

$where\ \alpha$ is scaling factor.

2) Define a multi-document cost function.

$$J = \sum_{i=1}^{V}\sum_{j=1}^{V} \eta.(tan^{-1}(\theta) * b_i w_i^T w_j + b_j w_i^T w_j + exp(\theta) * max\{b_m, b_c\} - (cos(X_{ij})/\sqrt[3]{||w_i|| * ||w_j||})^2$$

3) The Proposed Multi-document word embedding model is optimized by using the partial derivative w.r.t main words and contextual words as shown below.

$$\frac{\partial J}{\partial w_i} = b_i w_j C = b_i w_j(\tan^{-1}(\theta) * b_i w_i^T w_j + b_j w_i^T w_j + \exp(\theta) * \max(b_m, b_c) - (\cos X_{ij})/\sqrt[3]{||w_i|| * ||w_j||}$$

$$\frac{\partial J}{\partial w_j} = b_i w_i C = b_i w_i(\tan^{-1}(\theta) * b_i w_i^T w_j + b_j w_i^T w_j + \exp(\theta) * \max(b_m, b_c) - (\cos X_{ij})/\sqrt[3]{||w_i|| * ||w_j||}$$

update w_i and w_j using learning theta.

In the above optimized multi-domain glove optimization model, the cost function and its constraints are improved in order to find the essential key contextual features among the multiple domain document sets. Here, the multiweight factor is used to find the weighted document features among the main and contextual feature vectors. Finally, the multi-document cost function is based on multi-weights, main and contextual feature vectors on large contextual co-occurrence matrix.

B. Boosting Voting based Word Embedding Contextual Similarity

In this phase, a voted boosting method is used to compute the best similarity measure based on the multi-document glove main and contextual key vectors. In this phase, hash based similarity, string similarity and proposed multi-document main and contextual similarity measure are used to choose the majority voted similarity on the glove main and contextual feature vectors. The proposed main and contextual similarity measure is computed by using the following formula.

$let\ \omega(i)$ be the multi-document main word vector features ,

$\omega(j) represents$ the multi-document contextual word vector features.

$$\chi = \frac{cos(\omega(i), \omega(j)) \times \sum_{i=1}^{k}\left(\omega(i) - \overline{\omega(i)}\right)}{\sqrt{\sum_{i=1}^{k}\left|\omega(i) - \overline{\omega(i)}\right| \times \sum_{j=1}^{k}\left|\omega(j) - \overline{\omega(j)}\right|}}$$

$Multi-document$ main word similarity
$= tf(i) * log(\chi / max\{|\omega(i)|, |\omega(j)|\})$

$Multi-document$ contextual word similarity
$= log(1 + \chi / min\{|\omega(i)|, |\omega(j)|\})$

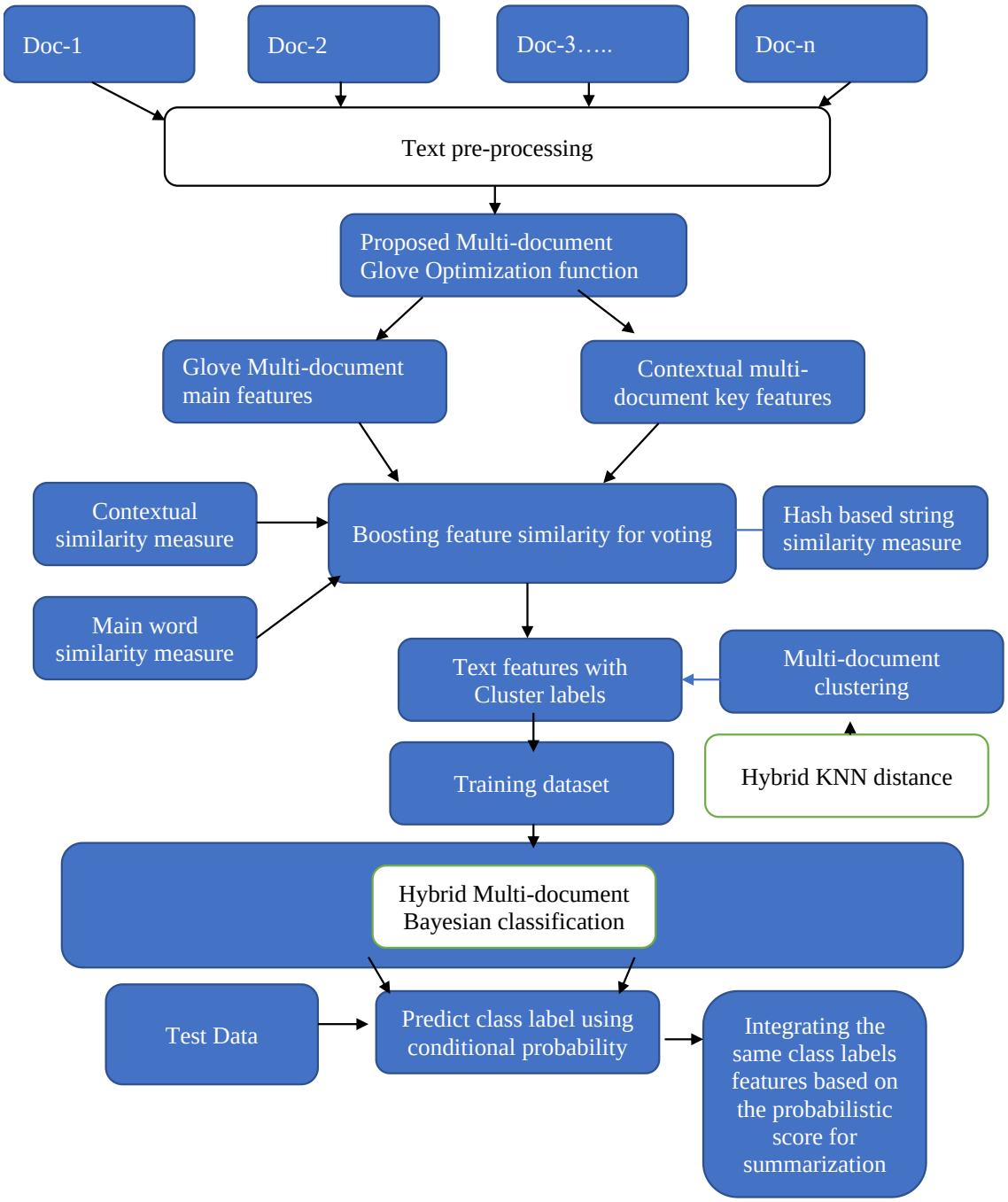


Fig. 1. Proposed Multi-Domain Multi-Document Summarization Framework.

C. Multi-Document Clustering based on KNN Approach

In this phase, a hybrid multi-document clustering based KNN approach is developed on the main and contextual key similarity features. This approach is used to group the multi-documents based on the domain main and contextual similarity vectors. Let k defines the user defined number of k-nearest objects for grouping.

let MD_t represents the multi-document sets.

Output: Clustered k documents with cluster class label

Procedure:

1) Read input data MD_t

2) Initialize k clusters for KNN and perform traditional k-means document clustering algorithm.

3) In the proposed document clustering approach, instead of using the conventional distance measures, a hybrid weighted distance measure is proposed between the main and contextual word vectors.

4) The weighted multi-document pair distance between the main and contextual word vectors is given as

$$\psi_M(TF_{t,d}) = P(w_m / D_i).\frac{tf_{t,d} \times \log \frac{\sqrt{\chi}}{|w_m|}}{[(\sum_{t=1}^{n}(1+\log(\frac{|D|}{\sqrt{T_m}})]}$$

$$\psi_C(TF_{t,d}) = P(w_c / D_i).\frac{tf_{t,d} \times \log \frac{\sqrt{\chi}}{|w_c|}}{[(\sum_{t=1}^{n}(1+\log(\frac{|D|}{\sqrt{T_c}})]}$$

Where

$$\text{where } tf_{t,d}, \eta > 0$$

and

$$T_c = \sum_{d=1}^{D}\sum_{t} tf_{t,d}, T_m = \sum_{d=1}^{D}\sum_{t} tf_{t,d}$$

$$\text{Weightedpairscore(WPS)} = \sum |\psi_M(TF_{t,d}) - \psi_C(TF_{t,d})|$$

Finally, the contextual similarity between the main word vectors and contextual word vectors are clustered using the following similarity score measure as

$$S(MW(d_i), CW(d_j)) = \psi_M(TF_{t,d}).\frac{\sum_{k=1}^{n} t_{ik} \times t_{jk}}{\sqrt{\sum_{k=1}^{n} t_{ik}^2}\sqrt{\sum_{k=1}^{n} t_{jk}^2}} + \psi_C(TF_{t,d}).\frac{\sum_{k=1}^{n} t_{ik} \times t_{jk}}{\sqrt{\sum_{k=1}^{n} t_{ik}^2}\sqrt{\sum_{k=1}^{n} t_{jk}^2}}$$

5) The kscore is used to find the document classification score in each domain filed for the class label prediction on the new test data. The kscore measure is computed using the following formula.

$$KScore(D_t, C_k) = \sum_{d \in DK} S(MW(d_i), CW(d_j)) \times P(D_i, C_j)$$

$$P(D_i, C_j) = \begin{cases} 1 & D_i \in Cj \\ 0 & D_i \notin C_j \end{cases}$$

D. Multi-Document Conditional Bayesian Estimation based Classification

In the multi-document summarization phase, the clustered training data which is generated in the previous section are taken as input to the multi-document base multi-domain classification process. Proposed Bayesian classification model is used to classify the key phrases for the multi-document summarization process. In this phase, two optimizations are performed on the traditional Bayesian text classification model. In the first optimization, a hybrid prior multi-document probability is developed to predict the multi-domain phase on the large textual document sets. In the second optimization, a hybrid posterior probability is proposed on the main and contextual word vectors in each class category. The main steps used in the proposed multi-document summarization are

1) Read contextual and main words clustered labelled document sets as input.

2) Compute prior multi-document classification probability as:

$$Pr(MW(d_i), CW(d_j)) = Multi - Doc((MW(d_i), CW(d_j), C(k)) = P(MW(d_i)/C(k)) * max\{P(CW \cap MW)/C(k)\}/|MW(d_i) + CW(d_j)|$$

3) Predict the posterior multi-document estimation using the maximization of the class labels as:

$$ClassPredict(MW(d_i), CW(d_j)) = argmax\{P(C(k)) * \{\prod P(CW \cup MW)/C(k)\}/||D| - (MW(d_i) + CW(d_j))|$$

4) To each document in the training documents sets,Merge the phrase with high posterior probability and contextual-main word similarity scores for summarization process.

IV. Experimental Results

The performance has been evaluated using the multi-document summarization datasets provided by Document Understanding Conferences (DUC) 2002, Document Understanding Conferences (DUC) 2004[38], multi-news [39], multi-biomedical datasets [40]. It is an open benchmark from the National Institute of Standards and Technology (NIST) for the evaluation of generic automatic summarization. The experiments have been carried out in amazon AWS server with 96 GB RAM.

In the experimental study, word embedding features, classification metrics and multi-document summarization rouge metrics are used to evaluate the performance of the proposed model to the conventional models.

Table 1, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for contextual similarity computation on DUC 2002 dataset. As represented in the table, the proposed approach has improved evaluation than the previous models in terms of contextual similarity between the main and contextual word vectors.

Figure 2, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for contextual similarity computation on DUC 2004 dataset. As represented in the table, the proposed approach has improved evaluation than the previous models in terms of contextual similarity between the main and contextual word vectors.

TABLE I. COMPARATIVE ANALYSIS OF PROPOSED MODEL TO CONVENTIONAL MODELS FOR OVERALL CONTEXTUAL SIMILARITY MEASURE (DUC:2002)

TestDoc	MI	Entropy	TextRank	Glove	Proposed Glove
TestCat1-1	0.86	0.89	0.92	0.93	0.96
TestCat1-2	0.88	0.9	0.91	0.93	0.95
TestCat1-3	0.87	0.89	0.92	0.94	0.95
TestCat1-4	0.86	0.89	0.9	0.93	0.96
TestCat1-5	0.87	0.89	0.92	0.93	0.95
TestCat1-6	0.87	0.89	0.92	0.94	0.95
TestCat1-7	0.88	0.9	0.91	0.92	0.95
TestCat1-8	0.86	0.91	0.91	0.93	0.95
TestCat1-9	0.89	0.89	0.92	0.93	0.95
TestCat1-10	0.85	0.88	0.92	0.93	0.95
TestCat1-11	0.86	0.88	0.91	0.92	0.95
TestCat1-12	0.86	0.9	0.91	0.93	0.95
TestCat1-13	0.85	0.88	0.91	0.92	0.97
TestCat1-14	0.86	0.91	0.92	0.93	0.97
TestCat1-15	0.85	0.9	0.92	0.93	0.96
TestCat1-16	0.88	0.89	0.91	0.94	0.95
TestCat1-17	0.86	0.88	0.92	0.94	0.96
TestCat1-18	0.85	0.88	0.91	0.94	0.96
TestCat1-19	0.88	0.89	0.9	0.94	0.96
TestCat1-20	0.87	0.89	0.91	0.93	0.95

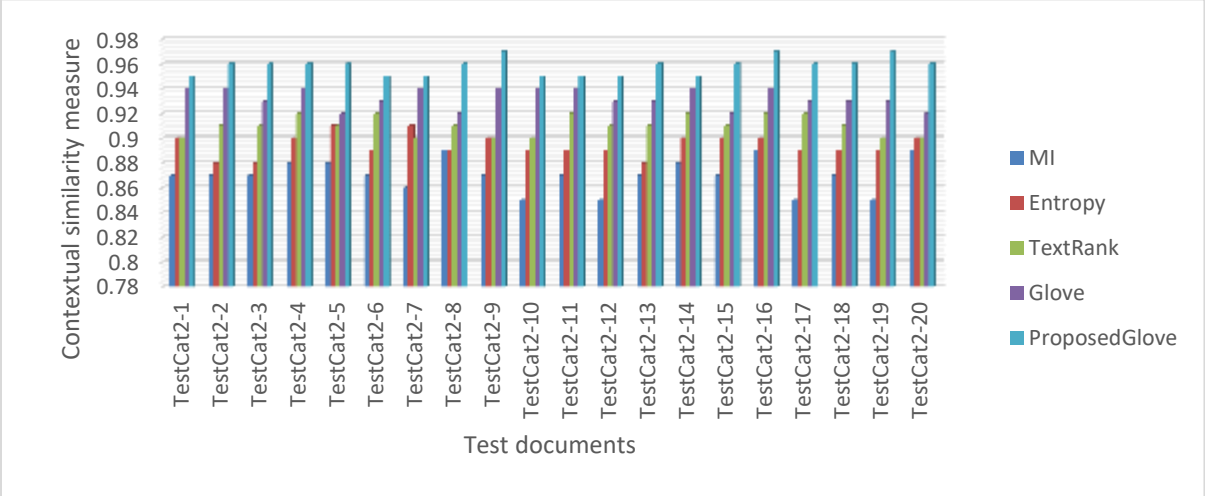


Fig. 2. Comparative Analysis of Proposed Model to Conventional Models for Overall Contextual Similarity Measure (DUC:2004).

Table 2, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for contextual similarity computation on multi-news dataset. As represented in the table, the proposed approach has improved evaluation than the previous models in terms of contextual similarity between the main and contextual word vectors.

Table 3, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for contextual similarity computation on multi-biomedical dataset. As represented in the table, the proposed approach has improved evaluation than the previous models in terms of contextual similarity between the main and contextual word vectors.

Figure 3, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for filtering optimal key features count for document clustering process on DUC 2002 dataset. As represented in the table, the proposed approach has improved evaluation than the previous models in terms of contextual keywords filtering among the large number of candidate key features space.

TABLE II. COMPARATIVE ANALYSIS OF PROPOSED MODEL TO CONVENTIONAL MODELS FOR OVERALL CONTEXTUAL SIMILARITY MEASURE (MULTI-NEWS)

TestDoc	MI	Entropy	TextRank	Glove	ProposedGlove
TestCat3-1	0.86	0.88	0.91	0.92	0.95
TestCat3-2	0.88	0.91	0.92	0.94	0.95
TestCat3-3	0.87	0.89	0.9	0.94	0.95
TestCat3-4	0.86	0.88	0.91	0.93	0.95
TestCat3-5	0.85	0.89	0.9	0.93	0.96
TestCat3-6	0.88	0.88	0.92	0.94	0.95
TestCat3-7	0.89	0.9	0.92	0.92	0.95
TestCat3-8	0.88	0.88	0.92	0.94	0.96
TestCat3-9	0.89	0.9	0.92	0.93	0.95
TestCat3-10	0.87	0.9	0.9	0.93	0.95
TestCat3-11	0.88	0.89	0.91	0.94	0.96
TestCat3-12	0.87	0.9	0.92	0.93	0.95
TestCat3-13	0.87	0.9	0.9	0.94	0.95
TestCat3-14	0.88	0.91	0.91	0.92	0.96
TestCat3-15	0.85	0.89	0.9	0.94	0.97
TestCat3-16	0.87	0.9	0.9	0.94	0.95
TestCat3-17	0.89	0.89	0.9	0.94	0.96
TestCat3-18	0.88	0.89	0.91	0.94	0.95
TestCat3-19	0.87	0.9	0.91	0.93	0.97
TestCat3-20	0.88	0.88	0.92	0.94	0.95

TABLE III. COMPARATIVE ANALYSIS OF PROPOSED MODEL TO CONVENTIONAL MODELS FOR OVERALL CONTEXTUAL SIMILARITY MEASURE (BIOMEDICAL DOCS)

TestDoc	MI	Entropy	TextRank	Glove	ProposedGlove
TestCat4-1	0.87	0.89	0.91	0.92	0.95
TestCat4-2	0.86	0.88	0.91	0.94	0.96
TestCat4-3	0.86	0.89	0.92	0.92	0.95
TestCat4-4	0.85	0.89	0.92	0.93	0.95
TestCat4-5	0.85	0.9	0.92	0.93	0.95
TestCat4-6	0.89	0.88	0.92	0.92	0.95
TestCat4-7	0.88	0.9	0.92	0.93	0.96
TestCat4-8	0.85	0.9	0.91	0.94	0.95
TestCat4-9	0.89	0.89	0.91	0.92	0.96
TestCat4-10	0.85	0.9	0.92	0.94	0.95
TestCat4-11	0.88	0.91	0.92	0.94	0.96
TestCat4-12	0.88	0.88	0.92	0.93	0.95
TestCat4-13	0.85	0.89	0.91	0.93	0.97
TestCat4-14	0.89	0.89	0.92	0.92	0.96
TestCat4-15	0.86	0.9	0.9	0.93	0.95
TestCat4-16	0.89	0.89	0.91	0.92	0.95
TestCat4-17	0.89	0.89	0.92	0.94	0.95
TestCat4-18	0.86	0.89	0.9	0.93	0.97
TestCat4-19	0.85	0.88	0.91	0.94	0.96
TestCat4-20	0.88	0.89	0.92	0.93	0.95

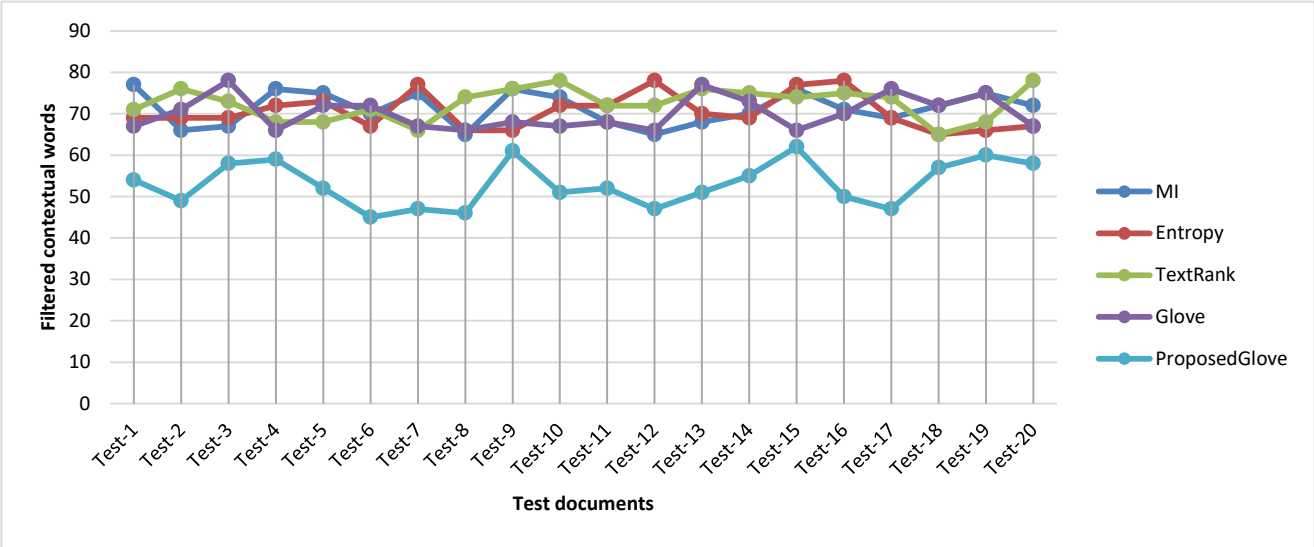


Fig. 3. Comparative Analysis of Proposed Model to Conventional Models for Overall Contextual Keywords Filtering (DUC 2002).

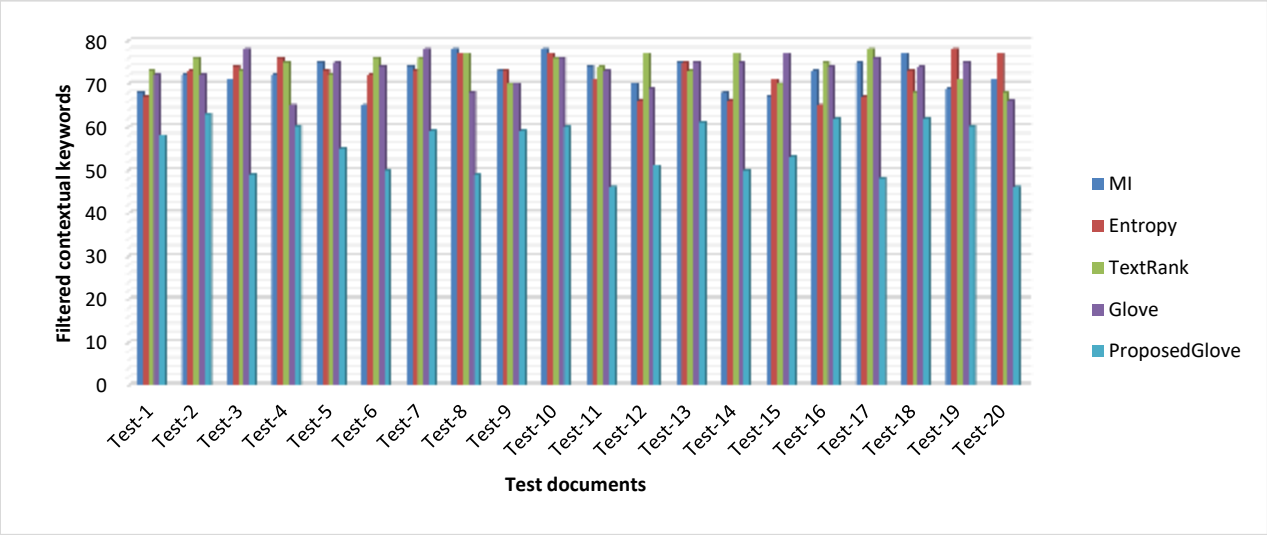


Fig. 4. Comparative Analysis of Proposed Model to Conventional Models for Overall Contextual Keywords Filtering (DUC 2004).

Figure 4, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for filtering optimal key features count for document clustering process on DUC 2004 dataset. From the figure, it is observed that the proposed approach has improved evaluation than the previous models in terms of contextual keywords filtering among the large number of candidate key features space.

Table 4, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for filtering optimal key features count for document clustering process on multi-news dataset. As represented in the table, the proposed approach has improved evaluation than the previous models in terms of contextual keywords filtering among the large number of candidate key features space.

Table 5, illustrates the performance evaluation of the proposed multi-domain glove optimization model to the conventional approaches for filtering optimal key features count for document clustering process on biomedical document sets. As represented in the table, the proposed approach has improved evaluation than the previous models in terms of contextual keywords filtering among the large number of candidate key features space.

Table 6, represents the performance evaluation of the proposed multi-document based Bayesian summarization model to the conventional models for classification precision on various domain databases. From the table, it is noted that the proposed multi-document based Bayesian summarization approach has improved precision than the previous approaches on different domain document sets.

Figure 5, represents the performance evaluation of the proposed multi-document based Bayesian summarization model to the conventional models for classification accuracy on various domain databases. From the figure, it is noted that the proposed multi-document based Bayesian summarization approach has improved accuracy than the previous approaches on different domain document sets.

Figure 6, represents the performance evaluation of the proposed multi-document based Bayesian summarization model to the conventional models for classification recall on various domain databases. From the figure, it is noted that the proposed multi-document based Bayesian summarization approach has improved recall than the previous approaches on different domain document sets.

TABLE IV. COMPARATIVE ANALYSIS OF PROPOSED MODEL TO CONVENTIONAL MODELS FOR OVERALL CONTEXTUAL KEYWORDS FILTERING (MULTI-NEWS)

TestDoc	MI	Entropy	TextRank	Glove	ProposedGlove
Test-1	73	73	70	77	61
Test-2	67	70	68	69	53
Test-3	65	66	76	68	57
Test-4	77	76	71	68	45
Test-5	74	67	70	77	54
Test-6	78	69	73	76	59
Test-7	73	73	68	71	57
Test-8	74	73	68	73	50
Test-9	67	71	68	76	62
Test-10	67	65	74	71	53
Test-11	72	71	74	70	51
Test-12	77	77	73	70	49
Test-13	68	65	66	65	54
Test-14	69	71	76	67	53
Test-15	70	72	71	76	57
Test-16	66	77	72	74	57
Test-17	71	75	71	77	57
Test-18	74	69	70	69	50
Test-19	67	75	77	70	62
Test-20	77	65	74	65	55

TABLE V. COMPARATIVE ANALYSIS OF PROPOSED MODEL TO CONVENTIONAL MODELS FOR OVERALL CONTEXTUAL KEYWORDS FILTERING (BIOMEDICAL DOCS)

TestDoc	MI	Entropy	TextRank	Glove	ProposedGlove
Test-1	72	75	66	76	60
Test-2	69	78	71	77	57
Test-3	72	69	68	69	55
Test-4	77	67	76	69	57
Test-5	76	77	66	72	55
Test-6	73	73	72	75	52
Test-7	67	66	72	73	56
Test-8	70	73	67	71	54
Test-9	68	74	71	70	61
Test-10	72	65	77	73	56
Test-11	69	69	69	65	48
Test-12	72	67	75	66	62
Test-13	71	75	72	73	52
Test-14	66	73	65	66	54
Test-15	75	68	73	69	52
Test-16	68	73	72	75	46
Test-17	71	71	67	77	48
Test-18	66	75	68	76	50
Test-19	69	67	72	70	52
Test-20	72	73	67	72	53

TABLE VI. COMPARATIVE EVALUATION OF PROPOSED MULTI-DOCUMENT BASED BAYESIAN SUMMARIZATION MODEL TO THE CONVENTIONAL MODELS FOR CLASSIFICATION PRECISION ON VARIOUS DOMAIN DATABASES

MultiDoc Test	CSTSumm	GistSumm	MultiLayer	ProposedMDBayesianSumm
DUC2002	0.86	0.89	0.9	0.97
DUC2004	0.87	0.89	0.91	0.96
Multi-News	0.86	0.9	0.9	0.95
Multi-Biomedical	0.85	0.89	0.9	0.95

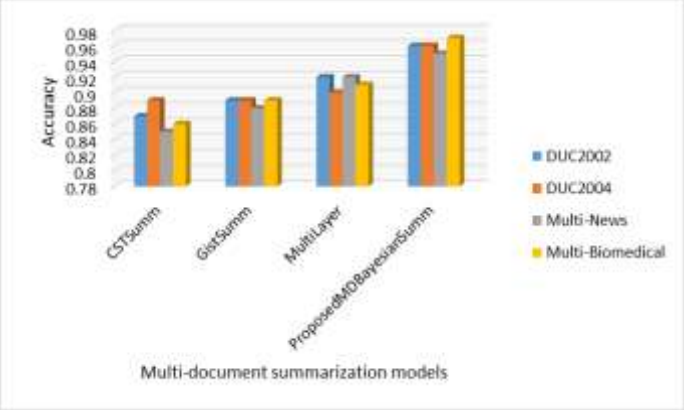


Fig. 5. Comparative Evaluation of Proposed Multi-Document based Bayesian Summarization Model to the Conventional Models for Classification Accuracy on Various Domain Databases.

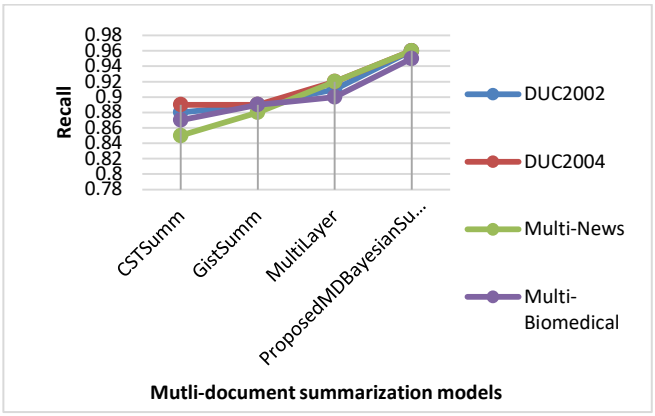


Fig. 6. Comparative Evaluation of Proposed Multi-Document based Bayesian Summarization Model to the Conventional Models for Classification Recall on Various Domain Databases.

For experimental evaluation, we use ROUGE (Recall-Oriented Understudy for Gisting Evaluation) in order to find the performance of the proposed multi-doc summarization process on various traditional models.

Table 7, represents the performance evaluation of the proposed multi-document based Bayesian summarization model to the conventional models for average rouge metrics on DUC 2002 domain database. From the table, it is noted that the proposed multi-document based Bayesian summarization approach has improved average rouge metrics than the previous approaches on different DUC 2002 document sets.Table 8, represents the performance evaluation of theproposed multi-document based Bayesian summarization model to the conventional models for average rouge metrics on multi-news domain database. From the table, it is noted that the proposed multi-document based Bayesian summarization approach has improved average rouge metrics than the previous approaches on different domain multi-news data.

TABLE VII. COMPARATIVE EVALUATION OF PROPOSED MULTI-DOCUMENT BASED BAYESIAN SUMMARIZATION MODEL TO THE CONVENTIONAL MODELS FOR AVERAGE ROUGE METRICS ON VARIOUS DUC 2002 DATABASE

Avg Rouge	Gensim	OPINOSIS	PyTextRank	Proposed BayesianSumm
Recall	0.05	0.065	0.087	0.17
Precision	0.04	0.075	0.075	0.14
F-measure	0.054	0.065	0.0734	0.12

TABLE VIII. COMPARATIVE EVALUATION OF PROPOSED MULTI-DOCUMENT BASED BAYESIAN SUMMARIZATION MODEL TO THE CONVENTIONAL MODELS FOR AVERAGE ROUGE METRICS ON MULTI-NEWS DOMAIN DATABASE

Avg Rouge	Gensim	OPINOSIS	PyTextRank	ProposedBayesianSumm
Recall	0.034	0.046	0.085	0.14
Precision	0.023	0.048	0.078	0.094
F-measure	0.036	0.05	0.09	0.12

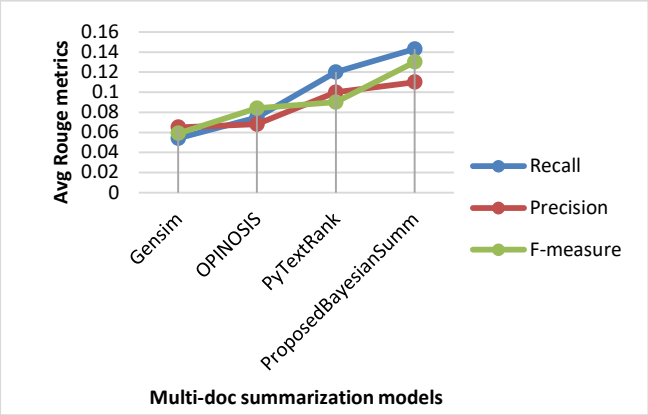


Fig. 7. Comparative Evaluation of Proposed Multi-Document based Bayesian Summarization Model to the Conventional Models for Average Rouge Metrics on Multi-Biomedical Domain Database.

Figure 7, represents the performance evaluation of the proposed multi-document based Bayesian summarization model to the conventional models for average rouge metrics on multi-biomedical domain database. From the table, it is noted that the proposed multi-document based Bayesian summarization approach has improved average rouge metrics than the previous approaches on biomedical document sets.

TABLE IX. COMPARATIVE EVALUATION OF PROPOSED MULTI-DOCUMENT BASED BAYESIAN SUMMARIZATION MODEL TO THE CONVENTIONAL MODELS FOR AVERAGE ROUGE METRICS ON VARIOUS DUC 2004 DATABASE

Avg Rouge	Gensim	OPINOSIS	PyTextRank	ProposedBayesianSumm
Recall	0.035	0.084	0.095	0.16
Precision	0.043	0.078	0.11	0.154
F-measure	0.049	0.069	0.12	0.158

Table 9, represents the performance evaluation of the proposed multi-document based Bayesian summarization model to the conventional models for average rouge metrics on DUC 2004 domain database. From the table, it is noted that the proposed multi-document based Bayesian summarization approach has improved average rouge metrics than the previous approaches on different DUC 2004 document sets.

A. Results Interpretation

In this work, different multi-document features and its correlated main and contextual words are used to analyze the multiple documents for summarization. From the experimental results it is noted that the average accuracy, recall and precision of the proposed multi-document summarization is better than the conventional models with nearly 1% improvement. Also, the contextual features of the proposed glove model has better optimization for the word to vector generation process.

V. CONCLUSION

Multi-document summarization plays a vital role in the multi-domain document sets due to variation in the feature space and inter and intra document cluster variations. Since, most of the conventional multi-document summarization models have large number of candidate feature sets for document clustering and classification process. In this work, a hybrid multi-document based glove optimization model is proposed in order to filter the key features on multi-domain document sets. Also, a hybrid document clustering and multi-document Bayesian classification model for multi-domain document summarization process is proposed on large document sets. Experimental evaluation represent the performance of the proposed Bayesian multi-document summarization approach has improved rouge evaluation metrics than the previous models with nearly 2-3% improvement on large multi-domain document sets. In the future scope, this work can be extended to improve the multi-level based dynamic multi-domain feature extraction and summarization process using the parallel processing framework.

REFERENCES

[1] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, "A decomposition-based multi-objective optimization approach for extractive multi-document text summarization," Applied Soft Computing, vol. 91, p. 106231, Jun. 2020, doi: 10.1016/j.asoc.2020.106231.

[2] A. Abdi, S. Hasan, S. M. Shamsuddin, N. Idris, and J. Piran, "A hybrid deep learning architecture for opinion-oriented multi-document summarization based on multi-feature fusion," Knowledge-Based Systems, vol. 213, p. 106658, Feb. 2021, doi: 10.1016/j.knosys.2020.106658.

[3] R. Ferreira et al., "A multi-document summarization system based on statistics and linguistic treatment," Expert Systems with Applications, vol. 41, no. 13, pp. 5780–5787, Oct. 2014, doi: 10.1016/j.eswa.2014.03.023.

[4] G. Yang, D. Wen, Kinshuk, N.-S. Chen, and E. Sutinen, "A novel contextual topic model for multi-document summarization," Expert Systems with Applications, vol. 42, no. 3, pp. 1340–1352, Feb. 2015, doi: 10.1016/j.eswa.2014.09.015.

[5] M. Mojrian and S. A. Mirroshandel, "A novel extractive multi-document text summarization system using quantum-inspired genetic algorithm: MTSQIGA," Expert Systems with Applications, vol. 171, p. 114555, Jun. 2021, doi: 10.1016/j.eswa.2020.114555.

[6] D. Bollegala, N. Okazaki, and M. Ishizuka, "A preference learning approach to sentence ordering for multi-document summarization," Information Sciences, vol. 217, pp. 78–95, Dec. 2012, doi: 10.1016/j.ins.2012.06.015.

[7] M. Bidoki, M. R. Moosavi, and M. Fakhrahmad, "A semantic approach to extractive multi-document summarization: Applying sentence expansion for tuning of conceptual densities," Information Processing & Management, vol. 57, no. 6, p. 102341, Nov. 2020, doi: 10.1016/j.ipm.2020.102341.

[8] A. Qaroush, I. Abu Farha, W. Ghanem, M. Washaha, and E. Maali, "An efficient single document Arabic text summarization using a combination of statistical and semantic features," Journal of King Saud University - Computer and Information Sciences, Mar. 2019, doi: 10.1016/j.jksuci.2019.03.010.

[9] R. Rautray and R. C. Balabantaray, "An evolutionary framework for multi document summarization using Cuckoo search approach: MDSCSA," Applied Computing and Informatics, vol. 14, no. 2, pp. 134–144, Jul. 2018, doi: 10.1016/j.aci.2017.05.003.

[10] S. Lamsiyah, A. El Mahdaouy, B. Espinasse, and S. El Alaoui Ouatik, "An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings," Expert Systems with Applications, vol. 167, p. 114152, Apr. 2021, doi: 10.1016/j.eswa.2020.114152.

[11] H. Oliveira et al., "Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization," Expert Systems with Applications, vol. 65, pp. 68–86, Dec. 2016, doi: 10.1016/j.eswa.2016.08.030.

[12] R. Rautray and R. C. Balabantaray, "Bio-inspired approaches for extractive document summarization: A comparative study," Karbala International Journal of Modern Science, vol. 3, no. 3, pp. 119–130, Jul. 2017, doi: 10.1016/j.kijoms.2017.06.001.

[13] R. Rautray and R. C. Balabantaray, "Cat swarm optimization based evolutionary framework for multi document summarization," Physica A: Statistical Mechanics and its Applications, vol. 477, pp. 174–186, Jul. 2017, doi: 10.1016/j.physa.2017.02.056.

[14] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, "Comparison of automatic methods for reducing the Pareto front to a single solution applied to multi-document text summarization," Knowledge-Based Systems, vol. 174, pp. 123–136, Jun. 2019, doi: 10.1016/j.knosys.2019.03.002.

[15] E. Linhares Pontes, S. Huet, J.-M. Torres-Moreno, and A. C. Linhares, "Compressive approaches for cross-language multi-document summarization," Data & Knowledge Engineering, vol. 125, p. 101763, Jan. 2020, doi: 10.1016/j.datak.2019.101763.

[16] K. Yao, L. Zhang, T. Luo, and Y. Wu, "Deep reinforcement learning for extractive document summarization," Neurocomputing, vol. 284, pp. 52–62, Apr. 2018, doi: 10.1016/j.neucom.2018.01.020.

[17] A. Ghadimi and H. Beigy, "Deep submodular network: An application to multi-document summarization," Expert Systems with Applications, vol. 152, p. 113392, Aug. 2020, doi: 10.1016/j.eswa.2020.113392.

[18] Y. Wu, Y. Li, and Y. Xu, "Dual pattern-enhanced representations model for query-focused multi-document summarisation," Knowledge-Based Systems, vol. 163, pp. 736–748, Jan. 2019, doi: 10.1016/j.knosys.2018.09.035.

[19] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, "Experimental analysis of multiple criteria for extractive multi-document text summarization," Expert Systems with Applications, vol. 140, p. 112904, Feb. 2020, doi: 10.1016/j.eswa.2019.112904.

[20] L. Marujo et al., "Exploring events and distributed representations of text in multi-document summarization," Knowledge-Based Systems, vol. 94, pp. 33–42, Feb. 2016, doi: 10.1016/j.knosys.2015.11.005.

[21] M. Rajangam and C. Annamalai, "Extractive document summarization using an adaptive, knowledge based cognitive model," Cognitive Systems Research, vol. 56, pp. 56–71, Aug. 2019, doi: 10.1016/j.cogsys.2018.11.005.

[22] J. V. Tohalino and D. R. Amancio, "Extractive multi-document summarization using multilayer networks," Physica A: Statistical Mechanics and its Applications, vol. 503, pp. 526–539, Aug. 2018, doi: 10.1016/j.physa.2018.03.013.

[23] A. John, P. S. Premjith, and M. Wilscy, "Extractive multi-document summarization using population-based multicriteria optimization," Expert Systems with Applications, vol. 86, pp. 385–397, Nov. 2017, doi: 10.1016/j.eswa.2017.05.075.

[24] T. Uçkan and A. Karcı, "Extractive multi-document text summarization based on graph independent sets," Egyptian Informatics Journal, vol. 21, no. 3, pp. 145–157, Sep. 2020, doi: 10.1016/j.eij.2019.12.002.

[25] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, "Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach," Knowledge-Based Systems, vol. 159, pp. 1–8, Nov. 2018, doi: 10.1016/j.knosys.2017.11.029.

[26] J. Chen and H. Zhuge, "Extractive summarization of documents with images based on multi-modal RNN," Future Generation Computer Systems, vol. 99, pp. 186–196, Oct. 2019, doi: 10.1016/j.future.2019.04.045.

[27] J.U. Heu, I. Qasim, and D.-H. Lee, "FoDoSu: Multi-document summarization exploiting semantic analysis based on social Folksonomy," Information Processing & Management, vol. 51, no. 1, pp. 212–225, Jan. 2015, doi: 10.1016/j.ipm.2014.06.003.

[28] D. Patel, S. Shah, and H. Chhinkaniwala, "Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique," Expert Systems with Applications, vol. 134, pp. 167–177, Nov. 2019, doi: 10.1016/j.eswa.2019.05.045.

[29] H. Kwon, B.-H. Go, J. Park, W. Lee, Y. Jeong, and J.-H. Lee, "Gated dynamic convolutions with deep layer fusion for abstractive document summarization," Computer Speech & Language, vol. 66, p. 101159, Mar. 2021, doi: 10.1016/j.csl.2020.101159.

[30] P. Verma and H. Om, "MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization," Expert Systems with Applications, vol. 120, pp. 43–56, Apr. 2019, doi: 10.1016/j.eswa.2018.11.022.

[31] A. Qaroush, I. Abu Farha, W. Ghanem, M. Washaha, and E. Maali, "An efficient single document Arabic text summarization using a combination of statistical and semantic features," Journal of King Saud University - Computer and Information Sciences, vol. 33, no. 6, pp. 677–692, Jul. 2021, doi: 10.1016/j.jksuci.2019.03.010.

[32] Z. Ji, Y. Zhao, Y. Pang, and X. Li, "Cross-modal guidance based auto-encoder for multi-video summarization," Pattern Recognition Letters, vol. 135, pp. 131–137, Jul. 2020, doi: 10.1016/j.patrec.2020.04.011.

[33] M. Rajangam and C. Annamalai, "Extractive document summarization using an adaptive, knowledge based cognitive model," Cognitive Systems Research, vol. 56, pp. 56–71, Aug. 2019, doi: 10.1016/j.cogsys.2018.11.005.

[34] D. Wang, H. Fan, and J. Liu, "Learning with joint cross-document information via multi-task learning for named entity recognition," Information Sciences, vol. 579, pp. 454–467, Nov. 2021, doi: 10.1016/j.ins.2021.08.015.

[35] S. Afantenos, V. Karkaletsis, and P. Stamatopoulos, "Summarization from medical documents: a survey," Artificial Intelligence in Medicine, vol. 33, no. 2, pp. 157–177, Feb. 2005, doi: 10.1016/j.artmed.2004.07.017.

[36] J. Chen and H. Zhuge, "Summarization of scientific documents by detecting common facts in citations," Future Generation Computer Systems, vol. 32, pp. 246–252, Mar. 2014, doi: 10.1016/j.future.2013.07.018.

[37] M. Mohd, R. Jan, and M. Shah, "Text document summarization using word embedding," Expert Systems with Applications, vol. 143, p. 112958, Apr. 2020, doi: 10.1016/j.eswa.2019.112958.

[38] https://duc.nist.gov/data.html.

[39] http://mlg.ucd.ie/datasets/bbc.html.

[40] https://www.ncbi.nlm.nih.gov/research/pubtator-api.

Integration of Value Co-creation into the e-Learning Platform

Eliza Annis Thangaiah, Ruzzakiah Jenal, Jamaiah Yahaya
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

***Abstract*—The e-learning platform is a technology used in most academic institutions. The e-learning platform provides services as an alternative to conventional methods. Previous studies have primarily focused on using and accepting e-learning among consumers and developing tangible attributes on the platform. Platform attributes should be available to engage with all users, leading to innovative ideas and improvement in the value offerings to users if well used. Therefore, this study explores the service science perspective in terms of co-creation for an e-learning platform. The concept of Service-Dominant Logic and value co-creation is adopted to explore and extract the elements and factors that are collectively applied to the model. The concepts illustrate how user value is co-created through the value propositions on the platform and value drivers for the users. The findings help identify the components for value proposition on the platform: enrichment, interaction, personalization, and environment. Meanwhile, the components of value drivers for actors are engagement, resources, experience, and goals. Then, the proposed components are used to develop an e-learning conceptual model. A service-driven model of e-learning will be a significant input to develop an effective platform that provides co-creation opportunities to its users. Future research is to identify the critical features available on the e-learning platform from the users' view.**

Keywords*—*e-Learning; co-creation; S-D Logic; value propositions; value drivers; actors

I. Introduction

Recently, e-learning technology has been in high demand since the Covid pandemic hit the world. It is widespread recognition that E-learning plays a decisive role in supporting and engaging teaching and learning among its users. Students and teachers initially perceived face-to-face learning as the ultimate way of teaching and learning, but they have to find alternatives since the restriction order is applied**.** The objective of e-learning systems is to improve the educational system, enhance students' and teachers' learning, boost the education system, and improve students' performance. Teachers and students, as primary users of the education system, need to adapt to survive in the ever-changing technology environment. They need to be agile to adjust, respond, and be resourceful in working together and find practical learning for the institutions. By collaborating to develop information, users promote to transform information into knowledge and eventually get values.

Since its introduction, e-learning has gone through various adaptation processes to strengthen the implementation process and technology evolving from time to time. Users can access learning materials, search for information, collaboration, and experience irrespective of physical location limitations to deliver knowledge. Besides, technology and computing help educators prepare students for 21st-century life lessons.

Most developers focus on tangible items to develop e-learning platforms or websites, such as technology [1] and internationalization [2], while value to the user and institutions is less considered [3]. Although online platforms are vastly rising, yet insufficient active contributors and continuous input of material or information [4]. A different perspective of the e-learning model from the value view can be considered so that e-learning feature development meets the implied value. It is reported that students stated e-learning did not have the desired features and was also harder to use compared to Facebook [5]. Thus, services are only valuable if used and not based on what is offered, and hence user involvement in system development may accomplish a successful system. The value of the system exists in the stakeholders' cognitive space. Values are unique among individuals but can influence the user's behaviour in using e-learning.

In Service-Dominant Logic (S-D Logic), creating value co-creation (VCC) together is key to generating value by integrating their value with others and developing a new one [6]. With VCC, all users are part of the creators of value when using the system. According to [7], the concept of value co-creation in academic virtual learning environments has been discussed less frequently. To convert the traditional teacher-student models, in which teachers determine the learning resources, they can participate in VCC's educational processes into a flexible structure and an active learning environment. Thus, to find the platform that caters based on user's active participation, value co-creation focusing on user value strategy is proposed. Positive values instilled in the user's mind will guide them towards a proper attitude. The e-learning platform services need to be used to deliver value [8]. A service has to have value to fulfill the user's needs. Further, studies linking education by using the concept of S-D Logic are still poorly implemented [9].

This paper will describe the factors of VCC in value proposition on the platform and value drivers from the users in using e-learning, which would lead to creating values for themselves and would benefit the institutions eventually. As [10] pointed out, service providers do not just make value proposition but also involves the actors who are users or contributors to produce values for themselves. It is needed to get feedback from the user to make them feel close, and there is a need to use the services on the platform.

This paper is structured as follows: Section 2, "Literature Review" outlines the topics on e-learning, service-dominant Logic, and value co-creation. Section 3, "Methodology" describes the methodology employed to develop the conceptual model. Next, Section 4, "Result and Discussion" draws the outcome and discussion about the e-learning conceptual model. Lastly, Section 5, "Conclusion" points toward the overall conclusion and following plot for the research.

II. Literature Review

A. E-learning

E-learning is a web-based technology and application to enable learning and teaching [11] delivered through technology such as the internet, intranets, audio, video and conferencing, virtual classrooms, and digital collaboration. Web-based or mobile-based learning has become common in education. It can take many platforms, from massive open online courses (MOOCs) to a virtual learning environment (VLE) and learning management system (LMS), and many others [12].

In learning, teachers determine the learning resources into a flexible structure and an active learning environment to participate in the educational processes in virtual academic learning environments [7]. Collaborative and virtual learning environments provide a dynamic platform of interaction and conversation for students and teachers. For a successful system, e-learning should provide a platform that enables qualitative two-way communication between students and teachers and amongst students themselves. Thus, teachers' perceptions as providers must change, allowing students to be active contributors and enabling collaboration to occur effortlessly.

B. Service-Dominant Logic (S-D Logic)

The traditional view in marketing was Goods-dominant Logic; consumers are provided with their value [10]. The latter S-D Logic claims that value is always co-created with the partnership with users. [13] added the value-in-use in Goods-dominant Logic lacks the firm-customer interactions. The customer must participate by utilizing a product or service in the value exchange process to create value.

Based on S-D Logic, the main component is 'the customer is always the co-creator of value, which means creating value is an interactional process requiring active participation of both the customer and the supplier. In this study, the researched relationship is between teacher and student, student and student, and teacher and teacher. They are the primary users of the e-learning platform and are called actors. Users are categorized as operant resources in S-D Logic, integrating skills and knowledge into the co-creation process into the e-learning platform's activities. Teachers' roles are not as providers and directors of the whole process, but they also become participants. Students' and teachers' perceptions and involvement as value co-creators can be linked to some foundational premises (FP). It is considered service value co-creators in S-D Logic based on ten FPs [14]. For example, FP1 is "Service is the fundamental basis of exchange," is a relevant premise for e-learning because e-learning provides services to all matters involving skills and knowledge. Also, FP10: "value if always uniquely and phenomenologically determined by the beneficiary" shows each user of e-learning platform are different, and their respective experience will add value to the platform.

C. Value co-creation (VCC)

In the value construct, there is a relationship with the co-creation element. Co-creation is needed, as various platforms provide space for virtual communities, yet many still lack sufficient active contributors and a continuous supply of knowledge content. Based on the value concept in the new marketing logic, value is not created solely by the firm but built with the customers during their usage of products and interactions with different actors. The value is produced in the customer sphere during consumption in response to the service provider's proposition value. Therefore, these value attributes are directly related to customer feelings and attitudes developed towards the service offered. [15] mentions VCC and the tools facilitating VCC activities have increased interest from information systems (IS) scholars and business practitioners. Researchers have been studying co-creation in various studies of information technology, such as in e-learning [16], [17], Internet of Things [18], information systems [15], and social media [19].

Showing co-created material is a real need with a real impact, so the student and teachers can see that working together is worthwhile**.** This active participation means incorporating engagement, experience, resources, and goals into the activities; they guide the co-creation process themselves and take more responsibility for the activities to be co-created with the other participants. Virtual communities' studies identify two co-creation behaviours: searching for information through the community and, secondly, by participating in the community by generating and sharing content with other members [8]. Thus, it is vital to identify the features that would enable users to use e-learning to have closeness and create a sensory and emotional connection while using e-learning. The firm can develop only a value proposition, but the user determines the value of the offering through its usage.

Users are more informed, connected, and empowered than before; due to Internet technologies, they have access to new tools that enable them to co-create with others. The level of user participation in co-creation varies depending on the user's experience, engagement, resources, and goals. Further, it will influence them whether to continue using the platform for good or not.

This study derived VCC for the e-learning platform from interaction, enrichment, environment, and personalization for its primary users: teachers and students. Every user is unique; they tend to have their minds and ideas, contributing to their desires, experience, and future expectations. When various users work together with their different backgrounds and create new knowledge, it would develop users' values. Students have individual opinions about learning; thus, they need a platform that enables them to share and contribute to education by participating, improving their learning, thinking skills, understanding concepts, and creating knowledge [7]. In

detail, the researchers' model illustrates how user value is co-created through the value propositions on the platform and value drivers for the users. Besides, a study by [20] identified the need for co-creation in the e-learning environment.

III. Methodology

The process of developing the conceptual model was organized into three phases: input, activity, and output, as in Table I. The flow of the model development is shown in Fig. 1.

TABLE I. Phases of Developing a Conceptual Model

Input	Activity	Output
• Selection of primary source studies • Identify theory and models	• Inspect and identify the relevant studies and models based on keywords. • Identify suitable constructs and components	• e-learning conceptual model based S-D Logic and VCC

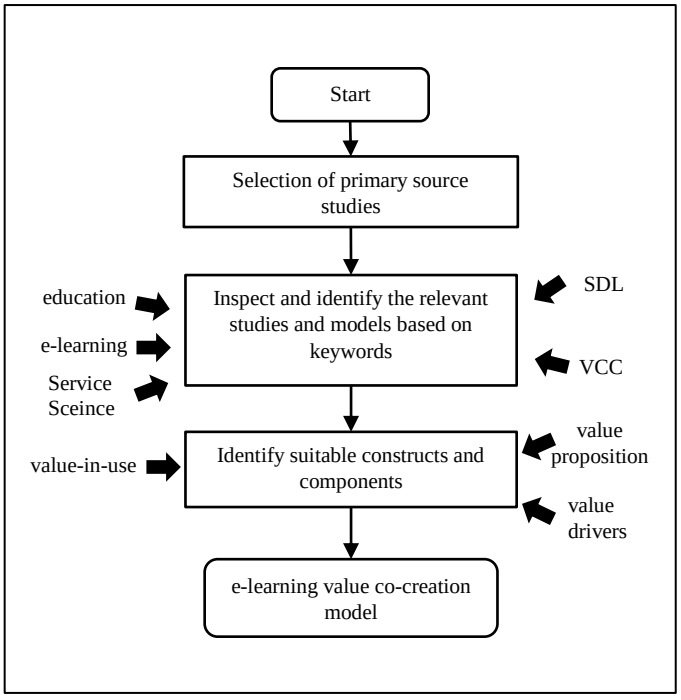


Fig. 1. The Flow of Model Development.

In selecting primary source studies, research databases, available search engines, and reliable websites were identified. This included the articles published by Sage Journals, Web of Science, Science Direct, Springer Link, Taylor and Francis, Emerald and Wiley. Next, academic papers, books, thesis, and related English and Malay language reports were selected. Keywords used were: service science, service-dominant Logic, e-learning, and education. All of the papers were available in electronic format, thus could be easily discovered using the keywords. Content analysis of all the items was handled to identify the elements and factors from the perspective of service science that could contribute to the value proposition on the e-learning platform and could lead value drivers to actors to contribute. Factors are examined in such can be used to develop the e-learning conceptual model embedding the concepts of service-dominant Logic and value co-creation.

IV. Result and Discussion

S-D Logic and VCC have been proposed as the vital factor to develop a successful service or product. Since e-learning is a service platform, it may benefit from involving users' input in the development stage. Users need to play an important role and create value with others. Therefore, the views of students and educators as stakeholders are needed in building an e-learning model. The joint contribution of ideas from students and instructors can build an e-learning model by determining users' values while using e-learning.

Therefore, a new approach needs to be considered in the e-learning platform by studying the value element of co-creation through students' and instructors' value element approaches. The conceptual model's elaboration will be discussed in Sub-section A. Value Proposition and B. Value Drivers.

A. Value Proposition

Platforms play a significant role in involving users, gathering them in the co-creation process, for sharing knowledge related to product or service usage. Platform with virtual environments for co-creation offers new opportunities by providing access to social media, i.e., content and interactions created via highly accessible web-based technologies. A platform can also create content that appeals to other users to use products or services while meaningfully enhancing the co-creation process. The technology on the platform must be taken advantage of to stimulate user's participation.

A value proposition is an opportunity for value-creating benefits delivered by an organization to its customers [21]. Value propositions facilitate the VCC process and do not have value by themselves. Once users accept a value proposition, they are an integrator of the value proposition (offered by the service provider) with their other resources to create value. In this study, four value propositions proposed are enrichment, interaction, personalization, and environment.

1) Enrichment: Enrichment on the e-learning platform refers to the information and features offered to generate knowledge to be more attractive, interactive, and suitable for learning. The content presentation provides easy management to users while using the e-learning platform. Information conveyed to users includes content management, how learning content is delivered, and enrichment of learning through more media towards users who aim to achieve goals [22]. Enrichment of information can shape individuals' development in conducting an independent collection of resources for current results and future planning [23]—mastery of information relating to users and systems, thus publishing value elements. Good learning management can help stimulate information delivery and create learning materials involving all users [24]. Enrichment on the platform can be presented on updated content, user information, an organized file system, achievement space, notification space,

reminders, and security. From the S-D Logic perspective, enrichment on the platform is a service offered involving skills and knowledge and enrichment of attributes/features, also known as operant resources. Information enrichment on platform can be represented through FP1 of S-D Logic, where the services offered on the platform involve skills and knowledge. As well as FP4, which states that operant resources are fundamental to the competitive advantage.

2) Interaction: Interaction through active user communication is a vital process of creating shared values, and the quality of interaction is fundamental to the creation of shared values of users. Online platforms such as e-learning enable interaction and collaboration between actors in co-creation [25]. From the point of view of S-D Logic, the users are co-creator of values and sources of integration, and they can share based on their level of experience, knowledge, and exposure [26]. Nowadays, students can access multiple resources to develop their understanding and further enhance sharing in the learning process [7]. Moreover, students are prone to search for information and no longer depend on just receiving end; they want their voices to be heard. In an e-learning platform, interaction creates an interactive learning environment for developing material content or learning material based on tacit knowledge about the subject matter, allowing VCC communication between teachers and students. Meaningful interaction between teachers and students, students and students, and teachers and teachers significantly strengthens students' sense of membership and influence.

There are various opportunities for users to communicate with available technology, such as comment sections, notification, gaming, frequently asked questions, and others [27]. It could promote using the e-learning platform for a more extended period. Interactions between users in the community are essential for value creation; therefore, it is structured to find opportunities to co-create with users. [8] recommended having maximum interactivity so that users feel crucial, their response is taken into consideration, and they play an essential role within the community. From the S-D Logic perspective, FP6 states that the client is a co-creator of values and through this interaction, VCC can be achieved and supported. Meanwhile, FP9 says that everyone involved in the activity is a source of integration. FP11 states the value of co-creation is the result of coordination between actors and organizational creation.

3) Personalization: Personalization in web-based interactive environments has been considered a key element in enabling effective and efficient user engagement and applied to learning content, learning resources, and development activities [28]. Personalization is a process that allows users to access and store information based on various personal characteristics of the user and provide unique value and benefits to users based on needs and backgrounds. Personalization of e-learning will enable users to experience and, in turn, also contribute to the learning process. The merging of the personalization concept to e-learning technology empowers the learner to have a learning process that can be adapted to their needs [29]. The attribute that can be represented by personalization on platforms such as customize the interface, personalized recommendation, progress bar, online personalized feedback, and a sticky note, to name a few. Therefore, e-learning equipped with personalization must meet a wide range of individual needs and preferences to enable VCC.

From the S-D Logic perspective, service orientation through user personalization achieves its benefits by making the user's own choices. E-Learning allows selecting learning materials that meet their level of knowledge, interest, and what they need to know to perform more effectively in an activity. E-Learning is more focused on the learner, and it is more interesting for the learner because it is information that they want to learn. From the S-D Logic perspective, FP10 states that values are unique, and phenomena depend on the user's benefits, while FP8 states the service is customer-centric. Its relationships and personalization enable users to achieve benefits by making their own choices.

4) Environment: A conducive e-learning platform environment is essential because it allows users to access information, resources, and services. From the S-D Logic point, an e-learning platform is a medium of service dissemination, and operant resource is the fundamental resource in inter-organizational and community. The environment on a quality e-learning platform plays a vital role in determining the use of e-learning [30]. The environment must give users a pleasant and acceptable environment to influence users regarding the platform's quality. The e-learning platform's environment refers to the medium or system developed to enable user operant and operand resources involving operational features that can create the process of integrating resources independently through value creation. Thus, the environment should consist of service quality elements, information quality and system, and technology quality. Well-designed e-learning may provide an environment that supports users' motivations, such as attractiveness, so users can choose to develop VCC based on their experience, resources, engagement, and goals [8]. FP3 can be associated with the e-learning platform as a service dissemination mechanism from the S-D Logic perspective, and FP4 states that operant resources are vital for competitive advantage.

B. Value Drivers

Actors as users are the person who uses services or products and their role in contributing on e-learning platform by using the services provided. Besides, actors derive value from their use of products or services, and therefore value delivery still represents a value creation type. Service exchanges can involve the co-creation or the self-creation of value, depending on the level of contribution provided by each actor in the service exchange [31]. Value-in-use means that value for the user is created or appears during participation [32] which indicates that the user is involved in the value

creation process [13]. Thus, as users use the platform, they together create co-creation value while benefiting through value-in-use. Actors have a role in contributing to the e-learning platform through value factors such as involvement, resources (skills and knowledge), experience, and goals.

1) Engagement: Engagement involves a commitment of time, energy, and the ability to configure operand and operant resources and an essential co-creation element. Actors' engagement on platforms includes posting comments and reviews, giving a rating, watching videos, and attending online classes. When users get engaged, it builds and strengthens the relationship and influences user faithfulness, and is expected to impact users' intentions toward future co-creation in other ways. Subsequently, they may gain knowledge related to the cognitive benefits of information acquisition. Students' participation in their learning process has been recognized as enhancing the educational process's quality and results [33]. Co-creators engagement in co-creation activities depends largely on their expectations and perceived motivations [34].

Further, actors would be motivated by rewards such as gifts, points, or social benefits from the title, status, and social esteem they may receive. They may desire to enhance their sense of self-improvement and enjoyment. Co-creation between actors can increase actor's engagement with the platform and further engage them to use it longer. A positive attitude builds good encouragement towards the institution [35]. Student engagement is considered crucial in optimizing the student experience, enhancing learning, and linking with high-quality learning outcomes and shared value creation.

2) Experience: Experience can be defined as an essential element in creating shared values, and the basis for working together is consumers' experience when using the products or services provided [16]. There is no value created in S-D Logic until product/services are used; thus, the experience is essential for value determination. Further, value comes from the knowledge shared, and continuous usage builds a collection of experiences. With experience, prepare users mentally to learn and apply personal learning to successfully perform current jobs and tasks. The user experience is affected when the user has had a direct or indirect relationship with a product/service and relates new behaviours to past experiences [36]. According to [37], participation in an online platform such as brand communities includes posting reviews and comments, give a rating, sharing experience, and others. Researchers acknowledge that positive experiences may benefit the organization while using the platform and increase user co-creation experiences [15]. Besides, when users have experience with a service, they would likely seek other activities on their next visit and enhance usage [38].

3) Resources: An actor who uses anything to establish value creation is called a resource [10] and anything an actor can draw on for support [39]. Resources comprise tangible, natural, and static resources and intangible and dynamic human ingenuity and appraisal functions. There are two types of S-D Logic resources: operand and operant resources [40]. From an e-learning perspective, operand resources are the tangible items that actors use to utilize, such as devices, internet connection, facilities provided, and users. Meanwhile, operant resources are resources that actors use to produce mental or physical skills and knowledge in increasing organizational competitiveness. Shortly, the user uses any operant resources to act on operand resources to get values.

In S-D Logic, two FPs, which are FP4 and FP9, are linked to resources. FP4 states, "operand resources are the fundamental source of strategic benefit, and FP9 states, "all social and economic actors are resource integrators [41]." In e-learning usage, users may invite other users to share their resources, cultivate resource sharing, and enable innovation to learn the material and extract value [42]. Tangible resources include lending equipment and intangible resources such as clicking like sharing, commenting, sharing ideas, and receiving opinions and statements. Shared learning material on the platform will allow users to retrieve and transfer material anytime and anywhere, enabling content to be more reusable and customizable.

The co-creation of the learning process and material can help produce improved material in line with other teachers' and students' actual requirements towards innovating the platform's services.

4) Goals: Goals refer to the desire to fulfil and motivate individuals to obtain knowledge. With a goal in mind, the user wants to learn and adapt to changes. [43] Studies indicate that students' achievement goals (performance or learning goals) are critical determinants of their attitude (e.g., satisfaction, perception, happiness), cognitive (e.g., critical thinking, information literacy), and affective goals (e.g., personal well-being, pleasure). When users' performance goals value highlighted and believed they had the high ability, they responded to overcoming the obstacles.

Interestingly, students' beliefs about their abilities, whether high or low, were irrelevant. Students sought to increase competence and opted for more challenging tasks to gain a learning goal's value. Motivation can influence learning, and the user tends to use products/services that they consider valuable. When users actively participate in learning activities, they prepare themselves for future academic pursuits and add exposure. Different users may have various underlying goals to use the VCC platform and engage in VCC [44]. Thus, the writer aims to find out the VCC in e-learning among vocational school teachers and students.

C. The E-learning Conceptual Model

Fig. 2 shows the e-learning conceptual model compromising the value proposition on the platform and value drivers for actors and their relationship. By adapting VCC as theoretical foundations, the model is divided into the value proposition and value drivers. The upper section shows the platform's value proposition elements, while the bottom section shows the value drivers for the platform's users. The model is developed to systematically illuminate how identified value drivers have the intention to use e-learning. The roles of users as main actors are to contribute to the e-learning

platform through engagement, share resources, experience, and goals. The concept of VCC prioritizes the value created when users use the service. At the same time, providers can enhance value-in-use by providing value proposition in platform features and supporting users to contribute continuously. Indirectly, a learning cycle exists to the user while using e-learning, and correspondingly, they together create VCC while benefiting through value-in-use.

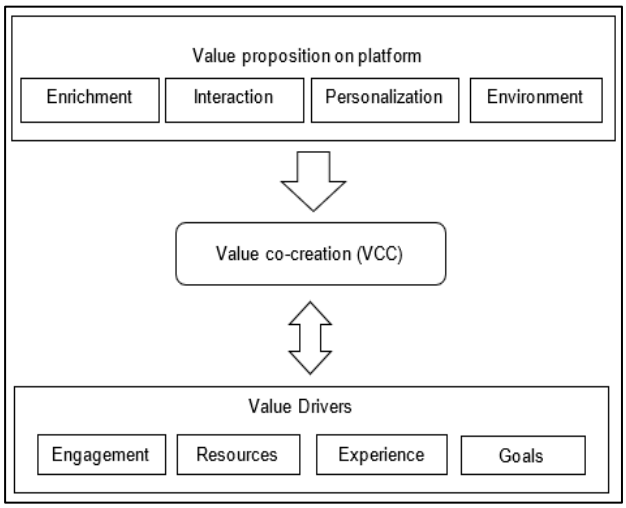


Fig. 2. E-learning Conceptual Model based on S-D Logic and VCC.

V. Conclusion and Future Plan

The Covid-19 pandemic has changed the lifestyle of many, including in the education system. Moving into online learning took some time for everyone since everyone is forced to use it. Most of the method used to develop e-learning platforms was focused on tangible items on e-learning platforms while value to the user and institutions is less considered. As mentioned, the concept of S-D Logic is still poorly implemented, especially with education. Thus, as this paper explored e-learning a platform that offers services in education, it would provide a new perspective in S-D Logic and VCC. The emergence of the area of S-D Logic and value co-creation, has thrived in imploring the research community to handle research the concepts, theory and methodologies of co-creation. This motivation has led us to analyze the idea of user's value in using e-learning services and search for a suitable value proposition on platform and value drivers for and from users that contribute the essence of value. This paper aims to identify value propositions components on the platform and value drivers for and by actors. When designing an e-learning platform, gathering VCC must be beneficial to the user's potential of value. However, the e-learning conceptual model does not cover all the other users, such as developers or system admin, due to the need to know the platform's primary users' values, teachers and students. Both the actors could contribute as participants and providers to the platform by being engaged, sharing experience and resources, and fulfilling their goals. Therefore, a new approach needs to be considered in the e-learning platform by studying the value element of co-creation through students' and teachers' value element approaches.

The next step is to identify the attribute on the e-learning platform that is crucial that can be implemented on the platform as a value proposition and lead actors to value. Further, the attribute of e-learning is matched to the dimension of the value proposition factors. Considering this research outcome would guide the developers of the critical features of the e-learning platform. Future research is to identify the critical features available on the e-learning platform from the users' view by using the means-end chain theory and laddering technique. Identifying the attributes with users will further motivate users to anticipate it is worthwhile using the service. The findings can help system developers to build a platform for institutes to manage the learning platform and understand how to create values for users.

Acknowledgment

The author would like to give appreciation to the Center for Software Technology and Management (SOFTAM), Faculty of Information Science and Technology (FTSM), Universiti Kebangsaan Malaysia (UKM), who have funded this paper through grants research (GGP-2020-038).

Reference

[1] P. C. Sun, K. C. Hsing, and F. Glenn, "Critical functionalities of a successful e-learning system - An analysis from instructors' cognitive structure toward system usage," Decis. Support Syst., vol. 48, no. 1, pp. 293–302, 2009.

[2] H. Nordin and D. Singh, "The Internationalization of E-Learning Websites : A Methodology," New Zeal. J. Comput. Interact. NZJCHI, vol. 1, no. 12, 2016.

[3] S. Motamarri, "Consumer Co-creation of Value in mHealth (Mobile Health) Service," J. Creat. Value, vol. 3, no. 1, pp. 63–76, 2017.

[4] C. J. Chen and S. W. Hung, "To give or to receive? Factors influencing members' knowledge sharing and community promotion in professional virtual communities," Inf. Manag., vol. 47, no. 4, pp. 226–236, 2010.

[5] I. T. Awidi, M. Paynter, and T. Vujosevic, "Facebook group in the learning design of a higher education course: An analysis of factors influencing positive learning experience for students," Comput. Educ., vol. 129, pp. 106–121, 2019.

[6] Y. Zhang, M. Zhang, N. Luo, Y. Wang, and T. Niu, "Understanding the formation mechanism of high-quality knowledge in social question and answer communities: A knowledge co-creation perspective," Int. J. Inf. Manage., vol. 48, no. July 2018, pp. 72–84, 2019.

[7] M. Ranjbarfard and M. H. Sureshjani, "Offering a framework for value co-creation in virtual academic learning environments," Interact. Technol. Smart Educ., vol. 15, no. 1, pp. 2–27, 2018.

[8] N. Rubio, N. Villaseñor, and M. J. Yague, "Does Use of Different Platforms Influence the Relationship between Cocreation Value-in-Use and Participants' Cocreation Behaviors? An Application in Third-Party Managed Virtual Communities," Complexity, vol. 2019, pp. 1–15, 2019.

[9] T. Beckman and A. Khare, "A Service-Dominant Logic and Value Co-creation Approach for Online Business Education," in Springer International Publishing AG 2018, 2018, pp. 21–35.

[10] R. F. Lusch and S. Nambisan, "Service Innovation: A Service-Dominant Logic Perspective," MIS Q., vol. 39, no. 1, pp. 155–175, 2015.

[11] S. Ghavifekr, "Factors affecting use of e-learning platform (SPeCTRUM) among University students in Malaysia," Educ. Inf. Technol., vol. 22, no. 1, pp. 75–100, 2017.

[12] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores," Comput. Intell. Neurosci., vol. 2018, 2018.

[13] C. Grönroos, "Value co-creation in service logic: A critical analysis," Mark. Theory, vol. 11, no. 3, pp. 279–301, 2011.

[14] S. L. Vargo and R. F. Lusch, "Why 'service'?," J. Acad. Mark. Sci., vol. 36, no. 1, pp. 25–38, 2008.

[15] D. Pacauskas, "The Role of ICT in the Value Co-Creation Process," Aalto University publication series, 2016.

[16] E. Y. Chou, C. Y. Lin, and H. C. Huang, "Fairness and devotion go far: Integrating online justice and value co-creation in virtual communities," Int. J. Inf. Manage., vol. 36, no. 1, pp. 60–72, 2016.

[17] Muriati Mukhtar, Mohamed Nazul Ismail, and Yazrina 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.

[18] M. T. Delgado et al., "Value Co-Creation Mechanisms," IoT Eur. Platf. Initiat., p. 75, 2016.

[19] W. A. Z. W. Ahmad, M. Mukhta, and Y. Yahya, "Evaluating the applicability of a social content management framework: A case analysis," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 12, pp. 339–345, 2018.

[20] E. A. Thangaiah, R. Jenal, and J. Yahaya, "Penerokaan Penggunaan E-Pembelajaran dalam Kalangan Pelajar dan Pengajar TVET - Satu Kajian Awal (Investigating the E-Learning Usage Among TVET Students and Teachers - A Preliminary Study)," Akademika, vol. 90, no. 3, pp. 5–18, 2020.

[21] F. Buttle and M. Stan, Customer relationship management: concepts and technologies. Routledge, 2019.

[22] T. Tuunanen and H. Govindji, "Understanding flow experience from users requirements," Behav. Inf. Technol., vol. 35, no. 2, pp. 134–150, 2016.

[23] Mohamed Nazul Ismail, Yazrina Yahya, and Muriati Mukhtar, "Nilai cipta-sama sistem pengurusan pembelajaran," J. Teknol. (Sciences Eng., vol. 60, pp. 21–29, 2013.

[24] R. Mcdaniel, J. R. Fanfarelli, and R. Lindgren, "Creative Content Management: Importance, Novelty, and Affect as Design Heuristics for Learning Management Systems," IEEE Trans. Prof. Commun., vol. 60, no. 2, pp. 183–200, 2017.

[25] R. Bidar, A. Barros, and J. Watson, "Co-creation of services: an online network perspective," Internet Res., vol. ahead-of-p, no. ahead-of-print, 2021.

[26] Prahalad C.K. and V. Ramaswamy, "Co-creating unique value with customers," Strateg. Leadersh., vol. 32, no. 3, pp. 4–9, 2004.

[27] M. M. Daniels, E. Sarte, and J. Dela Cruz, "Students' perception on e-learning: A basis for the development of e-learning framework in higher education institutions," IOP Conf. Ser. Mater. Sci. Eng., vol. 482, no. 1, 2019.

[28] H. M. Truong, "Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities," Comput. Human Behav., vol. 55, pp. 1185–1193, 2016.

[29] S. Ouf, M. Abd Ellatif, S. E. Salama, and Y. Helmy, "A proposed paradigm for smart learning environment based on semantic web," Comput. Human Behav., vol. 72, pp. 1–23, 2016.

[30] H. Ataburo, A. S. Muntaka, and E. K. Quansah, "Linkages among E-Service Quality, Satisfaction, and Usage of E-Services within Higher Educational Environments," Int. J. Bus. Soc. Res., vol. 7, no. 3, pp. 9–26, 2017.

[31] N. Zainuddin, L. Tam, and A. McCosker, "Serving yourself: value self-creation in health care service," J. Serv. Mark., vol. 30, no. 6, pp. 586–600, 2016.

[32] N. Bendapudi and R. P. Leone, "Psychological Implications of Customer Participation in Co-Production," J. Mark., vol. 67, pp. 14–28, 2003.

[33] S. A. Taylor, G. L. Hunter, H. Melton, and S. A. Goodwin, "Student engagement and marketing classes," J. Mark. Educ., vol. 33, no. 1, pp. 73–92, 2011.

[34] R. Bidar, J. Watson, and A. Barros, "Classification of service co-creation systems: An integrative approach," Int. Conf. Adv. Commun. Technol. ICACT, pp. 333–340, 2017.

[35] S. L. Huang and C. T. Chen, "How consumers become loyal fans on Facebook," Comput. Human Behav., vol. 82, pp. 124–135, 2018.

[36] A. F. Payne, K. Storbacka, and P. Frow, "Managing the co-creation of value," J. Acad. Mark. Sci., vol. 36, no. 1, pp. 83–96, 2008.

[37] S. Kamboj, B. Sarmah, S. Gupta, and Y. Dwivedi, "Examining branding co-creation in brand communities on social media: Applying the paradigm of Stimulus-Organism-Response," Int. J. Inf. Manage., vol. 39, no. October 2017, pp. 169–185, 2018.

[38] R. A. Rather, L. D. Hollebeek, and S. M. Rasoolimanesh, "First-Time versus Repeat Tourism Customer Engagement, Experience, and Value Cocreation: An Empirical Investigation," J. Travel Res., 2021.

[39] M. Blaschke, M. K. Haki, S. Aier, and R. Winter, "Value Co-creation Ontology — A Service-dominant Logic Perspective," Multikonferenz Wirtschaftsinformatik 2018, pp. 398–409, 2018.

[40] S. L. Vargo and R. F. Lusch, "Institutions and axioms: an extension and update of service-dominant logic," J. Acad. Mark. Sci., vol. 44, no. 1, pp. 5–23, 2015.

[41] S. L. Vargo and R. F. Lusch, "Service-dominant logic: continuing the evolution," J. Acad. Mark. Sci., vol. 36, no. 1, pp. 1–10, 2008.

[42] A. Sood, "Value As An Aid For Understanding Perceived Service Quality Of Digital Services : The Jyu Faculty Of Information Technology As A Case Study," University Of Jyväskylä, 2019.

[43] K. M. Judson and S. A. Taylor, "Moving from Marketization to Marketing of Higher Education : The Co-Creation of Value in Higher Education," High. Educ. Stud., vol. 4, no. 1, pp. 51–67, 2014.

[44] M. Zhou, X. Cai, Q. Liu, and W. Fan, "Examining continuance use on social network and micro-blogging sites: Different roles of self-image and peer influence," Int. J. Inf. Manage., vol. 47, no. January, pp. 215–232, 2019.

An Efficient Aspect based Sentiment Analysis Model by the Hybrid Fusion of Speech and Text Aspects

Maganti Syamala[1]
Research Scholar, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram, Tamil Nadu 608002, India
Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation Vaddeswaram, AP, India

N.J.Nalini[2]
Department of Computer Science and Engineering
Annamalai University
Annamalai Nagar, Chidambaram
Tamil Nadu 608002
India

***Abstract*—Aspect-based Sentiment Analysis (ABSA) is treated to be a challenging task in the domain of speech, as it needs the fusion of acoustic features and Linguistic features for information retrieval and decision making. The existing studies in speech are limited to speech and emotion recognition. The main objective of this work is to combine acoustic features in speech with linguistic features in text for ABSA. A deep learning and language model is implemented for acoustic feature extraction in speech. Different variants of text feature extraction techniques are used for aspect extraction in text. Trained Lexicons, Latent Dirichlet Allocation (LDA) model, Rule based approach and Efficient Named Entity Recognition (E-NER) guided dependency parsing approach has been used for aspect extraction. Sentiment with respect to the extracted aspect is analyzed using Natural Language Processing (NLP) techniques. The experimental results of the proposed model proved the effectiveness of hybrid level fusion by yielding improved results of 5.7% WER and 3% CER when compared with the traditional baseline individual linguistic and acoustic feature models.**

***Keywords*—Acoustic; aspect-based sentiment analysis; decision making; emotion; extraction; hybrid; lexicon; linguistic; natural language processing; speech**

I. Introduction

Sentiment analysis or opinion mining is the area of study in NLP where it helps to analyze the polarity with respect to the given context. Sentiment analysis depicts the state-of-art-of-mind to automate the process of analyzing the opinion, emotion, polarity, appraisal, interest, ideology, attitude, feelings towards an entity. Sentiment analysis plays an important role in our daily lives for analysis and decision making. In most of the existing studies, sentiment analysis is been carried out on text and the performance is been differentiated by varying the type of linguistic features extracted from text. The features on text are generally called as linguistic features and play a very crucial role in sentiment analysis. Due to the tremendous growth of data in World Wide Web, now-a-days traditional and web-based surveys are been replaced by sentiment analysis [1]. As WWW is a combination of text, audio and video, there is a need for analysis of sentiment on multimodal data. Feature extraction for sentiment analysis will be differed for different types of input like text, audio and video. The field of sentiment analysis in NLP had gained its popularity by implementing on text. By the evolution of massive data, research is been expanded and now it's confined not only to text but also had gained its popularity in different modalities. When sentimental analysis came into picture, it's been carried out only on text using NLP and machine learning techniques, where the polarity of the given document or sentence is classified as either positive, negative or neutral [1]. Next era of sentiment analysis is aspect-based sentiment analysis (ABSA) and had gained its popularity in recommender systems. Most of the recommender systems that used ABSA have identified the sentiment with respect to the aspect in the given text. Parts-of-Speech (POS) tagging was one of the widely used aspect identification technique for ABSA [4]. In this paper, aspect-based sentiment analysis was been carried out by combining both audio and text features.

Most of the research so far carried out on audio data is confined to speech analysis and emotion recognition. In the existing studies [6], various acoustic features are analyzed and are classified for speech emotion recognition. Identifying sentiment in speech is a challenging task because of following reasons.

- Even though both the terms emotion and sentiment express feelings with respect to the context but the way they are analyzed is different.
- Emotion is the one that can be analyzed in speech by means of various acoustic features and prosodic features like pitch, intensity, energy, loudness etc. Whereas in text the sentiment is defined as an adjective that qualifies the respective noun.
- There is a difficulty to map emotion in speech with parts-of-speech in text for analyzing the sentiment. Even though there are many existing studies carried out on speech for sentiment analysis, the work is limited in analyzing only the emotion in speech like happy, sad, angry, fear and etc.; but not the positivity and negativity in the given context.

As speech and text features are different, so there is a need to bridge the gap between them to perform sentiment analysis. Speech in call-centers and text in recommender systems has gained its popularity in the field of sentimental analysis [15]. Fig. 1 depicts the sentiment analysis model by considering bi-modal speech and text features.

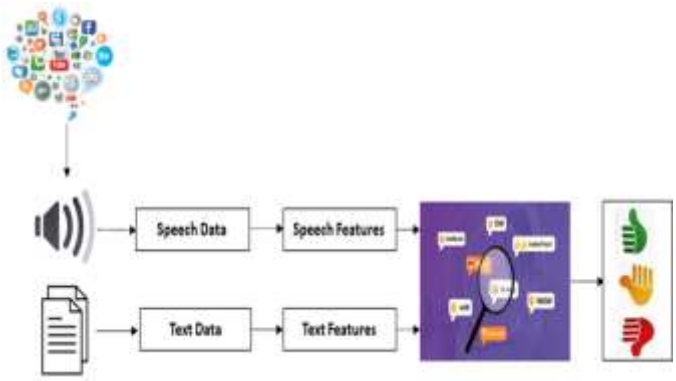


Fig. 1. Speech and Text based Bi-modal Sentiment Analysis.

The main contributions of the proposed work are:

- The importance of linguistic and acoustic features for ABSA is analyzed.
- A hybrid level fusion of acoustic and linguistic features for ABSA is evaluated using Word Error Rate (WER) metric and machine learning algorithms.
- The obtained results from proposed combined model are validated with the individual implementations of speech and text-based sentiment analysis.

II. Motivation

The field of sentiment analysis is catching everyone's attention in marketing, corporate and academia by executing the tasks in an easy and efficient manner. But most of the traditional frameworks are confined to work only either on text or audio or video. There is very limited study carried out on multimodal data. Now-a-days, sentiment is been analyzed as aspect-based sentiment analysis and major limitations are been identified in feature extraction and sentiment related aspect category identification. So, this made my work to drive towards implementing aspect-based sentiment analysis on multimodal speech and text data. Identification of sentiment with respect to the aspect helps to improve quality of service when compared with document and sentence level sentiment classification.

Now-a-days, tremendous growth of data available in social media and online commercial websites made everyone to provide online reviews demonstrated as a video in YouTube. Previously, consumers used take their decision for any purchase by analyzing the text reviews given by the customers [16]. In some cases, like where there is no customer who had already bought the product, there will be no rating and review provided for that product. In such cases consumer is not in a state to make a decision whether to go for it or not. So, this made me to develop an aspect-based sentiment analysis model on YouTube review data for improving the quality of service to consumers.

III. Related Work

The main objective of the proposed model is to analyze Aspect based sentiment analysis by combining both linguistic and acoustic features. Acoustic feature extraction techniques and Linguistic feature extraction techniques are applied for feature extraction [14] on YouTube product review dataset. The base line models are implemented by considering individual linguistic and acoustic features, validated using machine learning algorithms. A hybrid level fusion of acoustic and linguistic features for ABSA yields improved results when measured in terms of accuracy, precision, recall and F-score.

Existing methodology in sentiment analysis had used bag of words, parts-of-speech tagging as feature extraction techniques on text [2]. The work is limited to classify domain specific sentiment and resulted in document level sentiment classification with poor efficiency. Evolved topic modelling and by the use of LDA, it is made possible to classify sentiments by grouping into topics. But this approach is limited to automate the process of assigning labels by grouped topics, where manual assignment is needed. The literature in this paper is carried out to analyze the impact of aspect-based sentiment analysis by considering linguistic and acoustic features.

A. Linguistic Features: Aspect-based Sentiment Analysis on Text Data

Sentiment analysis had a wide variety of applications experimented on textual data. The evolution of sentiment analysis made the job of many real time applications easy in commercial markets for analyzing the customer, employee feedback in a working organization, recommending a product in ecommerce, decision making in any kind of purchase, political opinion, movie reviews, etc. Many studies have been carried to identify sentiments on text at various levels like document level, sentence level, aspect level, context level [1].

Md. E. Mowlaei et al; proposed adaptive lexicon-based ABSA [2] using three different types of lexicons like opinion lexicon, Sent-WordNet, Subjective to implement dynamic aspect-based sentiment analysis. The proposed methodology overcomes the limitations of existing domain dependent static lexicon approaches. The model lacks to identify implicit aspects even though it draws the attention to identify context dependent aspects in a dynamic way.

O. Alqaryouti et al; to improve the efficiency of sentiment classification proposed an integrated lexicon and rule-based approach [3] for aspect-based sentiment analysis to identify both implicit and explicit aspects. But the Lexicons used for generating the aspects are manually assigned to achieve higher efficiency in identifying the implicit and explicit aspects. A rule-based approach is used to integrate the extracted aspects and sentiments for classification. The model is implemented on government review data where general public post their opinions and it was suggested that it can be useful in mobile apps to analyze the feedback from public or customers.

V.S. Anoop et al; proposed an aspect-based sentiment analysis model on text using a topic modelling technique called LDA [4]. The input text by the use of LDA algorithm is segmented into topics, which then mapped manually to a relevant aspect. In case where there is a need to process huge data for sentiment analysis, it will be very difficult.

M. Shams et al; proposed a language independent aspect-based sentiment analysis model which undergoes through three phases of fine-grained operations [5]. The aspects are extracted

by having prior knowledge on dataset been used and used aspect word sets for mapping the polarity to the aspect. And finally used an expectation-maximization algorithm for calculating weightage of each word with respect to its aspect and assigned sentiment.

M. Syamala et al; to overcome the limitation of manual topic label assignment to the topics extracted from LDA proposed a deep fusion mechanism [19]. The extracted topics from LDA are converted into word embeddings and trained over a one-layer neural network to determine topic label for each set of extracted topics. The proposed sentiment classification model was compared against the models implemented with LDA and without LDA.

B. Acoustic Features: Aspect-based Sentiment Analysis on Audio Data

Most of the research carried out on speech for analysis is on either speech recognition or emotion recognition. Emotion recognition in speech differs from sentiment identification in text. Recognition of emotion in speech depends on various factors like pitch, volume, frequency, time, intensity, jitter, noise, and etc. But in case of text, identification of sentiment is independent of all the external environmental factors. So, there is a need to know the fusion mechanism between speech and text features for performing sentiment analysis. In this section, some of the existing works carried out on speech data for sentiment analysis is presented.

D. Griol et al; proposed a fusion mechanism between speech and text features [6]. The features extracted from both these modalities are trained for emotion classification in speech and sentiment classification in text. Acoustic and contextual features are been extracted from speech for emotion classification and semantic features are been extracted from transcriptions for sentiment identification. The proposed fusion model takes the account of peculiar context related errors in the transcriptions derived from speech.

Zhiyun Lu et al; proposed an end-to-end automatic speech emotion recognition model using pre-trained speech and text features from IEMOCAP dataset [7]. Build a speech sentiment database to enhance the sentiment in speech and also which is been considered as one of the current challenges in this field of research. The trained features are classified using a self-attention Recurrent Neural Network (RNN) to differentiate sentiment with respect to language model.

Bryan Li et al; combined acoustic and lexical features to develop a sentiment analysis model in order to analyse customer call services [8]. The acoustic low-level descriptors like MFCC, Intensity, pitch, loudness features are extracted using open SMILE. Lexical features are been extracted by considering n-grams. The Lexical classifier model was built on IEMOCAP dataset to make a comparison between the speech transcriptions and to choose the perfect speech recognition model. Implemented a decision-level fusion mechanism also known to be a late fusion to train the two modalities input to a classifier for decision making or classification.

Dong Zhang et al; proposed a REINFORCED approach which differs from self-attention model by concentrating on the word-level features in both speech, text and avoided the low level weighted and noisy features [9]. In this paper, the title of the paper is to depict sentiment on speech and text but actually emotions are been classified by training the extracted features into a deep learning model using SoftMax layer.

Maghilnan S et al; proposed speech sentiment analysis on speaker specific data [10]. In the proposed model, conversation between two entities is taken as input but can't able to handle if both the entities speak simultaneously. Two independent tasks are carried out to perform speaker identification and speech transcribes generation. Later both these outputs are used to map the transcribed text with respect to its speaker ID. Finally, the output text dialogue is classified into sentiment based on its polarity.

IV. Deep Fusion of Linguistic and Acoustic Features

As the proposed model in this paper analyses ABSA by considering both speech and text data, it's important to know the different ways of fusing linguistic and acoustic features. In general, the research that is been carried out in this area, defines three basic variants of fusing mechanisms like feature-level fusion, decision-level fusion, hybrid-level fusion.

A. Feature-Level Fusion

Feature-level fusion is also known as early-fusion where features from various modalities are extracted separately and a deep classification analysis was performed by fusing the models to enhance the performance. The main advantage with this type of fusion is, in the early stage it helps to derive or extract modality dependent features making the models to achieve more improvement. The main drawback with feature level fusion is that the aspects with respect to the modality may differ and accurate analysis can't be achieved when combined analysis is performed. For example, in speech the features are acoustic and in text features are linguistic. Poria S et al; in his paper multimodal emotion recognition and sentiment analysis [11], used feature-level fusion to fuse three modalities of YouTube data. A deep convolutional neural network was used to extract speech and visual features and word-embeddings, parts-of-speech tagging was used to extract textual features. A multiple kernel learning classifier is used to fuse and analyze the sentiment.

B. Decision-Level Fusion

In decision-level fusion, the features from different modalities are extracted separately and classified separately. The results obtained from each classification are merged into a feature vector for final decision making. The advantage of this approach is that the final feature vector obtained from decision fusion of individual modalities will be in same format so that no conversion is required. The drawback of this fusion is to perform classification on different modalities involves different types of classifiers.

Wöllmer M et al; used decision level fusion mechanism in his paper [12] to fuse audio, visual and text features of YouTube input data. The extracted acoustic, visual features are trained by a LSTM for sentiment score evaluation and Support Vector Machine (SVM) is used to train and derive the sentiment score of textual features. Final decision level late fusion was performed for final sentiment prediction by

calculating the weighted sum on the sentiment score obtained by assigning a weightage of about 1.2 to linguistic and 0.8 for audio and visual score.

C. Hybrid-Level Fusion

Hybrid-level fusion includes the model to use both feature and decision-level fusion mechanism in order to overcome the drawbacks in individual fusions.

Yue Gu et al, proposed an attention-based hybrid multimodal network for spoken language classification using hybrid fusion approach [13]. Word2Vec and Mel-frequency spectral coefficients (MFSCs) of text and audio features are been extracted. The extracted features are individually trained over a LSTM to obtain informative context related words and frames undergoing a feature level fusion. And finally, modality level fusion i.e., a decision-level fusion is performed by passing the extracted individual test and audio features through an attention layer to extract informative modality level features.

V. Proposed Model

In this paper, a novel Aspect-based Sentiment Analysis model was implemented on speech and text data. The dataset used for implementing the model is drawn from YouTube social platform. In order to evaluate the experimental results for sentiment modality comparison, both the speech and text models are been tested on the same dataset. The domain chosen for carrying out our experimental analysis is real-time product review data. In the initial phase, the raw audio format of the product review YouTube Video is trained over a speech analysis model. The speech analysis model maps the acoustic spectrogram features of the speech signal into the respective word utterances using a deep learning and language model. The word utterances from the speech analysis model are trained over different variants of text feature extraction techniques for deriving related and relevant aspects. The sentiment with respect to the derived aspect is analyzed for performing Aspect-based sentiment analysis. The components (features) in speech and text data are processed individually and are then fused. So, the whole process uses a hybrid fusion mechanism for mapping speech and text features for performing ABSA. Fig. 2 explains the work flow of the proposed speech and text analysis model for efficient Aspect based Sentiment Analysis.

A. YouTube Product Review Data Collection and Processing

In this phase, YouTube product reviews of Samsung M31mobile were downloaded as dataset. YouTube, a social platform where people share their live experience in the form of reviews have a natural, spontaneous speaking style. As the way the speaker speaks have a direct impact on describing the accuracy of the model, made me to motivate and download the dataset from YouTube for performing Aspect-based sentiment analysis on speech data. In total 40 YouTube reviews of size 90 KB on the Samsung M31 product having strong presence of subjectivity, positivity and negativity are randomly collected and converted to .wav files, are used for ABSSA.

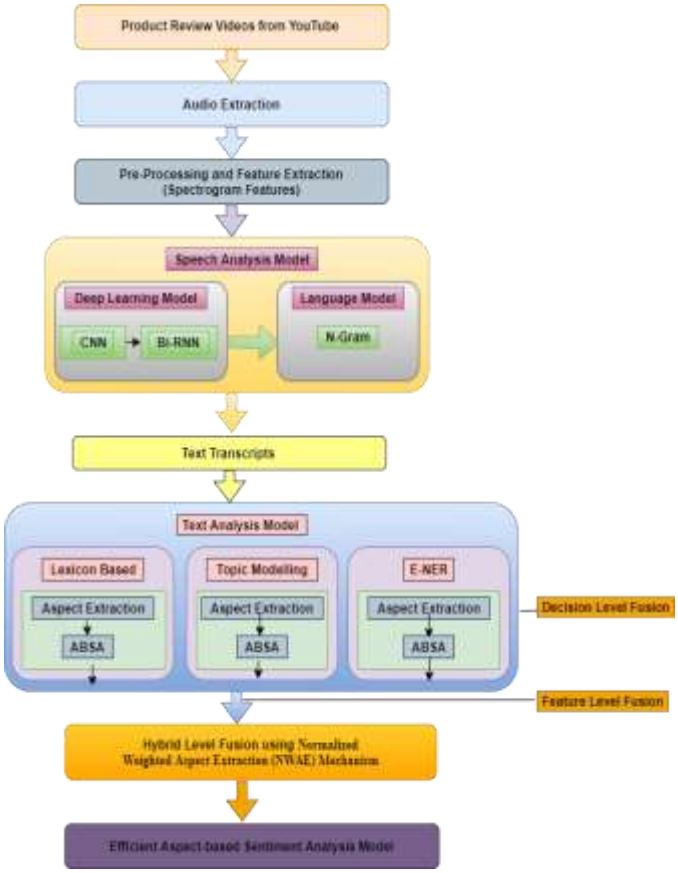


Fig. 2. Work flow of Proposed Speech and Text Analysis Model for Efficient Aspect-based Sentiment Analysis.

B. Speech Analysis Model

Emotion in speech is treated to be a kind of sentiment, which expresses an individual feeling in terms of happy, sad, fear, disgust, angry and etc. Sentiment in text differs from emotion in speech and there is a need to perform speech analysis in the form of automatic speech recognition to enhance sentiment from audio data. There are many ASR models and online speech - text conversion API's.

To enhance the performance of traditional ASR models and to overcome the limitations in online speech-text API's, in our proposed model used a deep learning framework for analyzing the acoustic features and a bi-gram language model to map the word utterances. As sentiment analysis is independent on the speech features like pitch, intensity, volume and etc. Initially, the spectrogram features are extracted from the input Wav audio file and are trained over a Convolutional Neural Network and a Bi-directional Recurrent Neural Network (Bi-RNN). The acoustic features when trained over these deep neural networks produces a character sequence of spoken utterances. A bi-gram language model by the use of chain rule retrieves the maximum occurrence of character sequences and the same are mapped into word utterances. Fig. 3 shows the text transcripts extracted from the proposed speech analysis model.

Source.Name	message
Samsung Galaxy M21 review _ Better than the Ga...	okay so what are the questions that I get ask...
Samsung Galaxy M31 - Full Review, 64 MP Quad C...	so after the galaxy m31 the m30 s we now have ...
Samsung Galaxy M31 Detailed Camera Review - En...	hey guys did saga from tech works and in this...
Samsung Galaxy M31 Full Review _ Better than R...	hey guys this is one from guiding dick and to...
Samsung Galaxy M31 Review - Decent Device, Dis...	some sense m-series our phones have always bee...
Samsung Galaxy M31 Review 2 Weeks Later! The B...	what's going on guys my name is Wade with tech...
Samsung Galaxy M31 Review with Pros.txt	hi guys this is Ranji them in this video let's...
Samsung Galaxy M31 Review_ Should You Buy_ - E...	so the galaxy m31 arrived in India recently a...
Samsung Galaxy M31 review_ Worth all the Hype_...	Samson's MCV smartphones have been quite popul...
Samsung Galaxy M31 The Only Review You Need - ...	you guys welcome to I again today we're going...

Fig. 3. Extracted Text Transcripts from the Proposed Speech Analysis Model.

1) Creation of Spectrogram

Input: Audio signal
Output: One-Time frame vector
Step 1: Dividing the input audio signal into time frames of frame size 1024, with a sampling rate of 16 kHz.
Step 2: Each frame signal is then split into its frequency components with a hop size of 512 samples between each successive Fast Fourier Transform window.
Step 3: Finally, each time frame is then represented as a one-time frame vector with a vector of amplitudes at each frequency.
Step 4: The one-time frame vectors obtained when lined up in time series order gives us the visual representation of input audio signal as a spectrogram.

2) Language model: The word sequences obtained from the above acoustic model need to be refined as the acoustic model finds the probability of character utterances based on sound and there are cases where two words can utter same sound. The use of language modelling followed by acoustic modelling helps to rectify this problem and increases the likelihood score of a particular sequence of word utterances. Equation (1) formulates the representation of a sequence of word utterances using N-gram language model. In this proposed speech analysis model, used a bi-gram language model by a chain of rule mechanism to find the respective sequence probability.

$$P(\mathrm{Word}_1, \mathrm{Word}_2, \ldots . \mathrm{Word}_n) = \pi\, \mathrm{P}(\mathrm{Word}_i | \mathrm{Word}_1, \mathrm{Word}_2, \ldots\, \mathrm{Word}_{i-1}) \quad (1)$$

3) Text analysis model: For improving the efficiency of Aspect-based sentiment analysis, in this paper three variants of text feature extraction techniques are applied for aspect level feature extraction. Sentiment was analyzed with respect to the extracted aspect at decision level. The way the decision level aspects are extracted and analyzed for sentiment are presented in detail in the below section.

a) Lexicon based semi-supervised pattern generation technique (Model 1): The opinion words representing the sentiment called as aspect terms are extracted by generating patterns from bi and tri grams. As defined, it is a Lexicon based approach, the similar aspect words related to the input dataset are assigned statistically. In addition to statistically stuffed aspect words, the hypernyms of aspect terms extracted from patterns are generated by using wordnet. The final aspect terms are obtained by considering statistically assigned words and its similar words generated from the patterns.

Consider a set of word sequences

$$\mathrm{Word}_1{}^n = \mathrm{Word}_1 \ldots\ldots\ldots \mathrm{Word}_n \quad (2)$$

Bigram approximation is represented as

$$\mathrm{P}(\mathrm{Word}_1{}^n) = \prod_{k=1}^{n} \mathrm{P}(\mathrm{Word}_k / \mathrm{Word}_{k-1}) \quad (3)$$

N-gram approximation is represented as

$$\mathrm{P}(\mathrm{Word}_1{}^n) = \prod_{k=1}^{n} \mathrm{P}(\mathrm{Word}_k / \mathrm{Word}_{k-N+1}^{k-1}) \quad (4)$$

The final aspect terms obtained are mapped with the patterns generated to extract the sentiment terms. Sentiment score was computed on the trained aspect and sentiment terms by importing 'testimonial. sentiment. polarity' library.

b) Topic Modelling Technique LDA (Latent Dirichlet Allocation) (Model 2): Text features are been extracted from the pe-processed input using LDA. Python libraries like gensim and ldamallet are used for extracting the dominant topics (aspect terms). Extraction is done by calculating the term document frequency on pre-processed lemmatized data by considering NOUN, ADJ, ADV and VERB from n-gram data. Dominant topic words termed to be as aspects with respect to the input qualifying the sentiment are extracted. Some of the examples of aspect terms in the context of electronic gadgets are battery, display, power etc.

Probability based topic extraction using LDA is formulated in (5).

$$\mathrm{P}(z = t/w) \infty \left(\alpha_t + n_{t/d}\right) \frac{\beta + n_{w/t}}{\beta V + n_{\cdot/t}} \quad (5)$$

Aspect category groups a list of aspect terms into its relevant category. For example, the aspect terms like battery, display, power can be categorized under the category mobiles and similarly taste, flavor, ambience can be categorized under the category restaurant. Aspect category is detected by training the extracted dominant/aspect terms into a Convolutional Neural Network (CNN). Using polarity as a measure, respective sentiment terms are extracted from the extracted aspect category terms.

c) Efficient Named Entity Recognition (E-NER) (Model 3): The aspect terms in this approach are extracted by a dependency parsing mechanism using POS tagging, an NLP technique.

A convolutional neural network was used to map the extracted aspect as relevant aspect categories. Word embeddings mechanism (6), (7) is used to train the input aspect terms as vectors to CNN (8). Filtering aspect related sentiment words and aspect sentiment classification uses the same

methodology as followed for aspect category detection for aspect term polarity extraction and sentiment classification.

$$m_i = \sum_{j=1,j\neq i}^{n}(a_{ij}, w_j) \quad (6)$$

$$a_{i,j} = \frac{\exp(score(w_i,w_j))}{\sum_{j1=1}^{n}\exp(score(w_i,w_j))} \quad (7)$$

$$score(w_i, w_j) = v_a^T \tanh(W_a[w_i \oplus w_j]) \quad (8)$$

4) Hybrid level fusion: In text analysis model, by the three variants of aspect extraction techniques we performed decision level fusion for analyzing the sentiment. The extracted aspects with respect to their sentiment have undergone a feature level fusion for enhancing the performance. By means of this decision level fusion followed by feature level fusion, helps to overcome the problems of dimensionality and filters the weighted aspects by deriving improved performance. In hybrid level fusion phase, employed a Normalized Weighted Aspect Extraction (NWAE) mechanism (10) in which the aspects extracted from each technique are filtered based on their weights. A decision rule was applied for classifying the polarity class of the derived weighted aspects (11).

$$tf - idf(d_i, v_j) = \frac{count(d_i v_j)}{\sum_{v^1 \in d_i} count(d_i, v^1)} \times \log \frac{n}{|\{v_j \in d^1, d^1 \in D\}|} \quad (9)$$

$$NWAE = \frac{1}{|P_j|}\sum_{d_i \in p_j} d_i \quad (10)$$

$$y_t = arg\ max_{pj} \sum_{x_i \in x_t}(1 - dist(x_t, x_i))I(x_i, c_j) \quad (11)$$

Input: Extracted Aspect terms A_t in d_1, d_2, d_3.
Output: Weighted Aspects W_a and its polarity.

```
file_path <- file.path(" d_1, d_2, d_3")
docs <- Corpus (DirSource (file_path))
docs <- Corpus (VectorSource(docs)) #This tells R to treat your preprocessed documents as text documents.
dtm <- DocumentTermMatrix(docs)
tdm <- TermDocumentMatrix(docs)
dtms <- removeSparseTerms(dtm, 0.1) #  Start by removing sparse terms
tf.idf<- weightTfIdf(dtm,normalize=TRUE)
x <- apply (tf.idf, 1, sum) #computing sum of the rows in tf-idf matrix
d <- NULL
for (i in seq(docs)) #NWAE
  d[i]<-x[i]/2
#Sum of the squares of the NWAE
df <- NULL, df1 <- NULL, df1 <-0
for (i in seq(docs))
  df[i] <- (d[i]) ^2
  df1<- df1+df[i]
#Sum of the squares of the documents
n <- ncol(dtm), n1 <- nrow(dtm), sf <- NULL, s <-NULL, w <- NULL
for (z in seq(docs))
  sf[z] <- 0
  for (i in seq(n))
    w[i]<- inspect (tf.idf[z [1], i])
    w[i]
    s[i] <- (w[i]) ^2
    sf[z] <- sf[z]+s[i]
#Similarity function for dimensionality reduction
sim <- NULL
for (i in seq(docs))
  sim[i] <- (x[i]*d[i])/((sqrt(sf[i])) * (sqrt(df1)))
#Projected documents values sum of the squares
p <- NULL, pro <- NULL, p <- 0
for (i in seq(docs))
  pro[i] <- (sim[i]) ^2
  p <- p+pro[i]
#Normalize the projected document vectors
v <- NULL
for (i in seq(n1-1))
  v[i] <- (sim[i]*sim[n1])/((sqrt(p)) * (sqrt((sim[n1]) ^2)))
#Compute Euclidean distance
m <- NULL
for (i in seq(n1-1))
 m[i]<-dist(v[i], sim[n1])
#Apply decision rule
y <- NULL
for (I in seq(n1-1))
 y[i] <- (1-m[i])
 max(y)
for (i in seq(n1-1))
 if(y[i]==max(y))
   print ("The aspect term is classified as:")
   print(i)
```

VI. Experimental Results

Experimental results in this paper are been evaluated on product review content drawn from YouTube videos. The data set in its initial video format is processed into Wav files for performing speech recognition using deep learning and language models. Word Error Rate (WER) and Character Error Rate (CER) are the two-evaluation metrics used for recognizing the performance of the speech recognition model. The proposed speech recognition model proved to improve the efficiency of the model by achieving 5.7% WER and 3% CER. The results proved to improve the performance of the proposed model when compared with the traditional state of art methods. The aim to perform aspect-based sentiment analysis on speech data made us to carry out the analysis on the derived text transcripts by using three different feature extraction techniques.

The application of feature extraction techniques improved the efficiency of the proposed model. The fusion of aspects at decision-level after undergoing individual feature-level fusion improvises the feature selection process and overcomes the problem of dimensionality.

Experimental results obtained from the three different text analysis models, discussed in Section 5 are made a comparison. Accuracy, precision, recall and f-score are the metrics used to measure the performance of the proposed model.

Fig. 4 lists the different aspects extracted from the patterns generated by means of bi-gram and tri-grams in text analysis model1.

Fig. 4. List of Aspects Extracted from Model 1.

Model 1, uses a semi-supervised pattern generation technique for aspect extraction, where it needs a list of statistically assigned aspects (Lexicon) relevant to the taken input context. So, Fig. 5 shows the list of statistically assigned aspects used in Model 1.

```
stuff = ['software', 'application', 'service', 'power supply', 'sim card', 'display',
         'storage space', 'sensor', 'wireless charging', 'design', 'cpu', 'accessories',
         'camera','quality','time','condition','screen','price','case','build','access',
         'battery','buy','power','switch','light','design','technology','radio','fashion'
'product','charging','feature','touch','profile','car','slot','tables','construction',
'period ','system','game','bottom','sound','blackberry charge','price anyone','price extra',
'cord length','charge port',' phone','horizon charge','fraction price','charge ','key',
'extension','internet','cheap','cover','speaker']
```

Fig. 5. Statistically Assigned Aspects in Model 1.

For effective aspect extraction in model 1, hypernyms are generated for the statistically assigned aspects. Extracted hypernyms for the statistically stuffed aspects are shown in the below Fig. 6.

```
Meaning 0 NLTK ID:  software.n.01
Hypernyms: code, computer_code
Hyponyms: alpha_software, authoring_language, beta_software, compatible_software, computer-aided_design,
CAD, database_management_system, DBMS, freeware, groupware, operating_system, OS, program, programme, computer_program, compu
ter_programme, routine, subroutine, subprogram, procedure, function, shareware, shrink-wrapped_software, software_documentati
on, documentation, spyware, supervisory_software, upgrade

Meaning 0 NLTK ID:  application.n.01
Hypernyms: use, usage, utilization, utilisation, employment, exercise
Hyponyms: misapplication, technology, engineering

Meaning 1 NLTK ID:  application.n.02
Hypernyms: request, petition, postulation
Hyponyms: credit_application, job_application, loan_application, patent_application
```

Fig. 6. Extracted Hypernyms for the Statistically Stuffed Aspects in Model1.

Table I and Fig. 7 shows the performance analysis comparison of model 1 when validated using machine learning algorithms in terms of accuracy, precision, recall and f1-score. From the analysis it shows that decision tree algorithm derived better accuracy of 73% among all the other compared machine learning algorithms.

TABLE I. PERFORMANCE ANALYSIS OF MODEL 1 USING MACHINE LEARNING ALGORITHMS

Machine Learning Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Support Vector Machine	65	65	65	65
Naïve Bayes	65	56	65	55
Logistic Regression	65	63	65	75
Decision Tree	**73**	**72**	**73**	**71**
K-Nearest Neighbor	67	44	67	53
Random Forest	69	68	69	68

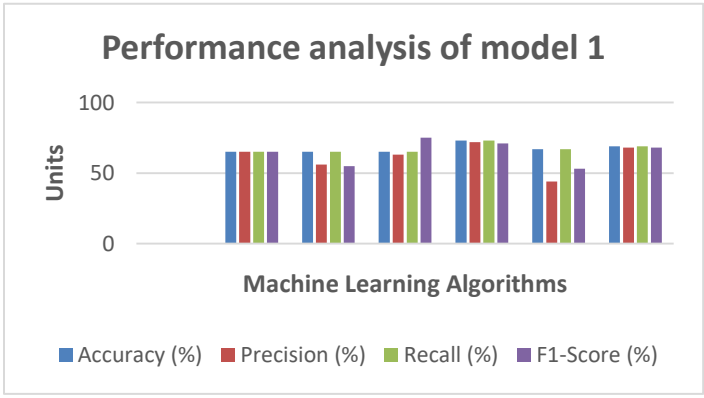


Fig. 7. Performance Analysis of Model 1 using Machine Learning Algorithms.

Fig. 8 presents the list of aspects extracted in model 2 based on probability using python libraries like gensim and ldamallet.

Fig. 9 presents the extracted aspect sentiment classification and plot in Fig. 10 presents the top most salient features extracted using model 2.

```
[(15, [('time', 0.12234042553191489), ('back', 0.0851063829787234), ('angle', 0.06914893617021277),
rocessor', 0.047872340425531915), ('single', 0.03723404255319149), ('bit', 0.03723404255319149), ('
ung', 0.02307692307692231), ('full', 0.06666666666666667), ('run', 0.06153846153846154), ('amole',
1025641025641026), ('notification', 0.041025641025641026), ('panel', 0.041025641025641026), ('hd',
1610486891385768), ('range', 0.11235955056179775), ('mode', 0.08239700374531835), ('detail', 0.0636
247191), ('update', 0.0299625468164794), ('hdr', 0.026217228464419477), ('light', 0.026217228464419
3659), ('photo', 0.07317073170731707), ('phone', 0.04065040650406504), ('performance', 0.036585365
('question', 0.032520325203252036), ('shoot', 0.032520325203252036), ('comparison', 0.032520325203
1527), ('make', 0.10843373493975904), ('bit', 0.08032128514056225), ('picture', 0.06827309236947791
o', 0.040160642570281124), ('portrait', 0.040160642570281124), ('expect', 0.040160642570281124), ('
y', 0.1188118811881188), ('quality', 0.07425742574257425), ('test', 0.0594059405940594), ('speaker'
5445544554455), ('pro', 0.03960396039604), ('lag', 0.034653465346534656), ('user', 0.024752475
6108594), ('series', 0.04524886877828054), ('offer', 0.04524886877828054), ('year', 0.040723981900
('line', 0.027149321266968326), ('network', 0.027149321266968326), ('cheap', 0.02262443438914027)])
('shot', 0.10989010989010989), ('ultra', 0.06227106227106227), ('focus', 0.047619047619047616), ('b
2197802197802198), ('turn', 0.02197802197802198), ('noise', 0.018315018315018316)]), (1, [('wide',
3506493), ('primary', 0.06060606060606061), ('side', 0.05627705627705628), ('great', 0.051948051948
('manage', 0.025974025974025976), ('switch', 0.025974025974025976)]), (17, [('thing', 0.14794520547
('work', 0.05205479452054795), ('guy', 0.05205479452054795), ('feel', 0.038356164383561646), ('he
136986301369864), ('mid', 0.024657534246575342)])]
```

Fig. 8. List of Aspects Extracted based on Probability from Model 2.

Dominant_Topic	Topic_Perc_Contrib	Keywords	Text	sentiment_terms	pol	sentiment
0.0	0.5640	phone, video, display, price, time, pretty, gu...	okay so what are the questions that I get ask...	ask awful good vary late break well launch pla...	0.9961	positive
1.0	0.6080	good, camera, battery, thing, game, screen, ta...	so after the galaxy m31 the m30 s we now have ...	s come adaptive fast charge case minimal bare...	0.9993	positive
1.0	0.8563	good, camera, battery, thing, game, screen, ta...	hey guys did saga from tech works and in this...	go detailed samsung current get take usual req...	0.9260	positive
0.0	0.6546	phone, video, display, price, time, pretty, gu...	hey guys this is one from guiding dick and to...	guide review recall watch check long think rig...	0.9882	positive
1.0	0.6116	good, camera, battery, thing, game, screen, ta...	some sense m-series our phones have always bee...	great amole create average bring let find know...	0.9986	positive
0.0	0.7038	phone, video, display, price, time, pretty, gu...	what's going on guys my name is Wade with tech...	go samsung vast base experience strike awesome...	0.9993	positive
1.0	0.8204	good, camera, battery, thing, game, screen, ta...	hi guys this is Ranji them in this video let's...	let test share feel divide quick big come mass...	0.9986	positive
1.0	0.6218	good, camera, battery, thing, game, screen, ta...	so the galaxy m31 arrived in India recently a...	arrive certain game compare buy galaxy let ans...	0.9991	positive
0.0	0.5300	phone, video, display, price, time, pretty, gu...	Samsoni's MCV smartphones have been quite popul...	popular different desirable hype mid range die...	0.9990	positive
0.0	0.6112	phone, video, display, price, time, pretty, gu...	you guys welcome to I again today we're going...	welcome go review new samsung launch start gre...	0.9990	positive
1.0	0.5094	good, camera, battery, thing, game, screen, ta...	Phone bckup is not good	good	0.4404	negative

Fig. 9. Aspect based Sentiment Classification in Model 2.

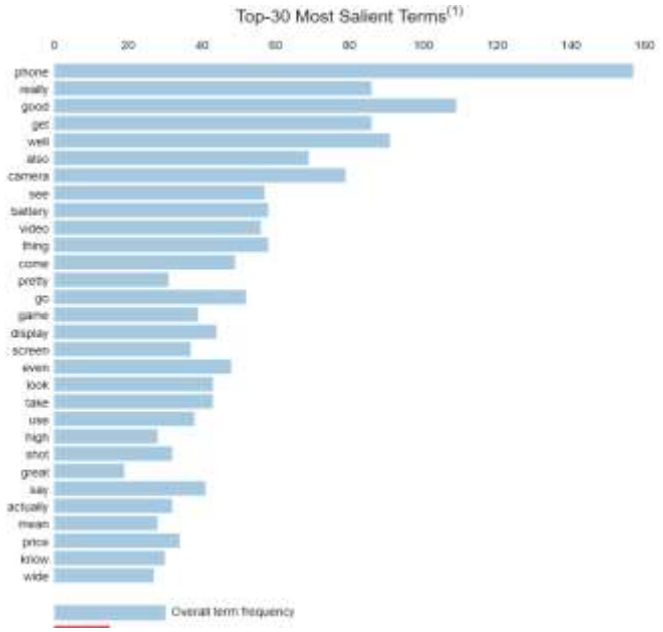


Fig. 10. Most Salient Aspects Extracted from Model 2.

Table II and Fig. 11 shows the performance analysis comparison of model 2 when validated using machine learning algorithms in terms of accuracy, precision, recall and f1-score. From the analysis it shows that Random Forest algorithm derived better accuracy of 89% among all the other compared machine learning algorithms.

TABLE II. PERFORMANCE ANALYSIS OF MODEL 2 USING MACHINE LEARNING ALGORITHMS

Machine Learning Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Support Vector Machine	74	69	67	64
Naïve Bayes	50	83	50	50
Logistic Regression	74	72	67	65
Decision Tree	75	88	75	77
K-Nearest Neighbor	75	56	75	64
Random Forest	89	87	85	85

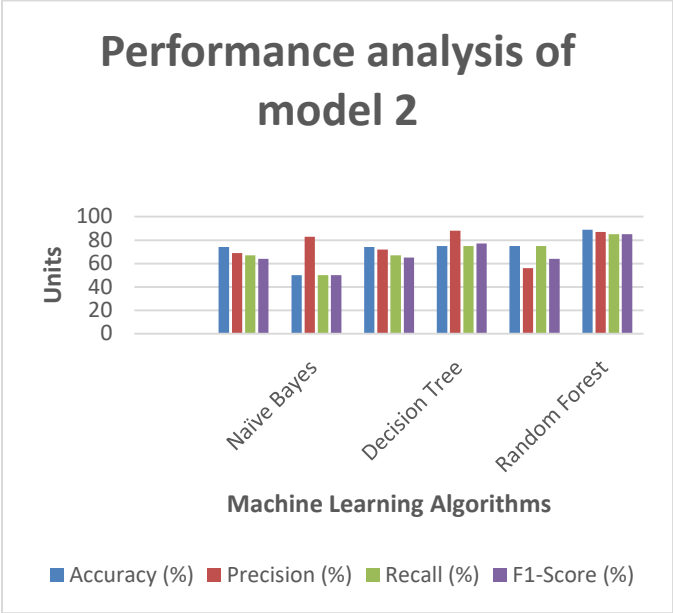


Fig. 11. Performance Analysis of Model 2 using Machine Learning Algorithms.

The Fig. 12 presents the way the aspects are extracted using POS tagging by dependency parsing mechanism and Fig. 13 presents the aspect-based sentiment analysis on the derived aspects using model 3.

Table III and Fig. 14 shows the performance analysis comparison of model 3 when validated using machine learning algorithms in terms of accuracy, precision, recall and f1-score. From the analysis it shows that Random Forest algorithm derived better accuracy of 95% among all the other compared machine learning algorithms.

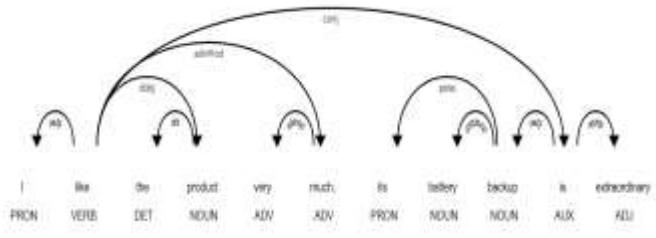


Fig. 12. Aspects Extracted from Model3.

Source_Name	message	Column2	aspect_terms	aspect_category	sentiment_terms	pol	sentiment
Samsung Galaxy M21 review _ Better than the Ga...	okay so what are the questions that I get ask...	100 and for the price this is the best Samsung	questions lot smartphone rupees answer time bi...	rupees smartphone n21 smudges buttons consumpt...	ask awful good vary late break well launch pla...	0.9981	positive
Samsung Galaxy M31 - Full Review, 64 MP Quad C...	so after the galaxy m31 the m30 s we now have ...	NaN	m30 m31 variants gigs ram storage box charger ...	phones case example smartphones issues sides g...	s come adaptive fast charge case minimal bare ...	0.9993	positive
Samsung Galaxy M31 Detailed Camera Review - En...	hey guys did saga from tech works and in this ...	000 rupees these have all been 16 megapixel im...	guys works video look cameras phone days silua...	guys samples sensor sensor sensor bones modes ...	go detailed samsung current get take usual req...	0.9260	positive
Samsung Galaxy M31 Full Review _ Better than R...	hey guys this is one from guiding dick and to...	NaN	guys guys video device link box device time mo...	guys device guys things body terms lights imag...	guide review recall watch check long think rig ...	0.9862	positive
Samsung Galaxy M31 Review - Decent Device, Dis...	some sense m-series our phones have always bee...	000 milliamp hours plus 15 watt fast charging ...	phones option displays life cameras m31 table ...	option levels flowers colors details booker ca...	great amoled create average bring let find know...	0.9985	positive
Samsung Galaxy M21 Review 2 Weeks Later! The B...	what's going on guys my name is Wade with tech...	NaN	guys name weeks majority youtubers phone peopl...	guys people m21 folks opinion support thing co...	go samsung vast base experience strike awesome ...	0.9993	positive
Samsung Galaxy M31 Review with Pros,txt	hi guys this is Ranji them in this video let's ...	000 anyways guys that's it for now for the rev...	guys video review guys device week experience ...	guys thing colors thing phones fingers device ...	let test share feel divide quick big come mass...	0.9886	positive
Samsung Galaxy M31 Review_ Should You Buy_ - E...	so the galaxy m31 arrived in India recently a ...	NaN	questions phone performance cameras review que...	design colors movies phones details photos thi...	arrive certain game compare buy galaxy let ans...	0.9991	positive
Samsung Galaxy M31 review_ Worth all the Hype_ ...	Samson's MCV smartphones have been quite popul...	NaN	smartphones segments centric qualities phone y...	smartphones qualities guys guys list changes m...	popular different desirable hype mid range die...	0.9990	positive
Samsung Galaxy M21 The Only Review You Need - ...	you guys welcome to I again today we're going ...	NaN	guys price features phone rupees line sale 5th...	price set specifications display display displ...	welcome go review new samsung launch start gre...	0.9990	positive
Samsung Galaxy M31 coder review	Phone bckup is not good	NaN	bckup		good	0.4404	negative
Samsung Galaxy M31 coder friend review	sleepy phone	NaN	phone	phone	sleepy	0.0000	negative

Fig. 13. Final Output of Aspect-based Sentiment Analysis.

TABLE III. PERFORMANCE ANALYSIS OF MODEL 2 USING MACHINE LEARNING ALGORITHMS

Machine Learning Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Support Vector Machine	75	72	69	65
Naïve Bayes	50	83	67	64
Logistic Regression	74	74	72	72
Decision Tree	50	77	65	72
K-Nearest Neighbor	72	64	62	60
Random Forest	95	92	90	90

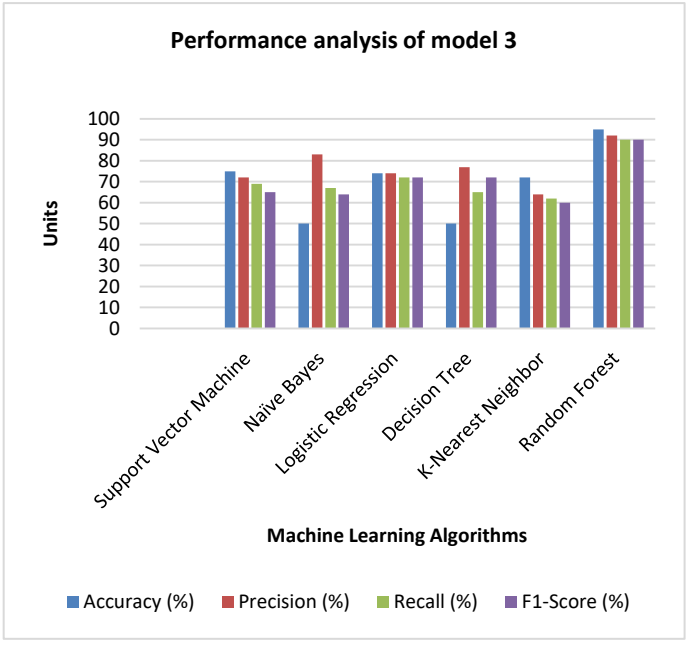


Fig. 14. Performance Analysis of Model 3 using Machine Learning Algorithms.

VII. CONCLUSION

The motivation to achieve fine-grained aspect-based sentiment analysis made us to propose an efficient hybrid aspect-based sentiment analysis model by the fusion of speech and text aspects. To enhance the performance of traditional ASR models, a deep learning framework and a bi-gram language model is employed for deriving the speech-to-text aspects. The results are made a comparison with the traditional ASR models and the proposed model achieved 5.7% WER and 3% CER [17]. Three variants of text feature extraction techniques were employed for improving the efficiency of ABSA. A decision level fusion was performed on the aspects extracted to enhance the sentiment in an efficient way. Feature level fusion was applied by proposing a NWAE mechanism as a feature selection measure to overcome the problem of dimensionality. The obtained results obtained from speech and three variants of text analysis models are compared with the individual text feature extraction techniques [18] [19] [20] and proved that the proposed hybrid level fusion mechanism improves the user readability by faster access and there by improves the performance.

REFERENCES

[1] M. Syamala, and N.J.Nalini, "A Deep Analysis on Aspect based Sentiment Text Classification Approaches," International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, No.5, pp.1795-1801 ,September - October 2019.

[2] Md. E. Mowlaei, Md. S. Abadeh, and H. Keshavarz, "Aspect-based sentiment analysis using adaptive aspect-based lexicons," Expert Systems with Applications, vol. 148, pp.1-27 ,15 June 2020.

[3] O. Alqaryouti, N. Siyam, A.A. Monem, and K. Shaalan, "Aspect-Based Sentiment Analysis Using Smart Government Review Data," Applied Computing and Informatics, pp.1-13, November 2019.

[4] V.S. Anoop and S. Asharaf, "Aspect-Oriented Sentiment Analysis: A Topic Modeling-Powered Approach," J. Intell. Syst., vol. 29(1), pp.1166–1178, December 2018.

[5] M. Shams, N. Khoshavi, and A. Baraani-Dastjerdi, "LISA: Language-Independent Method for Aspect-Based Sentiment Analysis," IEEE Access, vol.8, pp. 31034-31044, February 2020.

[6] D. Griol, J. M. Molina, and Z. Callejas, "combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances," Neurocomputing, vol. 326-327 ,pp. 132-140, January 2019.

[7] Z. Lu, L. Cao, Y. Zhang, Ch.Ch. Chiu, and James Fan, "Speech Sentiment Analysis Via Pre-Trained Features from End-To-End ASR Models," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, Spain, pp. 7149-7153, May 2020.

[8] B. Li, D. Dimitriadis, and A. Stolcke, "Acoustic and Lexical Sentiment Analysis for Customer Service Calls," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, United Kingdom, pp. 5876, 5880, May 2019.

[9] D. Zhang, S. Li, Q. Zhu, and G. Zhou, "Effective Sentiment-relevant Word Selection for Multi-modal Sentiment Analysis in Spoken Language," Proceedings of the 27th ACM International Conference on Multimedia, pp.148–156, October 2019.

[10] S. Maghilnan, and M. Rajesh Kumar, "Sentiment analysis on speaker specific speech data," International Conference on Intelligent Computing and Control (I2C2), Coimbatore, India, pp. 1-5, February 2018.

[11] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis In Data Mining", IEEE 16th International Conference on Data Mining (ICDM) , pp. 439–448, December 2016.

[12] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun C, K. Sagae, and LP. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," IEEE Intelligent Systems, vol. 28(3), pp. 46–53, March 2013.

[13] G. Yue, Y. Kangning, F. Shiyu, Ch. Shuhong, L. Xinyu and M. Ivan, "Hybrid Attention based Multimodal Network for Spoken Language Classification," Proceedings of the 27th International Conference on Computational Linguistics, pp. 2379–2390, August 20-26, 2018.

[14] S. Govindaraj and K. Gopalakrishnan, "Intensified Sentiment Analysis of Customer Product Reviews Using Acoustic and Textual Features," ETRI Journal, vol. 38 (3), pp. 494-501, 2016.

[15] L. Kaushik, A. Sangwan, and J. H. L. Hansen,"Sentiment extraction from natural audio streams," In Proceedings of International Conference on Acoustics, Speech and Signal Processing, pp. 8485-8489, 2013.

[16] H.H. Do, P.W.C. Prasad, A. Maag, and A. Alsadoon, "Deep learning for aspect-based sentiment analysis: a comparative review", Expert Systems with Applications, vol.118, pp.272-299, 2019.

[17] M. Syamala, and N.J. Nalini, "A Speech-based Sentiment Analysis using Combined Deep Learning and Language Model on Real-Time Product Review", International Journal of Engineering Trends and Technology, vol. 69 (1), pp. 172-178, January 2021.

[18] M. Syamala, and N.J. Nalini, "A filter-based improved decision tree sentiment classification model for real-time amazon product review data," International Journal of Intelligent Engineering and Systems, vol.13(1), pp. 191-202, January2020.

[19] M.Syamala, and N.J.Nalini, "LDA and Deep Learning: A Combined Approach for Feature Extraction and Sentiment Analysis," 10th ICCCNT, vol. 45670, pp. 1-5, December 2019.

[20] M. Syamala, and N.J. Nalini, "ABSA: Computational Measurement Analysis Approach for Prognosticated Aspect Extraction System", TEM JOURNAL - Technology, Education, vol. 10(1), pp. 82–94, February 2021.

Evaluating Chinese Potential e-Commerce Websites based on Analytic Hierarchy Process

Hemn Barzan Abdalla, Liwei Wang
Department of Computer Science, College of Science and Technology
Wenzhou-Kean University, China

***Abstract*—China has recently become the largest market for e-commerce at the global level. After the technological revolution and its widespread, As of December 2020, about 782.41 million people deal with this e-commerce through modern and advanced electronic devices, which are smartphones, computers, and others. The study of e-commerce is one of the branches of business administration established electronically through the use of Internet networks, which aim to carry out buying and selling operations. This article applied the Analytic hierarchy process (AHP) to evaluate three potential Chinese e-Commerce websites: JD, TAOBAO, and SUNING. We divided our model into three primary levels: Goal level, Strategical level, and criteria level. We take two main factors into account at the Strategical level: website (A1) and user (A2). Meanwhile, in the criteria level, we consider the total of six aspects: Visiting speed (B1), website stability (B2), page ranking (B3), average person visits (B4), period of average person visits (B5), and users' comments (B6). We also make all scores normalized; all scores are mapped into 0-1 to compare each website's performance. Our results show the TAOBAO is the best E-commerce website with a score of 0.8233 based on our algorithm, and JD is the second one with a score of 0.7895, while SUNING is the worst with a score of 0.5955.**

***Keywords*—Analytic hierarchy process (AHP); e-commerce; chinese e-commerce websites; JD (JING DONG); taobao; suning**

I. Introduction

E-commerce is indubitably a growing up industry around the world. Nowadays, E-commerce is regarded as a necessary part, and a cutting-edge shopping mode around the world [1]. with the development of the surge of E-commerce, modern society is undergoing a considerable change. From the Perspective of the e-commerce platform, there are mainly several business models: B2B, B2C, B2E, B2G, B2M, C2B, C2C, G2B, G2C, G2E, G2G, and P2P [2]. Meanwhile, various E-commerce platforms and related websites are showing up. The competition among all E-commerce is so radical that some websites will be eliminated through selection or competition. In China, the number of Internet users and the logistics and express industry is overgrowing, e-commerce enterprises are in fierce competition, and the platform situation is taking shape. In the C2C field, with the entry of search engine giant Baidu, online shopping users have more choices, and the industry competition is more intense. In this period, the pace of survival of the fittest in the e-commerce industry was accelerated, and the innovation of mode, product, and service appeared constantly [3].

With the growing maturity of online shopping, shopping demand has become diversified. In addition to continuously expanding categories, optimizing logistics, and after-sales service, manufacturers are increasingly developing their platform webpages. To compare and pick up high-level E-commerce, a good and robust evaluation Model should be applied. There are different types of evaluation models based on different algorithms. This study employs the Analytic hierarchy process(AHP) to evaluate three potential Chinese E-commerce websites: JD, TAOBAO, and SUNING. AHP has been bored out excellent performance in a wide range of research fields. It completely relies on subjective evaluation to prioritize the scheme, needs less data, and takes a short time to make a decision. On the whole, AHP introduces quantitative analysis into the complex decision-making process and makes full use of the preference information given by decision-makers in pairwise comparison for analysis and decision support. It not only effectively absorbs the results of qualitative analysis but also gives full play to the advantages of quantitative analysis, so that the decision-making process is highly organized and scientific, it is especially suitable for decision analysis of social, economic system [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14] utilized the AHP on project management. In other related fields, [15] 's research shows the APH could play an essential role in vendor selection of a telecommunications system. They found that AHP can be used to improve the group decision-making of selecting suppliers to meet customer requirements. AHP was applied to maritime transportation [16], Earth dam site selection [17], agricultural land suitability evaluation with the GIS platform [18] as well.

This paper's structures are as follows: first, the related work regarding the application of AHP on E-commerce will be introduced and discussed; next, a certain AHP model and its matrix are built for this scenario, and corresponding criteria and their definitions are selected; meanwhile, their measurements are quantified by scores. Finally, the results are compared and discussed.

II. Related Work

There are already existed abundant researches regarding the application of AHP on E-commerce platform evaluations. [19] proposed a new evaluation model based on Analytical Hierarchy Process (AHP) and Dempster Shafer (DS) theory and applied it to study E-commerce websites with 27 variables. AHP is used to analyze the ranking problem's structure and determine the weights of the criteria for ranking B2C E-

commerce websites [20]. Fuzzy AHP is employed as a product selection service in e-commerce [21]. System Quality, Content Quality, Usages, Table I: values of B1 and B2 at each time, the figure without () means average visiting time(s) of page, while the figure in () means ping time(ms) at each time.

Trust, Customer Support, Online Customer Feedback, and Personalization were considered when using ME-OWA and Fuzzy AHP to evaluationB2C E-commerce Websites. [22] proposed DSS based on an AHP method tcomprising three structured multi-criteria and sub-criteria for selecting the best possible choice from a set of alternatives. [23] evaluated the quality of e-commerce services using Fuzzy AHP and TOPSIS. [24] established a complete shipping efficiency evaluation system for cross-border e-commerce logistics and applies it to the Qingdao Port of China for solid evidence Based on the Analytic Hierarchy Process (AHP).[25] evaluated performance measurement indicators in the E-commerce logistics system based on the extent fuzzy analytic hierarchy process (AHP).[26]

Applied entropy method to establish a model to determine the objective weight of each index, and calculated the distance and close-degree between the e-commerce website and the ideal point. In addition, [27] was conducted to evaluate and select the quality of e-commerce services using Fuzzy AHP. [28] established B2C e-commerce site evaluation system, whose Perspective starts from consumer based on AHP.

This paper's contributions are mainly introducing some new related variables, especially the periodical variables like the Period of Average Person Visits which may be an essential index when studying the performance of customers. The wavelet technique is also employed in this study, which can filter periodical patterns [29].

TABLE I. ANALYZE THE STRUCTURE OF RANKING PROBLEM

Try times	TaoBao	JD	Suning
1	0.243 (16.502)	0.461 (3.699)	0.231 (282.537)
2	0.222 (15.689)	0.348 (3.802)	0.145 (321.117)
3	0.226 (14.696)	0.204 (4.523)	0.246 (373.535)
4	1.169 (15.880)	0.213 (5.132)	0.387 (307.473)
5	0.215 (16.782)	0.167 (4.412)	0.195 (350.582)
6	0.246 (16.333)	0.193 (4.742)	0.131 (289.817)
7	0.488 (15.563)	0.188 (6.170)	0.197 (328.913)
8	0.678 (14.967)	0.278 (5.044)	0.232 (368.091)
9	0.207 (16.511)	0.161 (4.792)	0.277 (307.994)
10	0.266 (16.176)	0.159 (5.174)	0.275 (346.720)

III. MODULE

We follow the judging criteria proposed by [15]. The basic idea of AHP is a multi-criteria decision-making and evaluation method combining qualitative and quantitative analysis. The elements of decision-making are decomposed into different levels, and the advantages and disadvantages of the decision-making schemes are sorted through some judgments. On this basis, qualitative and quantitative analysis is carried out. It makes people's thinking process hierarchical and quantitative and uses mathematics to provide a quantitative basis for analysis, decision-making, evaluation, prediction, and control. Let $\{A_1, A_2..., A_m\}$ is the set of evaluating elements in each level. The pairwise comparison matrix is defined as.

$$T = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix} \quad (1)$$

The interpretations of each element and importance grading criteria are shown in Table IV. The process of computing weight may be divided into three main steps:

multiply all elements in each row of judge matrix T noted by M*im.*

The interpretations of each element and importance grading criteria are shown in Table IV. The process of computing weight may be divided into three main steps:

multiply all elements in each row of judge matrix *T* noted by *Mi.*

$$M_i = \prod_{j=1}^{m} a_{ij}, j = 1,2,3 \cdots m. \quad (2)$$

compute the value of mth root of M*i*:

$$\overline{W_i} = \sqrt[m]{M_i}, i = 1,2,3, \cdots, m \quad (3)$$

the weight W_i can be acquired by normalizing the vector $(W_1, W_2, \dots, W_m)$, which can be finished by:

$$W_i = \frac{\overline{W_i}}{\sum_{j=1}^{m} \overline{W}_j} \quad (4)$$

We select two elements: Website (A1) and User(A2) as strategical level (Level2) while Six elements: Visiting speed (B1), website stability(B2), page ranking (B3), average person visit (B4), period of average person visit (B5) and users' comments (B6) are taken into account at criteria (Level3). The main outflow is shown in Fig. 1. After constructing a pairwise comparison judgment matrix. A necessary step is to perform an inconsistency test for each comparison judgment matrix. We follow the spirit of the Consistency Index (CI) proposed [28]. CI is defined as follows:

$$CI = \frac{\lambda_{max} - n}{n - n} \quad (5)$$

where λ_{max} is the maximum eigenvalue of matrix T. Furthermore, consistency ratio (CR) can be computed by employing the following equation:

$$CR = \frac{CI}{RI} \quad (6)$$

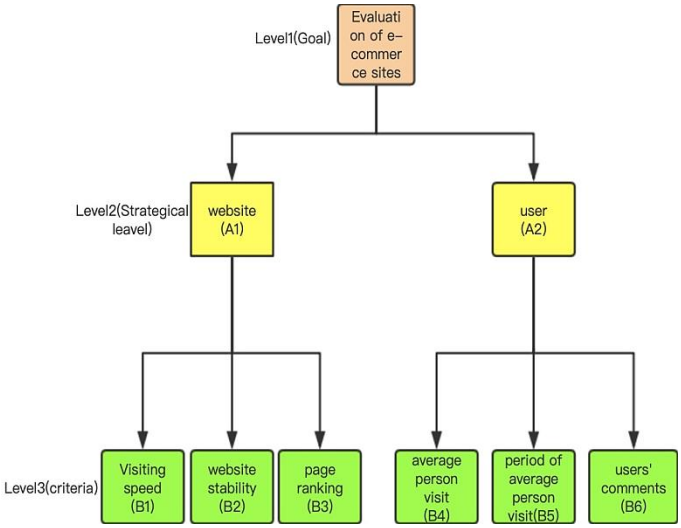


Fig. 1. Main Structure for AHP.

The RI is an index related to the dimension of matrix T. The value of RI is shown as follows:

Dimension	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46	1.49

(7)

After getting the value of CR, one may conclude that the pairwise comparison judgment matrix is desirable when CR < 0.1; otherwise, the pairwise comparison judgment matrix is unacceptable.

A. Visiting Speed (B1)

we use the same browser: Chrome: Mozilla/5.0. We employed the data from http://www.webkaka.com/WebCheck.aspx, where we set 10 times connected to a certain website and computed the average page-open time. Given that the faster visiting speed is, the higher grades should be acquired. Meanwhile, data normalization is necessary to eliminate the effect of different attributes' dimensions. Therefore, we set the final grade is calculated via the following formula:

$$Grade = 1 - \frac{x_{median} - min_{1\le i\le 30}\{x_j\}}{max_{1\le i\le 30}\{x_j\} - min_{1\le i\le 30}\{x_j\}} \quad (8)$$

where x_{media} is the median of each website's data. The result is shown in Table I.

B. Website Stability (B2)

we use Ping to make sure the network server is available because Ping IP is much faster than trying to connect to it through any number of protocols. This is also a good way to test Internet connection latency. Similarly, we try 10 times and record each response time. The data is normalized to [0,1] by applying Eq.1. The results are shown in Table I.

C. Page Ranking (B3)

It is referred to as PR (PageRank), from 0 to 10; the more significant the value, the more popular or important the page. In this study, we employ the google page rank algorithm(https://www.prchecker.info/checkpagerank.php). The data is mapped into [0,1] by di-viding10.

D. Average Person Visits (B4)

It is known person's visit could represent the popularity of a website. Therefore, we use the average page view (PV) from 1/10/2021-5/9/2021 to quantify the popularity, and the data is normalized to [0,1] by applying:

$$Grade = \frac{x_a - min_{1\le i\le 30}\{x_j\}}{max_{1\le i\le 30}\{x_j\} - min_{1\le i\le 30}\{x_j\}} \quad (9)$$

Where x_a is the average of each website's data.

E. Period of Average Person Visits (B5)

We use page view(PV) to represent the average person's visit. A page view is the number of page views; Every time a user visits each page in the website, it is recorded once. Users visit the same page many times, and the number of visits is accumulated. The PV data is available at https://www.prchecker.info/checkpagerank.phpInthisstudyweal sotak as grade to represent the frequency. So as to precisely reveal the potential period of PV, we utilize the Morlet wavelet as a method. The Morlet wavelet is a continuous plane wave [30] modulated by Gauss function based on the following mother function:

$$\Psi(t) = e^{ict}\left(e^{-\frac{t^2}{2}} - \sqrt{2}e^{-\frac{c^2}{4}}e^{-t^2}\right) \quad (10)$$

The formula of discrete wavelet transform is

$$W_f(a,b) = |a|^{-\frac{1}{2}}\sum_{i=1}^{N} f(i\delta t)\psi^*\left(\frac{i\delta t - b}{a}\right) \quad (11)$$

where * denotes complex conjugate, a is the scale factor (related to period and frequency), b is translation factor (time position), *i* is the time position label of data series, *f(t)* is variable time series, $W_{f(a,b)}$ is wavelet coefficient; δt is the difference of time series. The wavelet power spectrum is defined as:

$$E_{a,b} = |W_f(a,b)|^2 \quad (12)$$

The total wavelet power spectrum Ea represents the energy density corresponding to different scales a, which is defined as

$$E_a = \frac{1}{N}\sum_{b=1}^{N}\left|W_f(a,b)\right|^2 \quad (13)$$

Fig. 2 shows the results of Morlet wavelet analysis. There are two parts; the first is the contour of the real part of the wavelet coefficient, which has been mapped to color distribution, while the curve at the right region indicates the power spectrum of the wavelet coefficient as discussed in Eq.13. The power spectrum may help determine the main period in a certain time series; the peak in the wavelet power spectrum is regarded as the first/ main period discussed in this paper. In addition, the main period and its power have been marks as Y and X correspondingly. In general, all of (a), (b), (c) shows that cycle oscillations of a period of average person visit are very significant through the whole time domain.

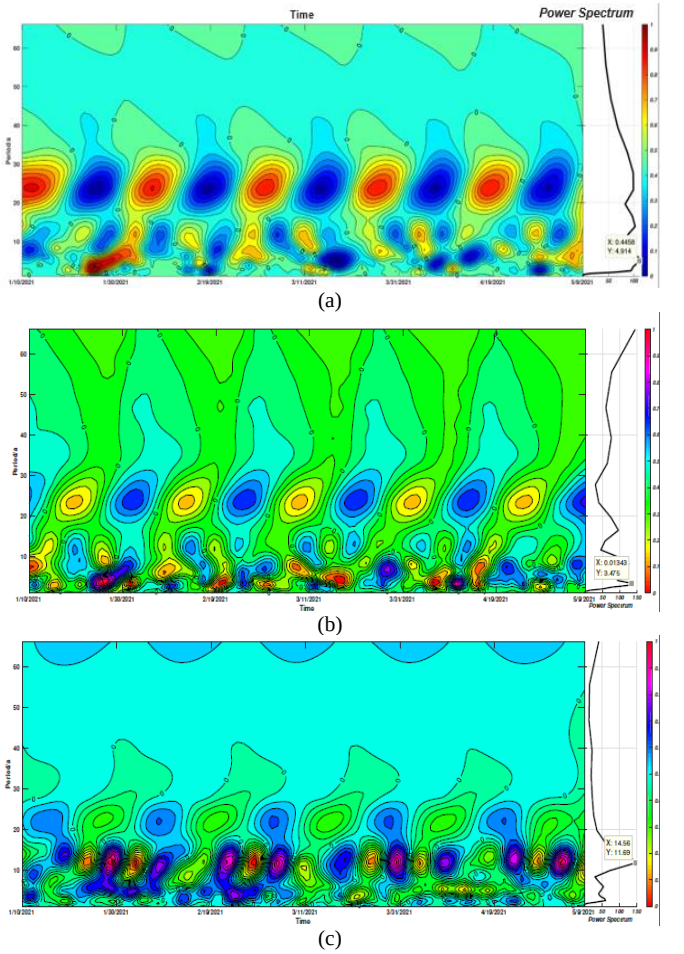
F. Users' Comments (B6)

The users' comments are also a vital index to evaluate an e-commerce website [31]. We choose the newest iPhone 12 product as our target, and then we search for this same iphone12 product on each website. 200 comments were

crawled from each website. Then we use the python package snownlp to grade each comment. The grade ranges from 0 to 1; if the grades are closer to 0, the comments are more negative and vice versa. The final grade is the average of those figures. The outflow of comments' sentimental analysis is shown in Fig. 3. The basic statistics of sentimental scores are shown in Table II.

TABLE II. BASIC STATISTICS OF SENTIMENTS MARKS FOR EACH WEBSITE

	SUNING	TAOBAO	JD
Mean	0.890912	0.896379	0.691769
Median	0.976914	0.986995	0.858268
Maximum	1	1	1
Minimum	0.011091	0.011038	0
Std. Dev.	0.187958	0.206043	0.352829
Skewness	-2.483815	-2.731886	-0.896659
Kurtosis	9.190327	10.16932	2.284826



(a)

(b)

(c)

Fig. 2. (a) Wavelet Analysis for the Period of Average Person visit JD (b) Wavelet Analysis for the Period of Average Person Visit TAOBAO (c) Wavelet Analysis for the Period of Average Person visit SUNING.

Fig. 4 shows the results of comments' sentimental analysis; firstly, the users' favorable rate is 72%. (b) and (c) are words cloud which may provide the potential reasons for the favorable rate. On the one hand, the majority of positive aspects concentrates on the following properties: High speed, nice appearance, high-quality photo; on the other hand, however, the positive comments mainly focus on the following aspects: the weak signal, the temperature of the phone is too hot, bad after-sale service, slow express.

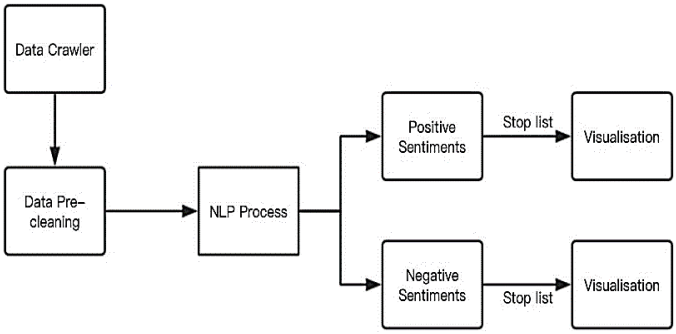
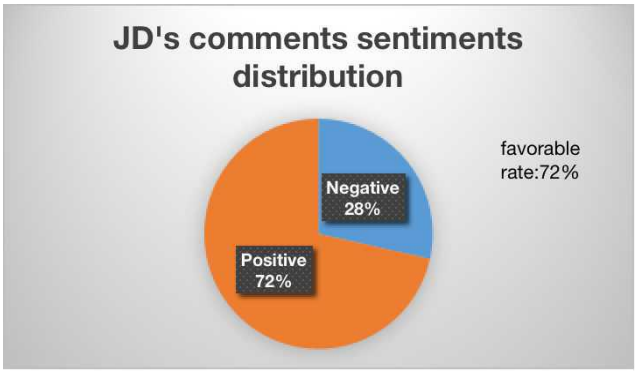
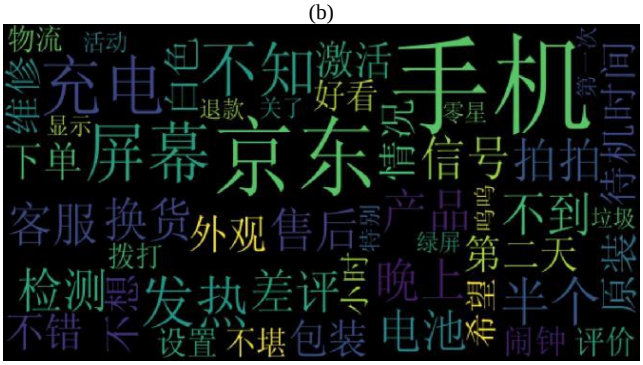


Fig. 3. Outflow of Comments' Sentimental Analysis.



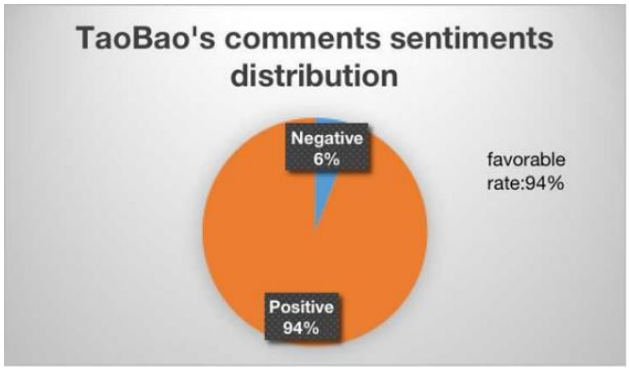
(a)

(b)



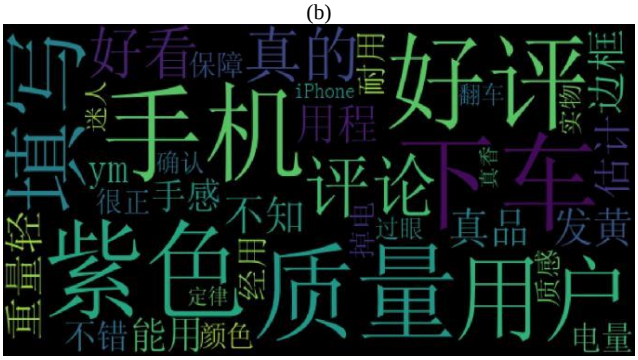
(c)

Fig. 4. (a) JD's Comments Sentiments Distribution (b) Positive Comments Words cloud of JD (c) Negative Comments Words cloud of JD.



(a)

(b)

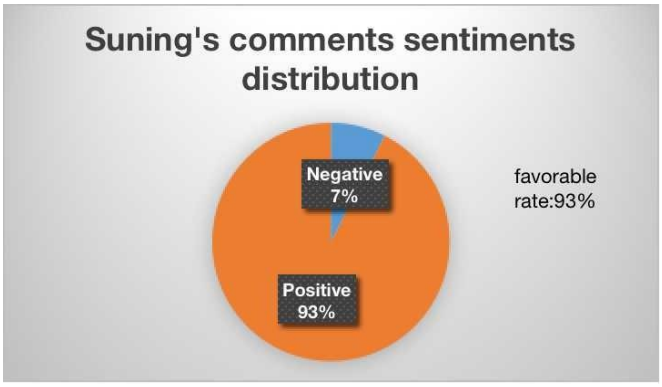


(c)

Fig. 5. (a) TAOBAO's Comments Sentiments Distribution (b) Positive Comments Words cloud of TAOBAO (c) Negative Comments Words cloud of TAOBAO.

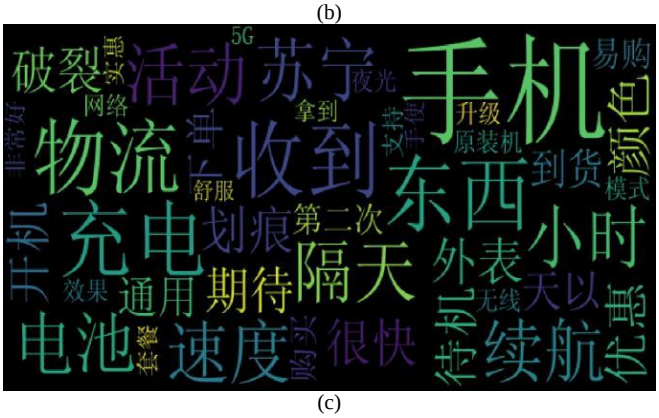
Fig. 5 shows the results of comments' sentimental analysis for TAOBAO; firstly, the users' favorable rate is 94% which is better than JD's. (b) and (c) are words cloud which may provide the potential reasons for the favorable rate. On the one hand, the majority of positive aspects concentrate on the following properties: long-endurance, good brand, high-quality photo; on the other hand, however, the positive comments mainly focus on the following aspects: bad texture, ugly appearance color.

Fig. 6 shows the results of comments' sentimental analysis for SUNING; firstly, the users' favorable rate is 93% which is better than JD's but worse than TAOBAO's. (b) and (c) are words cloud which may provide the potential reasons for the favorable rate. On the one hand, most positive aspects concentrate on the following properties: long endurance, high efficiency, good battery; on the other hand, however, the positive comments mainly focus on the following aspects: bad appearance, ugly appearance color, slow expression.



(a)

(b)



(c)

Fig. 6. (a) SUNING's Comments Sentiments Distribution (b) Positive Comments Words cloud of SUNING (c) Negative Comments Words cloud of SUNING.

IV. Concluding Remarks

We constructed the pairwise comparison judgment matrix after the consultation and survey based on related fields. We assume that the website(A1) and user(A2) are equally important; visiting speed(B1) is of equal importance compared with website stability(B2); visiting speed(B1) is moderately important over page ranking(B3); Visiting speed(B1) is slightly important than Page ranking(B3). On the other hand, average person visits(B4) is slightly important than the period of average person visit(B5); users' comments(B6) is moderately critical compared to average person visit(B5); users' comments(B6) is critical than the period of average person visit(B5). Grounded on these assumptions, the pair-wise comparison judgment matrix is built as follows:

$$T_{A_1} = \begin{pmatrix} A_1 & B_1 & B_2 & B_3 \\ B_1 & 1 & 1 & 3 \\ B_2 & 1 & 1 & 2 \\ B_3 & \frac{1}{3} & \frac{1}{2} & 1 \end{pmatrix}, T_{A_2} = \begin{pmatrix} A_2 & B_4 & B_5 & B_6 \\ B_4 & 1 & 2 & \frac{1}{3} \\ B_5 & \frac{1}{2} & 1 & \frac{1}{4} \\ B_6 & 3 & 4 & 1 \end{pmatrix} \quad (14)$$

The corresponding weight vectors: WA1 = [0.4434,0.3874,0.1692], WA2 = [0.2385,0.1365,0.6250].

The maximum eigenvalue is $\lambda_{A1}^{max} = 3.0183 = \lambda_{A2}^{max}, CI_{A1} = 0.0091 = CI_{A2}, CR_{A1}$ =0.0158 = CRA2. Notice that both CR are less than 0.1, which shows the consistency of each pairwise comparison judgment matrix. The Overall weight of each index can be easily calculated by multiplication: $W_{Bi} * W_{Aj}, i = 1..3, j = 1..2$. Table III shows the specific results. The total score of each e-commerce website is available in Table V. It is shown that TAOBAO is the best e-commerce website with a total score of 0.8233/1, while JD is the second one with a total score of 0.7895/1, and SUNING is the worst one compared with the TAOBAO and JD with a total score of 0.5955/1. Based on the results, some suggestions are proposed for SUNING: first, the website stability should be improved largely. Second, Advertising on other websites is an important way to attract visitors to the website. Advertising will enhance the brand awareness of the website. Compared with traditional advertising, online advertising has many advantages. First of all, online advertising is not limited by time and space; the use of the Internet can send ads to every corner, as long as there is a network that can also browse advertising information at any time. In the past, the traditional form of advertising is single, and it has strict requirements and constraints on the content and form of communication. The form of online advertising is diversified, and advertising can appear in many forms. Users can interact with each other by playing with advertising rather than just seeing it so that users can see more public information and experience more product services.

TABLE III. OVERALL JUDGMENT WEIGHT OF EACH INDEX

	Strategical Level	**Criteria**	**Overall Weight**
Evaluation of E-commerce websites	A1 website (0.5)	B1 Visiting speed (0.4434)	0.2217
		B2 Website stability (0.3874)	0.1937
		B3 Page ranking (0.1692)	0.0846
	A2 User (0.5)	B4 Average person visit (0.2385)	0.1193
		B5 period of average person visit (0.1365)	0.0683
		B6 Users' comments (0.6250)	0.3125

TABLE IV. THE CRITERION TABLE

The value of a_{ij}	**Interpretation**
1	equal importance
3	moderate importance of a*i* over a*j*
5	strong importance of a*i* over a*j*
7	very importance of a*i* over a*j*
9	extreme importance of a*i* over a*j*
2, 4, 6, 8	Intermediate values of 1,3,5,7
reciprocal	has reciprocal relations with respect to diagonally symmetric element : a*ij* · a*ji* = 1

TABLE V. THE TOTAL SCORE OF EACH E-COMMERCE WEBSITE

	TAOBAO	**JD**	**SUNING**
B1	0.8907	0.9350	0.9032
B2	0.9667	0.9971	0.1312
B3	0.7000	0.7000	0.7000
B4	0.6670	0.8361	0.2163
B5	0.2878	0.2035	0.0855
B6	0.8964	0.6918	0.8929
Total Score	0.8233	0.7895	0.5955

V. DISCUSSION

This work studied Chinese potential E-commerce Websites by applying AHP model. Three potential and competitive E-commerce websites platforms are compared. Our results reveal that TAOBAO is the most competitive with the highest score: 0.8233, JD with 07895 and SUNING with the lowest score:0.5955.

In the future, these works will be extended by following two aspects: First, other modern and effective evaluating models will be included and compared with current APH's results; Next, more related variables and hierarchies will be considered to develop a more completed E-commerce evaluating model.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support from with Wenzhou-Kean University and Leading Talents of Provincial Colleges and Universities, Zhejiang-China(#WB20200915000043).

REFERENCES

[1] Rania Nemat. Taking a look at different types of e-commerce. World Applied Programming, 1(2):100–104, 2011.

[2] Abdul Moktadir, Towfique Rahman, Charbel Jose Chiappetta Jabbour, Syed Mithun Ali, and Golam Kabir. Prioritiza-tion of drivers of corporate social responsibility in the footwear industry in an emerging economy: A fuzzy ahp ap-proach. Journal of cleaner production, 201:369–381, 2018.

[3] Hemn Barzan Abdalla, Lu Zhen and Zhang Yuantu, "A New Approach of e-Commerce Web Design for Accessibility based on Game Accessibility in Chinese Market" International Journal of Advanced Computer Science and Applications(IJACSA), 12(8), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120801.

[4] Maggie CY Tam and VM Rao Tummala. An application of the ahp in vendor selection of a telecommunications system. Omega, 29(2):171–182, 2001.

[5] Ali Emrouznejad and Marianna Marra. The state of the art development of ahp (1979–2017): a literature review with a social network analysis. International Journal of Production Research, 55(22):6653–6675, 2017.

[6] Seleshi G Yalew, Ann van Griensven, Marlous L Mul, and Pieter van der Zaag. Land suitability analysis for agriculture in the abbay basin using remote sensing, gis and ahp techniques. Modeling Earth Systems and Environment, 2(2):1–14, 2016.

[7] Hong Wei Wang, Yu Song Yan, and Hua Wei Duan. Measuring the performance of e-commerce logistics system based on fuzzy-ahp. In Applied Mechanics and Materials, volume 505, pages 898–901. Trans Tech Publ, 2014.

[8] Rania Nemat. Taking a look at different types of e-commerce. World Applied Programming, 1(2):100–104, 2011.

[9] Blanca Perez-Gladish and Bouchra M'Zali. An ahp-based approach to mutual funds' social performance measurement. International Journal of Multicriteria Decision Making, 1(1):103–127, 2010.

[10] GUO Kai. The competitiveness evaluation of e-commerce website based on ahp-entropy. In 2010 International Confer-ence on E-Business and EGovernment, pages 392–394. IEEE, 2010.

[11] Hai-Min Lyu, Yong-Xia Wu, Jack Shuilong Shen, and An-Nan Zhou. Assessment of social-economic risk of chinese dual land use system using fuzzy ahp. Sustainability, 10(7):2451, 2018.

[12] M´onica Garc´ıa-Mel´on, Blanca P´erez-Gladish, Tom´as G´omez-Navarro, and Paz Mendez-Rodriguez. Assessing mu-tual funds' corporate social responsibility: a multistakeholder-ahp based methodology. Annals of Operations Research, 244(2):475–503, 2016.

[13] Francesca Abastante, Salvatore Corrente, Salvatore Greco, Alessio Ishizaka, and Isabella M Lami. Choice architecture for architecture choices: Evaluating social housing initiatives putting together a parsimonious ahp methodology and the choquet integral. Land Use Policy, 78:748–762, 2018.

[14] Kamal M Al-Subhi Al-Harbi. Application of the ahp in project management. International journal of project manage-ment, 19(1):19–27, 2001.

[15] Thomas L Satty et al. The analytic hierarchy process, 1980.

[16] Aminuddin Md Arof. The application of a combined delphi-ahp method in maritime transport research-a review. Asian Social Science, 11(23):73, 2015.

[17] Gang Chen. Analyzing criteria and sub-criteria for the corporate social responsibility-based supplier selection process using ahp. The International Journal of Advanced Manufacturing Technology, 68(1-4):907–916, 2013.

[18] Xu, L, Kumar, DT, Shankar, KM, Kannan, D & Chen, G 2013, 'Analyzing criteria and sub-criteria for the corporate social responsibility-based supplier selection process using AHP', International Journal of Advanced Manufacturing Technology, vol. 68, no. 1-4, pp. 907-916. https://doi.org/10.1007/s00170-013-4952-7.

[19] Jia Yu, Jianrong Yao, and Yuangao Chen. Credit scoring with ahp and fuzzy comprehensive evaluation based on be-havioural data from weibo platform. Tehni˘cki vjesnik, 26(2):462–470, 2019.

[20] Xiong Ying, Guang-Ming Zeng, Gui-Qiu Chen, Lin Tang, Ke-Lin Wang, and Dao-You Huang. Combining ahp with gis in synthetic evaluation of eco-environment quality—a case study of hunan province, china. Ecological modelling, 209(2-4):97–109, 2007.

[21] Deng-Neng Chen, Chih-Wei Tseng, and Chia-Yi Lin. Applying fuzzy ahp on product selection service in e-commerce. In 2011 International Joint Conference on Service Sciences, pages 198–202. IEEE, 2011.

[22] Fereshteh Alizadeh and Mina Lahiji. Suitable delivery system in small ecommerce companies. Journal of Humanities Insights, 2(04):167–171, 2018.

[23] A Ishak, R Ginting, and W Wanli. Evaluation of e-commerce services quality using fuzzy ahp and topsis. In IOP Con-ference Series: MaterialsScience and Engineering, volume 1041, page 012042. IOP Publishing, 2021.

[24] Xuhua Chen. Marine transport efficiency evaluation of cross-border ecommerce logistics based on analytic hierarchy process. Journal of Coastal Research, 94(SI):682–686, 2019.

[25] Hong Wei Wang, Yu Song Yan, and Hua Wei Duan. Measuring the performance of e-commerce logistics system based on fuzzy-ahp. In Applied Mechanics and Materials, volume 505, pages 898–901. Trans Tech Publ, 2014.

[26] GUO Kai. The competitiveness evaluation of e-commerce website based on ahp-entropy. In 2010 International Confer-ence on E-Business and EGovernment, pages 392–394. IEEE, 2010.

[27] Aulia Ishak et al. Evaluation and selection of e-commerce service quality using fuzzy ahp method. In IOP Conference Series: Materials Science and Engineering, volume 1003, page 012152. IOP Publishing, 2020.

[28] Panova, Y., Tan, A., Hilmola, OP. et al. Evaluation of e-commerce location and entry to China – implications on shipping and trade. J. shipp. trd. 4, 6 (2019). https://doi.org/10.1186/s41072-019-0045-6.

[29] Christopher Torrence and Gilbert P Compo. A practical guide to wavelet analysis. Bulletin of the American Meteoro-logical society, 79(1):61–78, 1998.

[30] Thomas L Saaty. How to make a decision: the analytic hierarchy process. European journal of operational research, 48(1):9–26, 1990.

[31] Hong Wei Wang, Yu Song Yan, and Hua Wei Duan. Measuring the performance of e-commerce logistics system based on fuzzy-ahp. In Applied Mechanics and Materials, volume 505, pages 898–901. Trans Tech Publ, 2014.

Detection Technique and Mitigation Against a Phishing Attack

Haytham Tarek Mohammed Fetooh[1]
Information Security Prog. Faculty of Computers and Information
Mansoura University, Mansoura, Egypt

M. M. EL-GAYAR[2]
Dept. of Information Technology, Faculty of Computers, and Information, Mansoura University, Mansoura, Egypt

A. Aboelfetouh[3]
Dept. of Information Systems, Faculty of Computers and Information, Mansoura University, Mansoura, Egypt

***Abstract*—Wireless networking is a main part of our daily life during these days, each one wants to be connected. Nevertheless, the massive progress in the Wi-Fi trends and technologies leads most people to give no attention to the security issues. Also detecting a fake access point is a hard security issue over the wireless network. All the currently used methods are either in need of hardware installation, changing the protocol or needs analyzing frames. Moreover, these solutions mainly focus on a single digital attack identification. In this paper, we proposed an admin side way of detection of a not real access point. That works on multiple cyber-attacks especially the phishing attack. We shed the light on detecting WI-phishing or Evil Twin, DE authentication attack, KARMA attack, advanced WI-phishing attack and differentiate them from the normal packets. By performing the frame type analysis in real time and analyzing different static and dynamic parameters as any change in the static features will be considered as an evil twin attack. Also, providing that the value of the dynamic parameters surpasses the threshold, it reflects Evil Twin. The detector has been tested experimentally and it reflects average accuracy of 94.40%, 87.08% average precision and an average specificity of 96.39% for the five types of attack.**

***Keywords*—Rogue access point; phishing attacks; KARMA attack; social engineering; hacking**

I. Introduction and Background

Currently, the wireless techniques help users who are using terminals, phones, and tablets to use the internet services, in addition to being integrated in many interfaces and used implementation over the field of (IOT) Internet of Things. [1] Despite the growth of wi-fi technologies, users still do not care for security issues. As clients used to be online most of the time, this gives a higher availability of being victimized with many of the cyber security attacks. All these communications are done over the channel used for sending or receiving wireless waves in-between the access point (AP) and the user. Because of that, the attacker is in no need to physically access the victim's network. He can easily sniff, eavesdrop, resend frames using off the shelf tools [2]. While getting benefits from this technology, these vast numbers of non-smart connected cyber-physical devices have several properties that led to critical security issues, such as nodes mobility, wireless communications, lack of local security features, scale, and diversity. IoT botnets attack is considered a critical attack that affects IoT network infrastructure that launches a distributed denial of service (DDoS) [3].

Phishing is the attempt to acquire sensitive data or to inspire somebody to react in a desired method by simulating as a trusted one in the electronic atmosphere. As demanding a user to tap on a connection in an email or to give his Mastercard numbers or enter definite data as first, last name, address, age, and city. At that point, the hacker can access and use the data. Phishing assaults can be performed over different specialized strategies [2].

A. Impact on the Community and Motivation

Damage from cybercrime is expected to cost the world $6 trillion annually by 2021, raised from $3 trillion in 2015 according to Cybersecurity Ventures [3]. Phishing attacks are the most common type of cybersecurity breaches as stated by the official statistics from the cybersecurity breaches survey 2020 in the United Kingdom [4].

As Phishing attacks merge social psychology, technical systems, and security subjects. These attacks affect organizations and individuals alike, the loss for the organizations is significant as it includes the recovery cost, reputation loss, and productivity reduction [5].

According to a study named Proofpoint 2020, [6] around 90% of organizations suffered targeted phishing attacks during 2019. From which 88% experienced spear-phishing attacks, 83% faced voice phishing (Vishing), 86% dealt with social media attacks, 84% reported SMS/text phishing (SMishing), and 81% reported malicious USB drops.

According to the Phishing Activity Trends Report 1st Quarter 2021 [7] from the Anti-Phishing Working Group , APWG's records, with an unprecedented 245,771 attacks in one month which confirms that phishing attacks are on the rise.

This proofs that phishing attacks are in continuous raise in recent years and have become more sophisticated and gained more attention from cyber researchers. So that, this paper aims to contribute to solving a type of phishing attack which serves

in solving the phishing problem and mitigate its impact over the community.

Therefore, the research problem is to address the limitation of the previous studies and security scheme that may offer attack detection but fails to offer it in real time over the network. The problem's solution arises whereas there is an increasing evolution of network devices as well as smart appliance for WI-FI services. Hence, this acts as motivating factor towards working for improving the security of networks and connections as addressed in this research.

The aim of this study is to improve the detection of the attack and contribute to solving the problem of the phishing attack by present a solution that is not costly and in real-time in addition to achieving the best performance with high accuracy and the decreasing the cost. The main aim is to reduce the spread of this attack, improving the detection rate, improving accuracy, decreasing the false alarms, and decreasing the cost of the proposed method.

Therefore, this study addresses the following research questions:

- Research Question 1 (RQ1) How to enable administrators to detect WI phishing in real time without using a special or expensive hardware?
- Research Question 2 (RQ2) How to find a reliable forensic way that visualize the different attacks underway?

It is worth noting that in this research WI-phishing is referred to as Evil Twin (ET) or Rouge Access Point (RAP). As WI-phishing [2] or Evil Twin, DE authentication attack [8] , and KARMA attack [9] are considered types of phishing, we focus on detecting these types. WI-phishing [2] or ET is a procedure of phishing that uses a wireless network where the phisher is in between the client and illegitimate wireless AP, using a Rouge access point as in Fig. 1.

WI phishing is one of the most dangerous and severing attacks [10] that deceives the user to join a rogue access point (RAP) instead of Legitimate Access Point (LAP), while RAP is a malicious device used by an adversary as if it is a real AP.

In which the intruder always copies the same configurations of one or more nearby LAPs to broadcast the same Service Set Identifier (SSID) and always with even stronger transmitting power.

The DE authentication attack is when the attacker tries to sniff or break the connection between the victim and an AP by flooding the network with DE authentication frames to force the client to reauthenticate. Then, the attacker can save traffic during the authentication process and this step is the base of attack's phase one. The attacker decrypts the pre-shared secret to have the secret key and bypass security encryption. The second step of the DE authentication attack is to force the client to connect to a RAP to sniff the whole communication which needs special tools to be detected. While RAP based on DE authentication is perhaps the most well-known assaults in Wi-Fi networks [11] as shown in Fig. 2.

Fig. 1. Wi-phishing / Evil Twin Attack Detection.

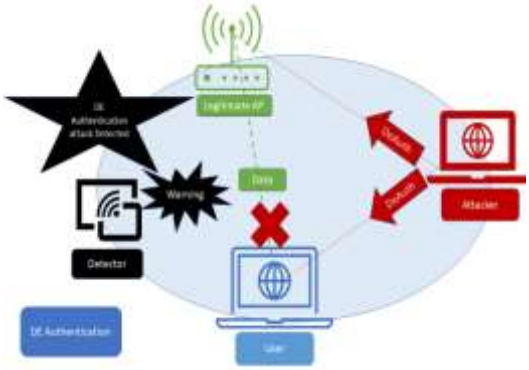


Fig. 2. De Authentication Attack Detection.

Karma attack depends on actively scanning the WLAN [12] to collect the probe frames requests from users' devices and then generate a corresponding probe response as if the required WLAN network is nearby. As the enormous growth of digital era more and more, many humans keep their own Wi-Fi settings in their device as it is for the device to automatically be connected to their known network if the network is in the nearby, devices send probe requests in probe frames to verify the existence of a network as the device does not know physically whether the network is in range or not. Resulting the targeted device is connected to the RAP which is made as a trap by the adversary.

Karma attacks can in any case influence customers that are utilizing active probing authentication. Also to perform it, the intruder can utilize a Pineapple AP 5 [13] that makes the assault a lot less difficult to achieve [9]. As shown in Fig. 3.

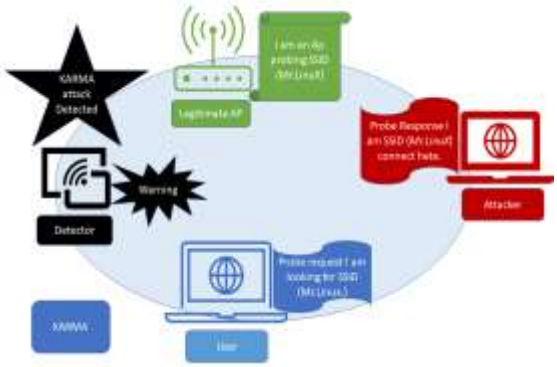


Fig. 3. KARMA Attack Detection.

B. Detecting Evil Twin Solutions

Detecting RAP solutions are categorized into two sets. The administrator-based set is a preliminary one that is based on observing the radio signals, and needs to be utilized on switches, servers, routers, or special devices. It may rely on the technique of whitelisting. From its name, it is deployed by the administrator of the network. But it has its downsides as most of it does not work in real time, needs protocol modification, or depends on a single point of failure as a server.

The second set of solutions is the client-based one. It is used by the user. Using the connection of TCP, Clock Skew, route option, IP packet header and data frame statistics are its main ways. But it has its con as it may needs a predefined data about the network [14].

For these reasons, such a study became more needed as it presents the following: Diverse detection logic to distinguish different kinds of attack signature. An empirical implementation of the proposed scheme for detecting WI-phishing [2]or ET, DE authentication attack , KARMA attack [9], advanced WI-phishing attack and the normal packets are prototyped under real attack. Recommending a better detection method by providing a real time admin side detection via specific parameters in the beacon frames as a signature of being under attack and then giving an alert. Furthermore, to send all beacon frames to a database for further processing and giving alert in case of any anomaly in the features of the previous packets. The final implementation is realized in the Python language using a Scapy [15] library to detect and classify frames.

Many studies have been published to address the problem of detecting phishing or WI Phishing attacks over the network, such as “Detecting fake access point or Evil Twin” as a relevant approach presented in [16] by Lovinger et al. , it depends on the network probing using Raspberry Pi 4. Capability, the author creates a logging system. Analyzing the wireless networks and scanning it, then filtering captured frames to create signatures and store it into the database. Upon the result, the alert is created after comparison. As for the challenges in this approach, it depends on a Raspberry Pi 4 as a device for detection which is a higher cost and limitations, while we developed a cross platform code with less cost. It depends on creating a unique fingerprint, he depends on a 256-bit hash function, which reflects more time, and he uses only static parameters while we depend on static and dynamic parameters for detecting the ET.

He depends on a log file which is not stable and harder to retrieve data rather than the visualized database used by our side. As a competitive advantage, we use all data from the database for further forensic purposes as it can store data for months which reflects stability and durability. Both of us work in real time detection. The other parts of the research are organized as follows: the related work is in Section 2; Section 3 describes the proposed method and the prototype. In Section 4, we present the experimental results. Then Section 5 contains the conclusion and the future work.

II. Related Work

A. Currently used Strategies for Detecting Evil Twin Attack are categorized in Five Groups as follows

1) Monitoring Wi-Fi traffic approach and limitation: Most wi-fi solutions attempt to spread sniffers over the network to accumulate fundamental data measurements such as MAC, SSID, working channel, RSS ID etc. The information accumulated deploying these sniffers enable the administrator to perceive the ET.

Kao et al. [17] detect the existence of ET by keeping a whitelist of the trusted MAC of Laps. Sniffers screen the remote traffic consistently, when an AP is discovered, whose MAC is not indexed in the whitelist, it is alerted as Evil Twin with accuracy of 96.4% average. As for the limitations of this approach, if the attacker sniffs the MAC address of the LAP, the approach has nothing to do in this case in addition of whitelisting only the MAC address disregarding other attributes that we have mentioned in our work.

Sachin et al. presented a way of identifying the malicious access point by setting up a whitelist of authentic APs' MAC address and IP. At that point send a broadcast packet over a central service to uncover the evil twin over the WLAN by receiving all replies from all access points and contrast them against the pre prepared whitelist. Yet, its downside is that it works over wireless networks when all terminals are in a similar range. It does not support the detection in real time as well as relying upon the whitelisting techniques [18] [19].

Sriram et al. [20]. and Chirumamilla et al [21] presented an agent-based intrusion detection system (IDS) to reveal RAP by screening the networks for the existence of new Aps and if these access points are not recorded in the pre-approved records, they will be flagged as RAP. As a downside in these methods, if an evil twin AP has a similar SSID and MAC, both these procedures are not as powerful as an insidious twin's MAC similar to the MAC of the approved AP. It depends on a server which is a single point of failure, and this method is useful only if the RAP is connected directly to the LAN and if the attacker has its own internet, the approach has nothing to do with it.

2) Timing based methods and feature extraction approach and limitation: In the frame analysis mechanism, the system aggregates all the frames using the mirror port of a core switch or by analyzing the frames obtained from the remote sensors distributed over the network. Utilizing the data from the gathered frames, many features are extracted to gain vital information regarding the existence of evil twin. The evil twin access point structures a scaffold between the real one and the customer to give Internet features. Due to crossing over an extra bounce, timing is put together and works with respect to the additional deferral happening because of the extra hop. This additional deferral gives proof to discovery of the evil twin.

Diogo M´onica et al. introduced an approach to distinguish a multi-hop evil twin via a real time detection device utilized by the client. It is not exposed to idleness or bandwidth with no need for a pre-prepared list. Using this way, channels are being scanned in less than 500ms. But it has constraints as a real access point can be distinguished as a fake one.it distinguishes evil twin in 30 seconds which is a major issue.[22].

Burns et al. in his paper [23] founded out the method of traceroute bidirectionally . To make a traceroute between a terminal and a remote server from terminal-to-server and then from server-back-to-terminal then to compare former both sides traceroute results. By comparing the number of hops both ways, if deference is found between the number of hops both ways that would be considered as an indicator of Evil Twin attack. The drawback of this method is that it has a single point of failure as it depends on an external server and this method is completely useless if the intruder uses his/her own internet connection like 4G.

Han et al. [24] have presented a technique that calculate the Round Trip Time (RTT) of a DNS query, while Mano et al. [25] utilizes a local RTT metric in addition to a frame payload slicing method to detect RAP too. As for the limitations of these solutions is that the RTT is affected by the congestion of the network. If there is a repeater in the network, this method is useless. In addition of depending on a server as it is an extra cost and considered as a single point of failure.

Yimin et al. [26] used the inter arrival time (IAT) to figure out the extra delay resulted by ET. Yet, the mentioned schemes fail when ET gives its own private connection causing the delay resulted because of the extra hop. It depends on training data to run the main algorithm of the detector and it is not easy for each admin to extract.

Fingerprinting a physical device remotely and passively in [27] is presented by F.Lanze et al. to mitigate the RAP but, it required a white listing with user interaction and a protocol modification for a spatial timestamp are utilized to mark beacon frames, resulting, the increase of false positive alarms probability as a result of the time synchronization problems, using only 50 observations for training, it detects RAP in 90% of all cases but it still depends on training data.

Based on the work of Kohno et al. in [27], Jana et al. [28] have presented the clock skew method to distinguish unapproved APs existence over the network. By calculating the clock skews of different APs using the IEEE 802.11-time synchronization function (TSF), as timestamps is a part of beacon frames. If the clock skew for a device does not equal the one kept previously, it is flagged as evil twin AP.

These solutions have many limitations as:

- It causes a higher load on the core switch because of the additional burden of feature processing.
- In case, an intruder uses his own Internet connectivity, the traffic doesn't arrive at the Centre switch, leaving the assault not detected.
- a spoofed response can be sent to the user to keep away from the time difference that may result from the extra hop.
- The approach results in a lot of false positives in case the frames are queued by a busy router which causes an additional delay.

3) Proprietary hardware approach and limitation: Pradip et al. [29] used a probing device that sends a pre-detection message to all connected users advising them to ignore the probe request. Afterwards, it sends another probe request and mark all responding APs as ET.

Eman et al. [30] used a chipset to detect the evil twin deauthentication attack depending on analyzing the packet frames especially the management frame named DE authentication frames.

These solutions have many limitations as:

- Ignoring the 802.11 probe request is a violation of the 802.11 standards [31].
- If attacker ignores reacting to the probe request to stay covered up and make, the method not useful.
- Special hardware means higher cost.

4) Signature-based and anomaly-based IDS's approach and limitation: It means using a database that contains the known intrusion patterns or signatures to be used for detection but in case of a new pattern or signature, it will never be detected. While an anomaly-based IDS creates profile for a host or network in the normal situation depending on statistics. It can recognize both known and unknown attacks. Both mentioned types lead to a large number of false positives. A survey of many anomaly based IDSs is mentioned in [32], It uses honeypot and anomaly analysis for making an IDS [2], it consists of filtering, IDS, and honeypot as the traffic after passing filtering and IDS. They rerouted all attacks to a honeypot for in-depth investigation, with False positive rate using anomaly detection system with Specificity of 0,62 and False Positive rate of 38% and it is a limitation.

5) RAP-based DE authentication/disassociation attack's approaches and limitations: **N**o single method recognizes all ET types. The practical technique is the one that identifies a wide range of RAP, needs no adjustment in protocol nor a determined equipment. All current methods have at least one of these highlights, yet none of them has all the features. As it is quite hard to detect Evil Twin, S. Jadhav et al. in [33] have modified the transmission protocol by additional timestamps, which are being observed for detection but protocol modification in itself is a limitation.

A. M. Alsahlany et al. in [34] have presented a good discussion and analysis for the security threats of RAP and its results shows that the RAP always comes in conjunction with DoS and MitM attacks in an experimental way but the author didn't provide any mitigation mechanism.

İ.F.KILINÇER et al. in [35] have presented a RAP mitigation method, an IoT-based approach depending on. A single board computer and a wireless antenna make a RAP detection system by detecting their media access control address (MAC) of the RAP which is assigned to an unauthorized (VLAN) Virtual Local Area Network. The detection mechanism depends on making a comparison between MAC and Basic Service Set Identifiers (BSSID) with identical SSID lists.

These solutions have many limitations as:

- The attacker can overcome the mitigation using an open-source tool (like mac changer) to obtain the LAP BSSID. Benefiting from the propagation of smartphones they used a simple approach to locate the RAP.
- The detection is based on simple comparison of BSSID for networks with identical SSID parameters. An attacker can easily obtain the BSSID and change it for the fake AP (tool mac changer).

B. Existing Solutions Summary and Limitations

So, the current techniques have many downsides: (1) Deployment is with high cost. (2) Protocol adjustment is needed. (3) In need of specified equipment or mainly works on a server or a special hardware. (4) Patching customer software. (5) Based on whitelisting, Result in a large number of false positives. These mentioned points are the reasons for choosing the proposed method which forms a need for addressing such kind of problems.

Therefore, this paper proposes an admin-side solution that defeats the issues related with the current strategies and distinguishes the RAP with higher accuracy and detection rate. It detects WI-phishing [2]or Evil Twin, DE authentication attack, KARMA attack [9], advanced WI-phishing attack and differentiate them from the normal packets. As a kind of phishing attack, based on performing the frame type analysis. In contrast with the previously mentioned solutions, our method of detection has many pros: (a) It needs no whitelisting technique of access points. (b) It provides real-time detection. (c) It does not need any training data of the targeted wireless network. (d) It does not depend on a remote server nor any hardware (e) No need for protocol modification. (f) Determine the attacker's MAC. (g) It needs no prior connection with any AP for detection.

III. Proposed Method

This section clarifies the utilized indicators for the proposed module via analyzing beacon frames and extracting features that are considered as a sign of attack. It helps in detecting WI-phishing [2] or Evil Twin, DE authentication attack , KARMA attack [9], advanced WI-phishing attack and differentiate them from the normal packets in real time and making a long-timed database that is used for forensics for detecting more sophisticated beacon-based attacks via a python language and SCAPY library. The research has a significant impact on the community of network system administrators as it will ease the process of detection in real time and forensics of the evil twin attack.

We have previously mentioned that there is a drawback in IEEE 802.11 protocol as the beacon is sent unencrypted, it helps in the occurrence of many attacks such as RAP. While methods given by researchers as some assisted in securing from the evil twin attack however had their own downsides. These methods range from the installation of special hardware [1],protocol modification [33] and measuring frame characteristics [34], [36], etc.

Limitations in the available Intrusion Detection/Prevention Systems such as Suricata [37] which works only on LAN, and Kismet [38] which has no sophisticated logging method as its pcap file cannot be analyzed in real-time in addition to its massive size.

In our approach, the proposed method countered these drawbacks as it does not need any change protocol, does not depend on learning data or expensive monitoring devices. It depends on a native, better, and real time detection method, depending on analyzing, storing, and visualizing sub types under beacon frames in real-time to figure out the anomaly which reflects the existence of RAP. As well as the long-term real time logging database analysis and visualization which gives more capability and elasticity for further forensics and threat analysis. This database based on Elasticsearch [39] and MongoDB [40], it provides a real time chart, detects anomalies, and generate an alert. It can even send it to the administrator by email. Our method implementation is realized using the language of Python -which enables cross platform implementation- that allows affordability and portability.

A. Detecting Beacon Frames

In the initial step, Scapy library is used for packet capturing phase. So, we can divide our detection algorithm into two different sub algorithms.

B. Real Time Detection Algorithm

1) Depends on IEEE 802.11 management frame -static parameters- that the attacker can sniff, any change in one or more static parameter would be considered as Evil Twin; These static parameters include BSSID, SSID, Channel, Encryption type, Country code, Supported channels and First Channel. We assume that the administrator knows all attributes of his network which should be assigned by the administrator at the first time.

a) Our algorithm can defend one or more Wi-Fi networks as the administrator can provide one or more network properties. These attributes are always being compared against all properties which are being captured in real time from all surrounding networks.

2) Also, there are dynamic Wi-Fi network's parameters, which are very hard for the attacker to imitate as the timestamp and signal strength of the network.

a) The algorithm depends on the fact that the timestamp increases regularly over time in the Wi-Fi access points that we are defending. This means, if we find another access point with the same static attributes but differ in timestamp as if it is less than or equals the last received timestamp from the access point, which means that the last access point is an evil twin.

b) Also for signal strength (RSSIs), we depend on the paper in [41] which is presented by Vanjale et al. as they stated that if the signal differs by 10 dB greater or less, that indicates an evil twin coexistence. Because the attacker may place his RAP nearby the LAP with stronger signal strength or closer to the target to lure them to use his access point. We considered 10 dB< or > it is considered as evil twin.

3) If there is a coexistence of two BSSIDs, in case of a DE authentication attack that exceeds the threshold, it reflects DE authentication attack which in this case considered as an indicator of evil twin. We have set our threshold in this phase as 10 DE authentication packets as mentioned in [42] by Chibiao Liu et al.

4) For detecting KARMA attack, which is only can be pursued in open WIFI networks. So, in case of new open WI-FI networks existence in the surrounding, it is considered as a KARMA Attack. Nevertheless, it will lead to high false positive rate, but it achieves higher recall.

C. Long-time Database for Logging and Forensics

1) We used a strong database that depends on MongoDB [40] Elasticsearch [39] to overcome the weakness in Kismet , like pcap. captured file analysis. As capturing a large file for many days will be extremely hard to be analyzed by using a normal computer.

2) By using this combination of MongoDB with Elasticsearch, we analyze and visualize the captured data for weeks; and in case of anomaly an alert is generated.

3) This database has many advantages for the administrator as knowing if the network is always receiving DE authentication attack from a particular MAC address, or if there is someone probing the users by using a famous Wi-Fi name which is open like the names for airports or cafes to lure the user to use it which is of course a KARMA attack. It enables the administrator to know about his physical location and if this open Wi-Fi is in the surroundings or not.

4) Using the mentioned database, the administrator can know how long the Wi phishing attack was underway. Whether the network targeted by a script kiddie, with a raspberry-pi, who floods the network with DE authentication frames and whether the DE authentication frames was targeting specific device or department.

D. Classes of the Proposed Algorithm

The proposed algorithm has seven correlated classes as follows:

1) Wireless interface management - enabling the monitoring mode in the wireless card and starting channel hopping over frequencies.

2) Scanning wireless networks - capturing transmitted frames of all the surrounding networks.

3) Frame Analysis and filtering - captured frames to find data frames out of the beacon ones, and to separate beacon frames into DE authentication frames beacons, to be compared against the threshold value, and other types of beacons.

4) Compare other types of beacon frames against the predefined parameters - IF difference found between real time captured frame attributes and the predefined attributes, an alert is generated. These attributes are BSSID, SSID, Channel, Encryption Type, Country Code, Supported Channels, First Channel.

5) Compare beacon frames' timestamp and signal strength- generate alert if the timestamp is not incremental or the difference in signal strength is > or < 10 dB than the previously recorded ones.

6) Listing all surrounding open WI-FI, if any, in case of new open Wi-Fi loomed, it is considered as a KARMA attack.

7) Store in database and start visualization - send all frames to a long-time database for forensics purposes and visualize it for further analysis. In case of anomaly, an alert is generated and then returned to the network scanning step.

For final implementation, Python 3.8 programming language was chosen, Fig. 4 figures out the seven phases of the algorithm of detection as follows.

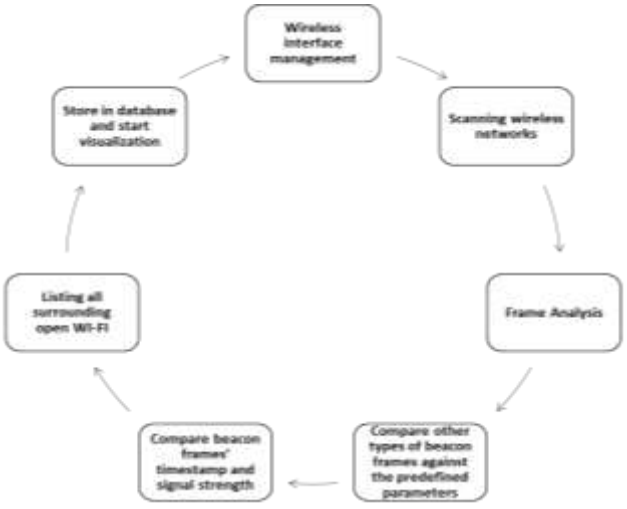


Fig. 4. Phases of the Detection Algorithm.

E. Detector's Pseudo Code

This section shows the proposed algorithm module. As shown in Fig. 5, we start the detection by setting the interface into monitoring mode, then start a continuous channel hopping that goes to each channel and scan it from channel 1 to channel 11. And define static parameters and calculate the dynamic parameters to detect attack in case of a difference occurrence. It also detects De authentication and KARMA attack.

- Start.
- Put interface into monitoring mode.
- Channel hop and scan wireless networks.
- Parse all captured beacon frames and send to logging database for further analysis.
- Start real time analyses.
  - Define the static attributes for network includes MAC_d, $SSID_d$, $Channel_d$, Encryption $Type_d$, Country $Code_d$, Supported $Channels_d$ and First $Channel_d$.
  - Define dynamic attributes for network $Timestamp_t$, Signal-$strength_t$.
  - Define all surrounding open Wi-Fi.
  - Start Frame analyses.
  - IF more than {SSID [coexists] AND DE authentication packets threshold detected}.
    - [Generate alert] "Warning.… Evil Twin detected."
    - Else IF deference found between real time captured frame attributes and the pre-defined attributes ({ MAC_d !=MAC} OR {$SSID_d$!= SSID} OR {$Channel_d$!= Channel} OR {Encryption $Type_d$!= Encryption Type} OR {Country $Code_d$!= Country Code} OR {Supported $Channels_d$!= Supported Channels} OR {First $Channel_d$!= First Channel})
    - [Generate alert] "Warning.… Evil Twin detected."
    - Else IF difference found in timestamp ({$Timestamp_t$ <= $Timestamp_{now}$} OR {Signal-$Strength_t$ [difference >10 dB] Signal-$Strength_{NOW}$ }.
    - [Generate alert] "Warning.… Advanced Evil Twin detected.
    - Else IF found new Open Wi-Fi network.
    - [Generate alert] "Warning.… KARMA attack detected.
  - AND start database visualization.
    - If found anomaly
    - [Generate alert] "Warning.… Evil Twin detected.".
    - Else return to network scanning

Fig. 5. Pseudo Code of the Proposed Method.

F. Flow Chart of the Proposed Method All Captured Beacon

In the following flow chart, Fig. 6, the program starts by putting interface in monitoring mode to scan all the nearby networks by making a channel hopping between channel 1 and 14. Then monitoring scanning and analyzing all beacon frames properties which captured by our network in promiscuous/monitoring mode; Firstly send all captured beacon frames and make a real time comparison between all features that is hard coded from the admin for the needed to be a defended network or networks against the fetched properties from captured beacon frames and in case of matched it generates an alert in real time. Furthermore, the long-term database which can handle, analyze, and visualize features of captured beacon frames to generate valuable statistics to forecast attacks.

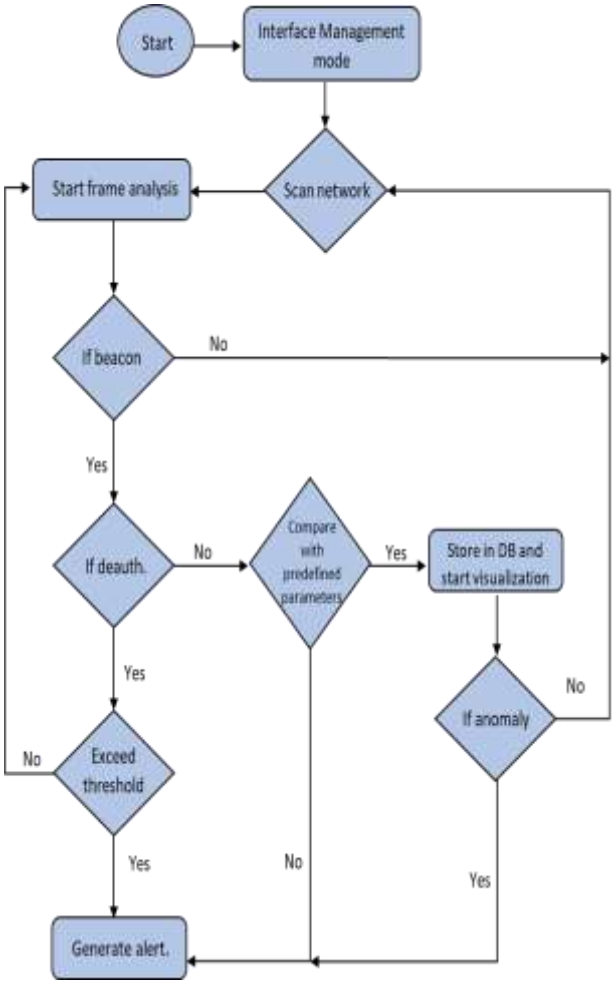


Fig. 6. The Flowchart of the Proposed Method.

IV. Results

This section is determined for the proposed technique's evaluation. In which we describe the laboratory, the design of the detector, the efficiency of the detector, evaluation measures and lastly the results are mapped over a confusion matrix as a predictive analysis method.

A. Laboratory Description

The experiment has been implemented over a network named "MR. Linux" for evaluating the proposed methodology. Python 3.8 programming language, Scapy library and an ALFA model AWUS036H [43] in the promiscuous mode have been used for monitoring and analyzing the packets sent and received. We used airgeddon [44] and Wi-Fi pumpkin [45] to launch the attack, using a Lenovo G4080 Core I5 , 4th generation with 4 gigabyte RAM laptop loaded by Wifislax 2.4 64 bits [46] which is a distribution GNU / Linux. The attacker is launched using a wireless interface card ALFA model AWUS036H. [43] The device that is used for detection is hp EliteBook 745 G3 with AMD64 A10, 8 gigabyte RAM loaded with a Ubuntu 20.4 OS [46] having a Pre-installed Python 3.8, We also use Elasticsearch [39] and MongoDB [40].

B. Proposed Detector's Design

Our proposed algorithm overcomes admin side vulnerabilities in solutions as mentioned in section II as the efficiency of the algorithm does not depend on protocol modification, data sampling, machine learning algorithms, dedicated server, or RTT parameters. As well as it is real-time that does not depend upon training knowledge or Wi-Fi network's fingerprint.

- (KARMA) attack. We simulate a KARMA attack by simulating an open Wi-Fi of a well-known place that is not on the range, if detected in real time.
- (Common Wi-Fi phishing) attack/ replacement WI phishing: We simulate the attack by cloning the BSSID address to be similar to the legitimate AP and flood DE authentication / disassociation frames over all the Wi-Fi network and create another similar fake Wi-Fi network with a different one or more attribute (Encryption type - Channel - First Supported channel or Country Code), it is detected in real time.
- (Common Wi-Fi phishing) attack/ WI phishing coexistence: We simulate the attack by cloning the BSSID to be similar to the legitimate AP and create a coexisting fake Wi-Fi with different one or more attribute (Encryption type - Channel - First Supported channel or Country Code), it is detected in real time.
- (Advanced Wi-Fi phishing) attack with higher signal strength. We simulate the attack by cloning all parameters of the real AP, but with higher signal strength to lure the users to connect to the fake one, it is detected in real time.
- (Advanced Wi-Fi phishing) attack with time difference: We simulate the attack by cloning all parameters of the real AP, but with less timestamp value and based on time difference, it is detected in real time.
- (Real time database is made for more analysis and forensic purposes).
    - To know which of our clients was connected to the Wi-phishing AP, as all beacon frames are logged.
    - Using the mentioned database, we know how long the different attacks were underway.
    - By analysing the Realtime database, we can answer the question, was our network targeted or were DE authentication frames targeting specific devices or departments.

C. Efficiency of Proposed Algorithm

This section is dedicated for evaluating the proposed method. The section is separated into two main parts. The first part evaluates the performance of the proposed detector for classifying the types of attacks on detecting (1) KARMA attack (2) DE authentication attack [9] (3) WI-phishing [2] or Evil Twin, (4) advanced WI-phishing, (5) and differentiating them from the normal packets in real time, furthermore to database all beacon frame for visualization, forensics, and further anomaly inspection. The second part compares the results against the method in [16] proposed by Lovinger et al. , Zeeshan Afzal et al. in [47] and Mayank Agarwal et al. in [1].

D. Evaluation Measures

The performance of the proposed method is analyzed via the estimation of different evaluation metrics like TN rate, TP rate, Specificity Accuracy, false negative rate, and false positive rate, Precision, Recall and F-Measure which are detailed in the subsequent descriptions:

These measures are calculated over a confusion matrix classification as a predictive analysis method based on equation number. 1, 2, 3, 4 and 5. In which TP, FN and FP represent numbers of true positives, false negatives, and false positives, respectively.

1) Specificity: The parameter of specificity is defined as the ratio of total true negatives to the summation of total true negative and false positive value. True negative rate is called specificity.

$$Specificity = \frac{TNs}{TNs + FPs} \quad (1)$$

2) Accuracy: The accuracy metrics are estimated by the parameters value of specificity and sensitivity, which are expressed by equation number 1. Also accuracy refers to how accurate the proposed method can classify frame types in a correct way, and this is expressed by equation number 2, which is applied to return the accuracy value. The accuracy value expresses a comparison between frames that are correctly classified with the whole frames.

$$Accuracy = \frac{(TP + TN)}{TP + TN + FP + FN} \quad (2)$$

3) Precision: The value of precision refers to the number of frames or a category frame that is classified correctly divided by the total frames classified of the same type. Precision is calculated by equation 3. And precision is also referred to as positive predictive value; other related measures used in classification include true negative rate and accuracy.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

4) Recall: Nevertheless, recall shows how many percent of mentioned attacks are correctly classified by the classification. Equation 4 is used for resulting the value of recall. Recall in this context is also referred to as the true positive rate or sensitivity, and it is defined as the ratio of total true positives to the summation of total false negative and false positive value.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

5) F1: is the harmonic average value of precision and recall and calculated by equation number 5.

$$F - score = \frac{(2*Recall*Precision)}{Precision+Recall} \quad (5)$$

E. Evaluation of the Proposed Detector

For evaluation, a comprehensive analysis is conducted. Results showed the FP, TP, FN, and TN presented and analyzed via a confusion matrix as a predictive analysis method. The used set of data for evaluation is outlined in Table I to make sure that the predicted attacks are the actual ones that were sent by the attacking tool. The calculations done via the confusion matrix, Table II shows results with average accuracy of 94.40%, 87.08% average precision and an average specificity of 96.39%.

Table II shows the classifications performance based on the number of each frame type by classifying each type of attack. While Fig. 7 shows the high value of TN with higher value than the TP which increases the overall algorithm detection accuracy which is reflected in Fig. 8.

F. Testing

We have tested our solution against airgeddon [44] and wifipumpkin3 [45] for launching the attack to run the previously mentioned attacks against a network that we have permission to, and calculate the response. We used the OS Wifislax 2.4 64 bits [46] which is a distribution GNU / Linux. The alert reflects the kind of attack underway as shown in Fig. 9 that shows the real time detection. Fig. 10, 11, 12 and 13 show the detection of different types of attack, Fig. 14 shows a sample of anomaly detection through the database. While Fig. 15 represents the database visualization in real time.

TABLE I. CONFUSION MATRIX OF THE PROPOSED METHOD

	KARMA	DE auth.	WI-phishing	advanced WI-phishing	Normal
KARMA	100	0	0	0	0
DE auth.	0	89	3	5	3
WI-phishing	0	2	90	4	4
advanced WI-phishing	0	2	8	87	3
Normal	16	6	0	9	69

TABLE II. CLASSIFICATION AND RESULTS OF THE PROPOSED DETECTOR

	KARMA	DE auth.	WI-phishing	advanced WI-phishing	Normal
TP	100	89	90	87	69
TN	335	346	345	348	366
FP	16	10	11	18	10
FN	0	11	10	13	31
Accuracy	96.45%	95.39%	95.39%	93.35%	91.39%
precision	86.21%	89.90%	89.11%	82.86%	87.34%
Recall	100.00%	89.00%	90.00%	87.00%	69.00%
specificity	95.44%	97.19%	96.91%	95.08%	97.34%
F-score	92.59%	89.45%	89.55%	84.88%	77.09%

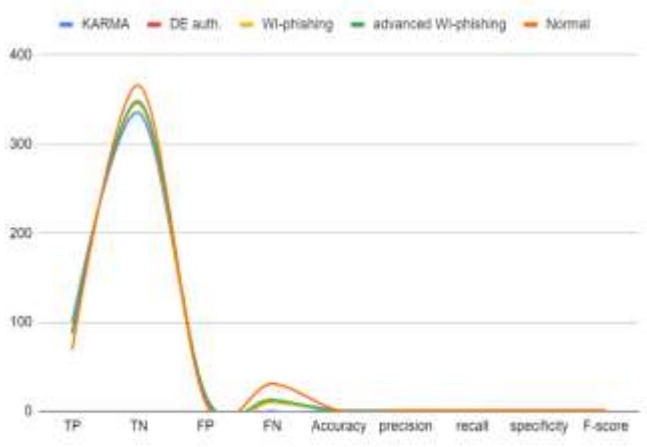


Fig. 7. Matrix Representation Graph.

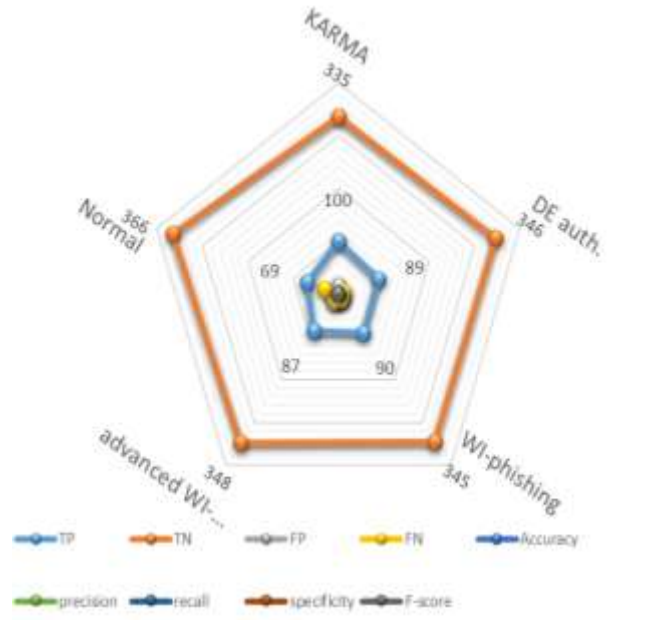


Fig. 8. Algorithm Detection Accuracy.

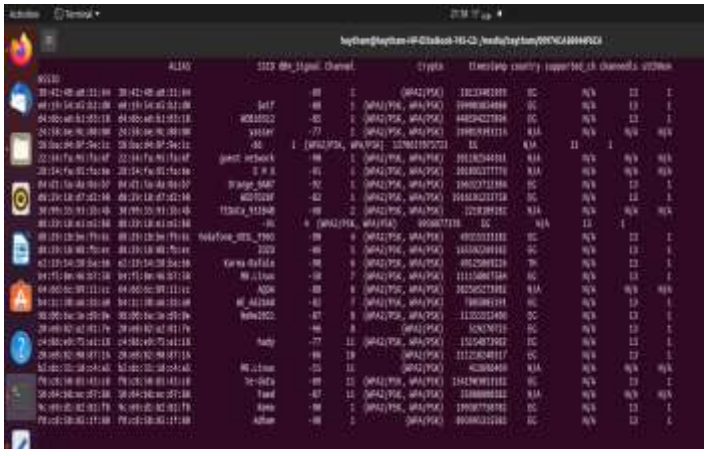


Fig. 9. Real Time Detection.

Fig. 10. KARMA Attack Detected.

Fig. 11. Advanced WI Phishing Attack Detection.

Fig. 12. De Authentication Attack Detected.

Fig. 13. WI Phishing Attack Detected.

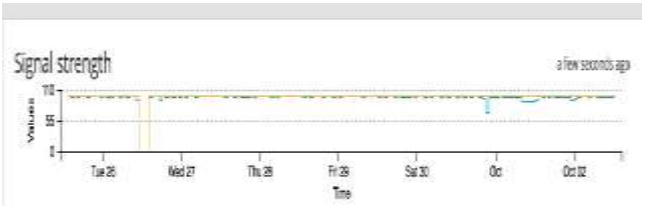


Fig. 14. Anomaly Detection in Signal Strength.

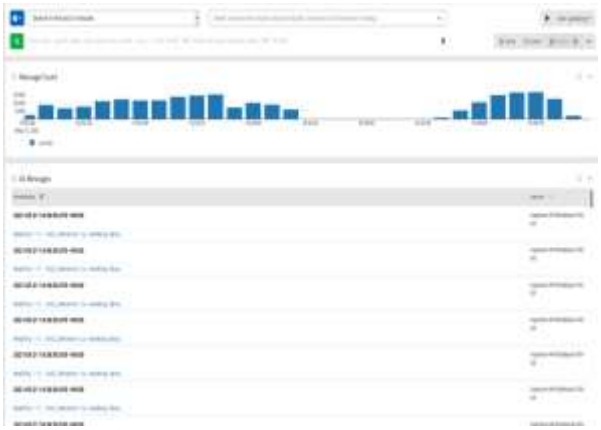


Fig. 15. Real Time Database Visualization.

G. Comparison and Evaluation

Here is a comparison between the proposed method and the method presented in [16] by Lovinger et al. As this approach depends on the network probe using the Raspberry Pi 4. Capability, and creates a logging system. Analyzing the wireless networks and scanning it, then filtering captured frames to create signatures and store it into the database, based on the result obtained the alert is created after comparison. As for the challenges in this approach, it depends on a Raspberry Pi 4 as a device for detection which means more cost and higher limitations, while we developed a cross platform code with less cost. It depends on the log file which is not stable and harder to retrieve data. It depends on creating a unique fingerprint, he depends on a 256-bit hash function, which reflects more time, he uses only static parameters.

While our proposed method depends on static and dynamic parameters for detecting the evil twin, it provides real time detection, it is not passive, but it is active, it is an admin solution, it is cheap, it does not store SSID in DB nor perform bookkeeping of all the APs in the neighborhood. It detects more than one type of attack as WI-phishing or Evil Twin, DE authentication attack, KARMA attack and differentiate them from the normal packets as a kind of phishing attack by depending on Sniffing and analyzing the wireless frames. It has an average accuracy of 94.40%, 87.08% average precision and an average specificity of 96.39% for the five types of attack. The average of the false positive rate is 13 % for all tested types, and it also detects the attacker's MAC address. The results prove that the detector's accuracy is quite high and provide most of the expected features. It also shows that the proposed system can be used for forensic purposes as it can store data for a long time which reflects stability and durability, for data that is stored in the database and start visualization. Table III represents a comparison between the proposed method and three methods for different authors, Levinger et al.[16], Zeeshan Afzal [47] and Mayank Agarwal [1].

To wrap up, we conclude that we have achieved best results after comparing with the previously proposed methods in addition to performing the detection in real time.

TABLE III. A COMPARISON BETWEEN THE PROPOSED METHOD AND THREE OTHER METHODS FOR DIFFERENT AUTHORS

	Lovinger et al.[16]	Zeeshan Afzal et al. [47]	Mayank Agarwal et al. [1]	proposed method
real time detection	y	n	n	y
perform channel hopping	y	n	n	y
using a special device	y	n	n	n
creates a logging system	y	y	y	n
database for forensics	n	n	n	y
analyzing the wireless networks	y	y	y	y
creating network signature	y	y	n	n
based on whitelisting	y	n	n	y
using log file	y	y	y	n
using static parameters	y	-	y	y
using dynamic parameters	n	-	n	y
detecting evil twin	y	y	y	y
detecting KARMA attack	y	n	n	y
detecting DE authentication	y	y	y	y
advanced WI-phishing	n	y	n	y
perform bookkeeping of APs	y	n	y	n
accuracy	-	89%	100%	94.4% av.
false positive rate	-	14.6%.	-	13% av.

V. Conclusion

The wireless network is a primary portion in our world, on account of being used in many life aspects. In this paper, a real time attack detection method has been proposed and helped in detecting different types of wireless attacks as detecting WI-phishing or Evil Twin, DE authentication attack, KARMA attack, advanced WI-phishing attack and differentiate them from the normal packets. While the previously mentioned algorithms of other researchers are either outdated, limited in their detection methods, architecture and/or scope of detection. The implementation was written in Python using the Scapy library by analyzing beacon frames properties in real time and extracting features to be compared against the pre-stored features of LAPs beacon properties and consider any change or a threshold exceeding as a sign of attack. The proposed detector has the advantages of being stable, working in real time, low cost, it does not need extra hardware. It is also powered by a database that can store frames for a long time, which by analyzing them the detector has an added value of forensics, forecasting and detecting anomaly. The detector's efficiency was modelled in a mathematical way and implemented in real life scenarios, returning average accuracy of 94.40%, a value of 87.08% average precision and an average specificity of 96.39% for the different attack scenarios.

VI. Future Work

In the future, we need to analyze our collected data using AI, machine learning and deep learning to generate attack vectors that will help for faster and better detection of the different types of attack. In addition to being willing to deploy the mentioned algorithm for detecting other types of attack rather than the previously mentioned. Also, we can use semantic analysis and ranking technique evaluated in [48] for detecting and ranking other types of attack. We need to make a real time probe request analysis to reduce the value of false positive for the real time detection of the KARMA. Trying not only to detect the attack, but also make a counterattack.

References

[1] M. Agarwal, S. Biswas, and S. Nandi, "An Efficient Scheme to Detect Evil Twin Rogue Access Point Attack in 802.11 Wi-Fi Networks," Int. J. Wirel. Inf. Networks, vol. 25, no. 2, pp. 130–145, 2018, doi: 10.1007/s10776-018-0396-1.

[2] K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: Their types, vectors and technical approaches," Expert Syst. Appl., vol. 106, pp. 1–20, 2018, doi: 10.1016/j.eswa.2018.03.050.

[3] S. Morgan, "2019 Official Annual Cybercrime Report," Cybersecurity Ventur., p. 12, 2019.

[4] "Cyber Security Breaches Survey 2020 - GOV.UK." https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2020/cyber-security-breaches-survey-2020 (accessed Sep. 18, 2021).

[5] E. Medvet, E. Kirda, and C. Kruegel, "Visual-similarity-based phishing detection," Proc. 4th Int. Conf. Secur. Priv. Commun. Networks, Secur., 2008, doi: 10.1145/1460877.1460905.

[6] Proofpoint, "2020 State of the Phish," Proofpoint, pp. 1–48, 2020, [Online]. Available: https://www.proofpoint.com/sites/default/files/gtd-pfpt-us-tr-state-of-the-phish-2020.pdf.

[7] Anti-Phishing WoPhishing Activity Trends Report 1 Quarterrking Group, "Phishing Activity Trends Report 1 Quarter," Most, no. March, pp. 1–12, 2021.

[8] A. Arora, "Preventing wireless deauthentication attacks over 802.11 Networks." 2018, [Online]. Available: http://arxiv.org/abs/1901.07301.

[9] R. Gonçalves, M. E. Correia, and P. Brandão, "A flexible framework for rogue access point detection," ICETE 2018 - Proc. 15th Int. Jt. Conf. E-bus. Telecommun., vol. 2, 2018.

[10] S. M. Hussain, "Impact of DDoS Attack (UDP Flooding) on Queuing Models," pp. 210–216, 2013.

[11] M. A. C. Aung and K. P. Thant, "Detection and mitigation of wireless link layer attacks," Proc. - 2017 15th IEEE/ACIS Int. Conf. Softw. Eng. Res. Manag. Appl. SERA 2017, pp. 173–178, 2017, doi: 10.1109/SERA.2017.7965725.

[12] A. A. Al-zubi, "The International Congress for global Science and Technology ICGST International Journal on Computer Network and Special Issue on Network Security Techniques," 2015.

[13] "WiFi Pineapple - Hak5." https://shop.hak5.org/products/wifi-pineapple (accessed Jul. 18, 2021).

[14] C. Benzaïd, A. Boulgheraif, F. Z. Dahmane, A. Al-Nemrat, and K. Zeraoulia, "Intelligent detection of MAC spoofing attack in 802.11 network," ACM Int. Conf. Proceeding Ser., vol. 04-07-Janu, 2016, doi: 10.1145/2833312.2850446.

[15] "Scapy." https://scapy.net/ (accessed Jul. 20, 2021).

[16] N. Lovinger, T. Gerlich, Z. Martinasek, and L. Malina, "Detection of wireless fake access points," Int. Congr. Ultra Mod. Telecommun. Control Syst. Work., vol. 2020-Octob, pp. 113–118, 2020, doi: 10.1109/ICUMT51630.2020.9222455.

[17] K. Kao, T. Yeo, W. Yong, H. C. the 2011 A. S. on, and undefined 2011, "A location-aware rogue AP detection system based on wireless packet sniffing of sensor APs," dl.acm.org, Accessed: Nov. 07, 2019. [Online]. Available: https://dl.acm.org/citation.cfm?id=1982195.

[18] S. R. Sonawane and S. Vanjale, "Wireless LAN Intrusion Prevention System (WLIPS) for Evil Twin Access Points," IJCST - Int. J. Comput. Sci. Technol., vol. 4–8491, no. 2, pp. 2–5, 2013.

[19] Sachin R. Sonawane, Sandeep P. Chavan, and Ajeet A. Ghodeswar, "Study of Different Rogue Access Point Detection and Prevention Techniques in WLAN," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 3, no. 10, pp. 1232–1237, 2013.

[20] V. Sriram, G. S.-2010 I. 2nd, and undefined 2010, "Detecting and eliminating Rogue Access Points in IEEE-802.11 WLAN-a multi-agent sourcing Methodology," ieeexplore.ieee.org, Accessed: Nov. 07, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5422999/.

[21] M. K. Chirumamilla and B. Ramamurthy, "Agent based intrusion detection and response system for wireless LANs," IEEE Int. Conf. Commun., vol. 1, pp. 492–496, 2003, doi: 10.1109/icc.2003.1204225.

[22] D. Mónica and C. Ribeiro, "WiFiHop - Mitigating the evil twin attack through multi-hop detection," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011, vol. 6879 LNCS, pp. 21–39, doi: 10.1007/978-3-642-23822-2_2.

[23] A. Burns, L. Wu, X. Du, and L. Zhu, "A novel traceroute-based detection scheme for Wi-Fi Evil twin attacks," 2017 IEEE Glob. Commun. Conf. GLOBECOM 2017 - Proc., vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1109/GLOCOM.2017.8253957.

[24] H. Han, B. Sheng, C. Tan, Q. Li, and S. Lu, "HAN ET AL.: A TIMING-BASED SCHEME FOR ROGUE AP DETECTION A Timing-Based Scheme for Rogue AP Detection," MobiCom, vol. 11, pp. 104–115, 2018, [Online]. Available: https://pdfs.semanticscholar.org/1dd9/786e51dd4fbe5df185f4a6ae3e1d70113207.pdf.

[25] C. Mano, A. Blaich, Q. Liao, Y. J.-A. T. on, and undefined 2008, "RIPPS: Rogue identifying packet payload slicer detecting unauthorized wireless hosts through network traffic conditioning," dl.acm.org, Accessed: Nov. 07, 2019. [Online]. Available: https://dl.acm.org/citation.cfm?id=1330334.

[26] Y. Song, C. Yang, and G. Gu, "Who is peeping at your passwords at starbucks? - To catch an evil twin access point," 2010, doi: 10.1109/DSN.2010.5544302.

[27] F. Lanze, A. Panchenko, B. Braatz, and T. Engel, "Letting the puss in boots sweat: Detecting fake access points using dependency of clock

skews on temperature," ASIA CCS 2014 - Proc. 9th ACM Symp. Information, Comput. Commun. Secur., pp. 3–14, 2014, doi: 10.1145/2590296.2590333.

[28] S. Jana and S. K. Kasera, "On fast and accurate detection of unauthorized wireless access points using clock skews," IEEE Trans. Mob. Comput., 2010, doi: 10.1109/TMC.2009.145.

[29] "US20120124665A1 - Method and apparatus for detecting a rogue access point in a communication network - Google Patents." https://patents.google.com/patent/US20120124665 (accessed Nov. 04, 2019).

[30] E. A. Metwally, N. A. Haikal, and H. H. Soliman, "Detecting Semantic Social Engineering Attack in the Context of Information Security," pp. 43–65, 2022, doi: 10.1007/978-981-16-2275-5_3.

[31] IEEE Standard for Information Technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements - ANSI/IEEE Std 802.11, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications - p.447 Table 18-9., vol. 1999, no. June. 2007.

[32] N. Agrawal and S. Tapaswi, "The Performance Analysis of Honeypot Based Intrusion Detection System for Wireless Network," Int. J. Wirel. Inf. Networks, vol. 24, no. 1, pp. 14–26, Mar. 2017, doi: 10.1007/s10776-016-0330-3.

[33] S. Jadhav, S. B. Vanjale, and P. B. Mane, "Illegal access point detection using clock skews method in wireless LAN," in 2014 International Conference on Computing for Sustainable Global Development, INDIACom 2014, 2014, pp. 724–729, doi: 10.1109/IndiaCom.2014.6828057.

[34] A. M. Alsahlany, A. R. Almusawy, and Z. H. Alfatlawy, "Risk analysis of a fake access point attack against Wi-Fi network," vol. 9, no. 5, pp. 322–326, 2018.

[35] İ. F. KILINÇER, F. ERTAM, and A. ŞENGÜR, "Automated Fake Access Point Attack Detection and Prevention System with IoT Devices," Balk. J. Electr. Comput. Eng., no. January, 2020, doi: 10.17694/bajece.634104.

[36] K. F. Kao, W. C. Chen, J. C. Chang, and H. Te Chu, "An accurate fake access point detection method based on deviation of beacon time interval," Proc. - 8th Int. Conf. Softw. Secur. Reliab. - Companion, SERE-C 2014, pp. 1–2, 2014, doi: 10.1109/SERE-C.2014.13.

[37] D. Day and B. Burns, "A Performance Analysis of Snort and Suricata Network Intrusion Detection and Prevention Engines," ICDS 2011, Fifth Int. Conf. Digit. Soc., no. c, pp. 187–192, 2011, [Online]. Available: http://www.thinkmind.org/index.php?view=article&articleid=icds_2011_7_40_90007.

[38] M. Kor, J. Lámer, and F. Jakab, "Intrusion Prevention / I Ntrusion Detection System (IPS / IDS) F OR W I F I N ETWORKS," vol. 6, no. 4, pp. 83–95, 2014.

[39] "GitHub - elastic/elasticsearch: Free and Open, Distributed, RESTful Search Engine.".

[40] "The most popular database for modern apps | MongoDB.".

[41] S. Vanjale and P. B. Mane, "A novel approach for elimination of rogue access point in wireless network," 2015, doi: 10.1109/INDICON.2014.7030418.

[42] C. Liu and J. Yu, "Rogue access point based DoS attacks against 802.11 WLANs," Proc. - 4th Adv. Int. Conf. Telecommun. AICT 2008, pp. 271–276, 2008, doi: 10.1109/AICT.2008.54.

[43] "AWUS036NH (EOL) – ALFA Network Inc." https://www.alfa.com.tw/products/awus036nh?variant=36481029374024 (accessed Jan. 23, 2021).

[44] "GitHub - v1s1t0r1sh3r3/airgeddon: This is a multi-use bash script for Linux systems to audit wireless networks." https://github.com/v1s1t0r1sh3r3/airgeddon (accessed Jan. 23, 2021).

[45] "GitHub - P0cL4bs/wifipumpkin3: Powerful framework for rogue access point attack." https://github.com/P0cL4bs/wifipumpkin3 (accessed Jan. 23, 2021).

[46] "Live Wifislax." https://www.wifislax.com/ (accessed Jan. 23, 2021).

[47] Z. Afzal, J. Rossebo, B. Talha, and M. Chowdhury, "A Wireless Intrusion Detection System for 802.11 networks," Proc. 2016 IEEE Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2016, pp. 828–834, 2016, doi: 10.1109/WiSPNET.2016.7566249.

[48] M. M. El-Gayar, N. E. Mekky, A. Atwan, and H. Soliman, "Enhanced search engine using proposed framework and ranking algorithm based on semantic relations," IEEE Access, vol. 7, pp. 139337–139349, 2019, doi: 10.1109/ACCESS.2019.2941937.

A PSNR Review of ESTARFM Cloud Removal Method with Sentinel 2 and Landsat 8 Combination

Dietrich G. P. Tarigan, Sani M. Isa
Computer Science Department
BINUS Graduate Program-Master of Computer Science, Jakarta, Indonesia

***Abstract*—Remote sensing images with high spatial and temporal resolution (HSHT) for GIS land use monitoring are crucial data sources. When trying to get HSHT resolution images, cloud cover is a typical problem. The effects of cloud cover reduction using the ESTARFM, one of spatiotemporal image fusion technique, is examined in this study. By merging two satellite photos of low-resolution and medium-resolution images, the Enhanced Spatial and Temporal Adaptive Reflectance Fusion Method (ESTARFM), predicts the reflectance value of the cloud cover region. ESTARFM, on the other hand, employs both medium and high-resolution satellite pictures in this study. Using Sentinel 2 and Landsat 8, the Peak Signal Noise Ratio (PSNR) statistical methods are then utilized to evaluate the ESTARFM. The PSNR explain ESTARFM cloud removal performance by comparing the level of similarity of the reference image with the reconstructed image. In remote sensing, this hypothesis was established to get high-quality HSHT pictures. Based on this study, Landsat 8 images that have been cloud removed with ESTARFM may be classed as good. The PSNR value of 21.8 to 26 backs this up, and the ESTARFM result seems good on visual examination.**

Keywords*—Cloud removal; RS-Remote sensing; PSNR-Peak signal noise ratio; GIS-Geographic information system; spatiotemporal image fusion

I. Introduction

An effective way to monitor land-use changes on the earth's surface is remote sensing technology [1]. However, there is still a significant gap between the obtained image and the required image. A grave challenge to maximize the usage of small sensing images (RS) is the cloud problem. Satellite sensors during the acquisition process are unable to capture passive radiation energy in cloud-covered areas, resulting in missed information in satellite images [2]. Electromagnetic waves from ground features are blocked from reaching the sensor system by the cloud. It substantially limits knowledge gained from optical RS images, thus putting a halt to future analysis. Since a result, cloud removal has become a hot issue in the RS community, as it expands the applications and research that optical RS data may be used in Geographic Information System (GIS).

The ESTARFM method is one of many cloud removal methods in remote sensing. The improved spatial and temporal adaptive reflectance fusion model (ESTARFM), which forecasts cloud-covered regions by fusing two satellite pictures, is abbreviated as this model [3]. ESTARFM requires two sets of low and high spatial resolution images captured on two different days (t1 and t2), as well as one collection of images for cloud removal (tp). In the main paper, ESTARFM using Landsat 7 and MODIS imagery. For this research, authors use Sentinel 2 imagery combined with Landsat 8 imagery to run the ESTARFM doing image fusion to correct the cloud effect. The combination of Landsat 8 and Sentinel 2 is chosen to maximize the cloud-covered area result of the ESTARFM method, especially in the area of interest Jakarta, Indonesia. This option combination needs to be investigated because Sentinel 2 imagery is high-resolution imagery, and Landsat 8 is low-resolution imagery. This hypothesis is then evaluated by comparing the ESTARFM result imagery to the actual Landsat 8 imagery image.

Evaluating ESTARFM performance commonly uses statistical analysis like RMSE or MSE [3]. PSNR is another statistical analysis that is possibly used. It is an objective evaluation parameter comparing a reconstructed image to its original image [5]. Besides the final quantitative analysis of PSNR, this paper would also like to visually analyze the influence of Sentinel imagery, cloud density, and specific parameter used. This paper consists of several sections. In Section 2, we presented some current methods used in cloud removal. Section 3, we introduce the ESTARFM method. Then Section 4 explains the research methodology. Section 5 contains the results and experiments. Finally, in Section 6, we bring the paper to a close.

II. Research History

Several prior research has looked towards cloud removal. Cloud removal in remote sensing is the process of reconstructing lost information [4]. According to the previous study, cloud removal methods may be divided into four categories: spatial-based, spectral-based, temporal-based, and multi-source-based [5]. These four approaches are explained below.

The primary purpose of the spatial-based cloud removal model is to provide a complimentary area of cloud covered by exploiting no cloud area of the target image. It estimates pixel loss by spatial interpolation approach as the first type method [6, 7, 8]. Unfortunately, the spatial cloud removal model only performs well on small initial cover gaps and poorly for extensive cloud contamination. So, spatial-based cloud removal can only produce good cloud-free visualization for small holes but is less suitable for quantitative analysis.

Spectral-based cloud removal model can use the correlation of auxiliary clear band and cloud contaminated band to reconstruct the new cloud-free imagery [9]. One of many

approaches is a fog-optimized transformation method [10], correcting the visible radiometric bands from contamination of thin clouds and haze in Landsat imagery. This method is known as the HOT (haze optimization transformation) approach. The HOT change is a visual band space analysis used to see various classes of land use cover. Unfortunately, the spectral response of fog in the visible band space is extremely sensitive to the fog's optical wavelength and depth. The reconstruction result of this approach is quite good in a thin layer of moisture and cannot overcome thick cloud cover. In addition, the spectral-based cloud removal method is unable to distinguish sure land covers such as ice/snow, terrain, and water bodies.

Multitemporal imagery for a given place can be produced by RS systems with frequent satellite revisit cycles (n-day cycles). The temporal cloud removal approach fully uses the temporal correlation between multitemporal images for reconstruction [11, 12]. The material method can effectively reconstruct the lost area due to cloud contamination, especially in congested acquisition frequency of time series data. This method follows solid time series displaying common cloud pollutants when the images are chronologically ordered [4]. Unfortunately, the temporal process is limited to congested data sets, often provided by coarse spatial resolution imagery. On the other hand, the material approach implies that spatial land coverage is determined during the acquisition interval [5]. Temporal-based techniques are appropriate for den situations, according to this general idea.

Combining the advantages of the above methods can solve several problems in cloud removal methods. This approach is called the multisource-based or hybrid approach. The auxiliary picture must have the same wavelength and spatial resolution as the target image in this method [13]. The spatiotemporal method of cloud removal is an example of a hybrid approach that Zhu [14] has been developed. The ESTARFM Model is a commonly used benchmark for developing cloud removal models that employ the spatiotemporal fusion approach. [13]. This algorithm was developed based on STARFM as an improvement.

III. ESTARFM METHOD

Various satellite sensors acquired on the same day and in the exact location can be compared and correlated, mainly when preprocessing steps such as radiometric correction, geometric correction, and atmospheric adjustment have been completed. However, there can be a systematic bias in surface reflection across various sensor images due to sensor system variations such as orbital parameters, bandwidth, acquisition time, and spectral response function. This condition was the primary purpose of ESTARFM to utilize the correlation of various satellite sensor data. Then, integrate multi-source data while minimizing system bias ESTARFM looking at the heterogeneity of the land surface, pure pixels, and mixed pixels [14]. Images with low spatial resolution but high temporal resolution are referred to as "coarse-resolution" by ESTARFM.

In contrast, images with high spatial resolution but low temporal resolution are referred to as "fine resolution" images. [13]. Eq. 1 examines the first steps and equations for spectrally comparable homogeneous pixels within a moving window.

$$\left|L(x_i, y_i, t_k) - L\left(x_{\frac{w}{2}}, y_{\frac{w}{2}}, t_k\right)\right| \leq 2 \times \frac{\sigma(B)}{n} \quad (1)$$

where, $L(x_i, y_i, t_k)$ represent the i-th pixels surface reflectance with the location (x_i, y_i) in the moving window on the observed data t_k. $L\left(x_{\frac{w}{2}}, y_{\frac{w}{2}}, t_k\right)$ is the surface reflectance position of the center pixel $\left(x_{\frac{w}{2}}, y_{\frac{w}{2}}\right)$ in the moving window on the observed data t_k [13]. $\sigma(B)$ displays the whole image's standard deviations for band B.; n shows how many classes are in the study area. A homogeneous pixel is spectrally similar if it fulfills the conditions in equation (1). After a homogeneous pixel is obtained, the weight and conversion coefficient is calculated. Equation 2 is then used to estimate the expected reflection of a center pixel in a moving window.

$$F\left(x_{\frac{w}{2}}, y_{\frac{w}{2}}, t_p, \mathrm{B}\right) = F\left(x_{\frac{w}{2}}, y_{\frac{w}{2}}, t_0, \mathrm{B}\right) + \sum_{i=1}^{N} W_i \times V_i \times \left(C\left(x_i, y_i, t_p, \mathrm{B}\right) - C(x_i, y_i, t_0, \mathrm{B})\right) \quad (2)$$

where N is the quantity of spectrally uniform pixels with similar spectral characteristics, including the moving window's center pixel; (x_i, y_i) indicates the pixel location similar to i-th; w indicates the width of the search window; t_0 and t_p are the observation date and the predicted date; W shows the corresponding weight and V_i is conversion coefficient of the i-similar pixel; $F\left(x_{\frac{w}{2}}, y_{\frac{w}{2}}, t_p, \mathrm{B}\right)$, $F\left(x_{\frac{w}{2}}, y_{\frac{w}{2}}, t_0, \mathrm{B}\right)$ indicate the reflectance of acceptable spatial resolution of the center pixel on the predicted and observed dates, respectively; $C\left(x_i, y_i, t_p, \mathrm{B}\right)$, $C(x_i, y_i, t_0, \mathrm{B})$ represent the coarse spatial resolution reflectance of the pixels located at (x_i, y_i) at the predicted and observed data [13]. The conversion coefficients are calculated using a linear regression model into a similar pixel of high and low spatial resolution imagery [15]. W_i is the weight for similar pixels is calculated based on the spatial distance from the middle pixel of the moving window and the spectral similarity between fine and coarse resolution pixels [13]. W_i is calculated by equation.

$$d_i = 1 + \frac{\sqrt{(x_i - x_{w/2})^2 + (y_i - y_{w/2})^2}}{\frac{w}{2}} \quad (3)$$

$$D_i = (1 - R_i) \times d_i \quad (4)$$

$$W_i = \frac{1/D_i}{\sum_{i=1}^{N} 1/D_i} \quad (5)$$

where, d_i denotes the geographic distance between the center and a spectrally similar uniform pixel in a moving window; R_i denotes the spectral correlation coefficient calculated using the correlation between any similar pixels in a fine spatial-resolution image and a corresponding coarse spatial-resolution pixel; and D_i is an index integrating spectral and spatial similarity [13]. ESTARFM additionally considers the temporal weight when predicting the center pixel's reflection. The following equation is used to determine the final projected reflectance.

$$F\left(x_{\frac{w}{2}}, y_{\frac{w}{2}}, t_p, \mathrm{B}\right) = T_m \times F_m\left(x_{\frac{w}{2}}, y_{\frac{w}{2}}, t_p, \mathrm{B}\right) + T_n \times F_n\left(x_{\frac{w}{2}}, y_{\frac{w}{2}}, t_p, \mathrm{B}\right) \quad (6)$$

where, T_k $(k \in (m,n))$ shows the temporal weight between observed and estimated dates.

IV. Research Methodology

This research consists of four stages, namely: the data collection stage, preprocessing step, the implementation stage, and the evaluation stage. At the data collection stage, the location is the province of Jakarta and timespan around 2020. The images to be used are Sentinel 2 and Landsat 8. The reason for Sentinel 2 images is chosen because its 10-day temporal resolution is expected to be able as an auxiliary image in the spatial, and the spectral resolution of Sentinel 2, which reaches 10 meters, is expected to improve cloud-covered result data of ESTARFM.

In preprocessing stage, the initial steps are done to Sentinel 2 and Landsat 8 imagery. Radiometric, atmospheric, and geometric corrections in the preprocessing setting are performed on SNAP (Sentinel Application Platform) software for Sentinel 2 images and SCP (Semi-Automatic Classification Plugin) plugin [16] for Landsat 8 images in QGIS. Specifically for Sentinel 2 images, an up-sampling process [17] was carried out to equalize the spatial resolution with Landsat 8.

Pairs of fine and coarse resolution images taken on the neighboring date and a set of rough resolution images for the target forecast date are required to execute the ESTARFM model [16]. Fig. 1 states the process flow of the ESTARFM image fusion. Before applying ESTARFM, all images must be processed first to reflect the geo-registered surface.

When it comes to putting ESTARFM into action, there are four key stages [14]. To begin, two fine-resolution and coarse-resolution imagery search for pixels in the local window that is similar to the center pixel. Second, the weights (Wi) of all matched pixels are computed. Then, two pairs of fine and coarse images in the designated frame are used separately to find similar pixels. This step can be seen in Fig. 2 (A). Third, linear regression is used to calculate the conversion coefficient Vi of the fine and coarse images. The conversion coefficient conforms to the yield obtained using regression analysis of identified similar pixels. This step can be seen in Fig. 2 (B). On the target date, the reflectance value of each pair of input imagery is collected. This step can be seen at Fig. 3. So, at the appropriate forecast date, Wi and Vi are utilized to compute the fine-resolution reflection of the coarse-resolution imagery [13]. The weighted average is used in the fourth step, which takes into account three factors: the spectral difference between fine and coarse images taken on the same date, the temporal difference between the images, and the spatial distance between similar pixels in the moving window and the center pixel [14].

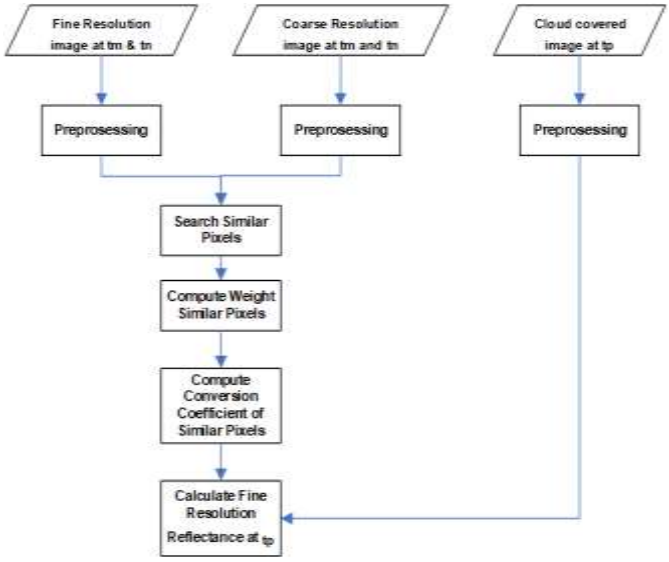


Fig. 1. ESTARFM Image Fusion Method.

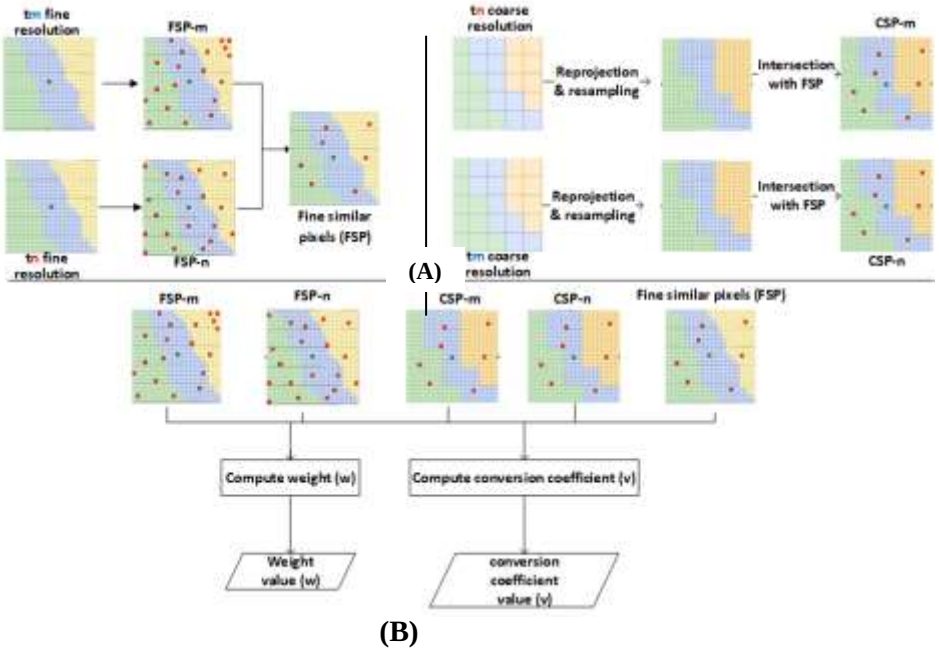


Fig. 2. (A) - Identify Similar Pixels Process; (B) - The Process of obtaining Conversion Coefficient.

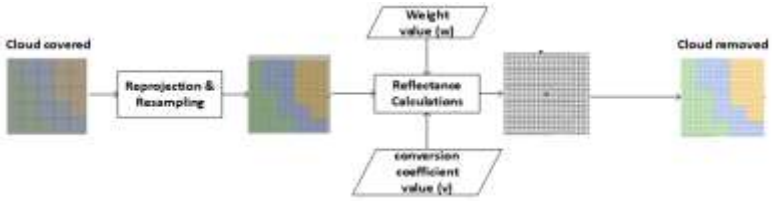


Fig. 3. Reflectance Calculation Process of Cloud-covered Area.

The requirement of two pairs of input images limits the possible combinations of inputs [13]. Also, because of the long processing times for these methods, only one input date combination per target is tested. The weighted average technique obtains the fine image on the date the Landsat 8 image is cloud-covered. Similar pixels play the primary role in the performance of weighted-function-based spatiotemporal fusion methods [18]. This study quantitatively assessed the performance of ESTARFM from the perspective of similar pixels.

ESTARFM also studies influencing factors of similar-pixel identification like the methods of similar-pixel identification, the number of classes (m), and moving window size (w) [13]. Enlarging the period between the base and forecast dates additionally assesses the effect of input base picture pairings. This study could provide a method to analyze the performance of weighted-function-based methods and guidance for the future applications of ESTARFM in a specific area. This entire cycle has been done using a python-based algorithm taken from Zhu's web page in https://xiaolinzhu.weebly.com/.

Following that descriptive/qualitative and analytical/quantitative assessments of the fusion product quality are done when the cloud-free image results have been acquired. The qualitative analysis focuses on a visual examination of variations in reflection and spatial patterns in observed (actual) and expected (cloud-free) pictures [19, 20]. Meanwhile, the PSNR is used for quantitative evaluation. The "Peak Signal-to-Noise Ratio" (PSNR) is a statistical technique used in remote sensing applications to assess data quality following picture fusion, mainly from various satellite sensors [21]. The PSNR value may be used to evaluate the amount of similarity among the original image and the reconstructed/predicted image after fusion to determine the quality of the reconstructed image. The PSNR is calculated using the equation below.

$$PSNR = 20 \times log10 \frac{N}{\sqrt{MSE}} \quad (7)$$

and

$$MSE = \frac{1}{mn}\sum_{y=1}^{m}\sum_{x=1}^{n}(I(x,y) - I'(x,y))^2 \quad (8)$$

N is the most significant potential pixel value in a picture, I (x, y) is the pixel value difference between the original and predicted image. At the same time, m and n are the image's length and breadth, respectively. PSNR is a compression parameter that is often utilized [22]. The amount of resemblance between the compressed or reconstructed picture and the original image is considered this parameter. The mean squared error (MSE) of the two pictures being compared determines the PSNR value. This PSNR value is comparable to the MSE value. Thus, if the MSE between the compressed image and the slice image is less, the PSNR value between the two images will be higher [23]. This paper assumes noise as a cloud-covered area, so a satellite sensor can't receive a signal. The peak ratio between noise and signal also represents the performance of ESTARFM, especially when use Landsat 8 and Sentinel 2 combination in Jakarta province during 2020.

V. Experiments and Discussion

A. Data

The research area defined in this paper is Jakarta, Indonesia, as seen in Fig. 4. This area was chosen because of the frequent occurrence of cloud cover, and the site is relatively small, around 661.5 km^2. Jakarta is determined to reduce mistakes of image fusion due to land-use change. Because Jakarta, as a metropolitan city, is considered to have insignificant land-use changes. In Landsat 8 imagery, Jakarta is covered in tile path/row 122/64 and Sentinel 2 in two tile paths/rows, namely 49 MXU and 49 MYU. The technical specifications of these two images can be seen in Table I. The Sentinel 2 image is chosen as a fine-resolution image because of its five days temporal resolution, compared to Landsat 8 image, which has 16 days temporal resolution and is considered coarse resolution. The smaller the temporal resolution, the higher the chance of getting an image without cloud cover. In addition, Sentinel 2 has a spatial resolution of 10m on the RGB band compared to the 30m that Landsat 8 has on the same band. The more detailed spatial resolution is expected to increase the ESTARFM model's ability to correct cloud-covered Landsat 8 images.

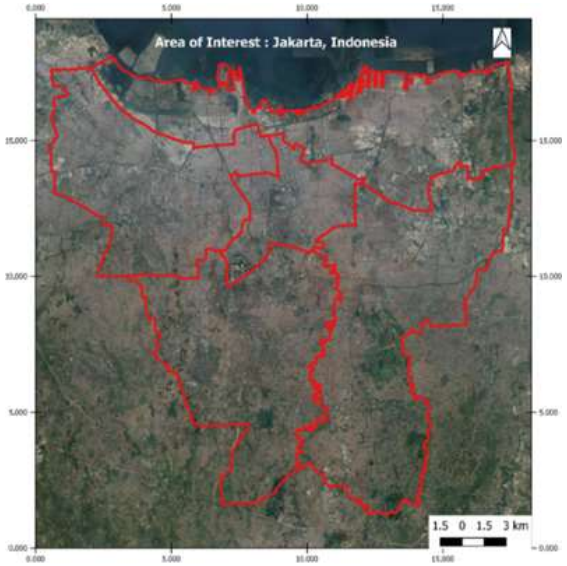


Fig. 4. Area of Interest: Jakarta, Indonesia.

TABLE I. THE FEATURES OF THE LANDSAT-8 AND SENTINEL-2 DATASETS USED IN THIS STUDY

Characteristic	Landsat-8 OLI	Sentinel-2 MSI
Bands (spectral range) used	B2 (0.452 - 0.512 µm)	B2 (0.439 - 0.535 µm)
	B3 (0.533 - 0.590 µm)	B3 (0.537 - 0.582 µm)
	B4 (0.636 - 0.673 µm)	B4 (0.646 - 0.685 µm)
Spatial Resolution	30m	10m
Temporal Resolution	16 days	5 days
Research Location Tile	WRS Path/Row: 122/64	Path/Row: 49MXU 49 MYU

The Sentinel 2 and Landsat 8 satellite imagery data used are the 2020 acquisitions year. Direct visual checks are carried out on the web https://glovis.usgs.gov to select satellite images with certain clouds cover manually. According to the findings, obtaining satellite imagery that is cloud-free by 2020 would be challenging. Correcting and identifying cloud-covered areas in Landsat 8 data, a cloud-free picture is selected from Sentinel 2 and Landsat 8 satellite images with a cloud cover percentage of 0% to 5%.

Several images were used on three simulations, as shown in Table II and Table III. Visually, the condition of Landsat 8 and Sentinel 2 data can be seen in Fig. 5 and the target cloud removal in Fig. 6. As long as in 2020, based on the author's visual analysis via the web https://glovis.usgs.gov, these two Landsat 8 and Sentinel 2 images are considered cloud-free, so they can be used as correcting Landsat 8 photos in the ESTARFM model.

All Landsat 8 and Sentinel 2 data are preprocessed first. Landsat 8 imagery is done in the Semi-Automatic Classification Plugin (SCP) plugin in QGIS, and Sentinel 2 is done in SNAP. Landsat 8 imagery that has been preprocessed, such as radiometric correction and reflectance conversion, is then cropped based on Jakarta's area of interest. After that, reprojection is carried out to the UTM WGS84 48S coordinates according to the location of the Jakarta area.

This treatment is carried out on Band 2 3 4 of Landsat 8 and combined into one complete RGB image on each date. The Sentinel 2 image is preprocessed in SNAP for radiometric correcting and up-sampling the resolution from 10m to 30m. Because Jakarta is located on two tiles, Sentinel 2, namely, 48MXU and 48MYU, mosaicking is done on the SNAP program to combine two tile images on each date. Furthermore, after the two tiles are combined, they are cropped based on Jakarta's area of interest. This treatment is also carried out on band 2 3 4 Sentinel 2 and then combined into one complete RGB image on each date.

B. Implementation Details

ESTARFM [17] python source code has been provided on Zhu's web page. This research uses the PyCharm IDE platform to run the python code. PyCharm is a free and open-source integrated development environment (IDE) for the Python programming language. JetBrains, a Czech Republic-based firm, created this program.

It should be noted that the parameter set in the ESTARFM model needs to be determined first due to Jakarta, as the area of research is different from the original ESTARFM model. The preliminary trial is done to determine the optimal value to be used for the Jakarta area. After parameters are set, the ESTARFM model can be run immediately.

When running, the writer simply points to the desired GeoTIFF file image directory. After the process is finished, a cloud-removed image is generated with the GeoTIFF file extension.

In addition, QGIS is used to extract the reflectance value of the reference image and the model result image. The extraction procedure begins with creating a point-shaped shapefile spread out throughout the Jakarta region of interest. About 10 000 scattered points are assumed as a ground check for the reflectance value of the image. The reflectance values obtained are for RGB or band 2 3 4. Each reflectance value is processed in a spreadsheet to make a scatterplot and PSNR analysis based on the ESTARFM method section.

TABLE II. THE LANDSAT-8 AND SENTINEL-2 REFLECTANCE PRODUCTS USED AS THE BASE IMAGERY

Satellite Imagery	Product ID	Date Acquisition	Remarks
Landsat-8 OLI	LC08_L1TP_122064_20200422_20200508_01_T1	April 22, 2020	Coarse (C) t1
	LC08_L1TP_122064_20200913_20200920_01_T1	September 13, 2020	Coarse (C) t2
Sentinel-2 MSI	S2A_MSIL2A_20200728T025551_T48MXU_20210213T103728.SAFE	July 28, 2020	Fine (F) t1 - stacked image
	S2A_MSIL2A_20200728T025551_T48MYU_20210213T103728.SAFE		
	S2A_MSIL2A_20200827T025551_T48MXU_20210213T111615.SAFE	August 27, 2020	Fine (F) t2 - stacked image
	S2A_MSIL2A_20200827T025551_T48MYU_20210213T111615.SAFE		

TABLE III. THE LANDSAT-8 CLOUD-COVERED TARGET IMAGE

Satellite Imagery	Product ID	Date Acquisition	Remarks
Landsat-8 OLI	LC08_L1TP_122064_20200625_20200707_01_T1	June 25, 2020	Cloud covered (tp) -sim1
	LC08_L1TP_122064_20201015_20201104_01_T1	October 15, 2020	Cloud covered (tp) -sim2
	LC08_L1TP_122064_20201202_20201217_01_T1	December 2, 2020	Cloud covered (tp) -sim3

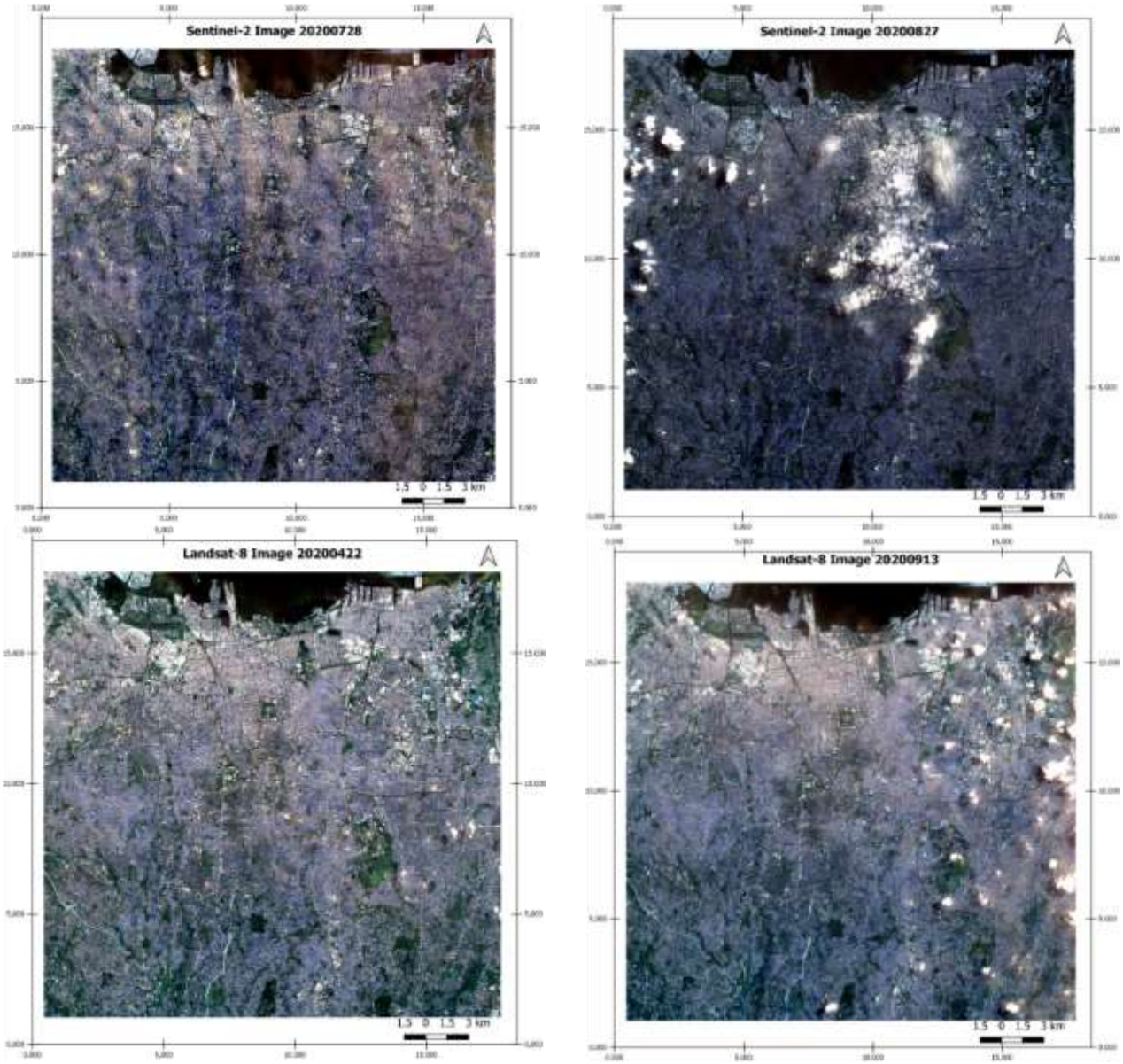


Fig. 5. Landsat 8 and Sentinel 2 relatively Cloud Free Image as based Image.

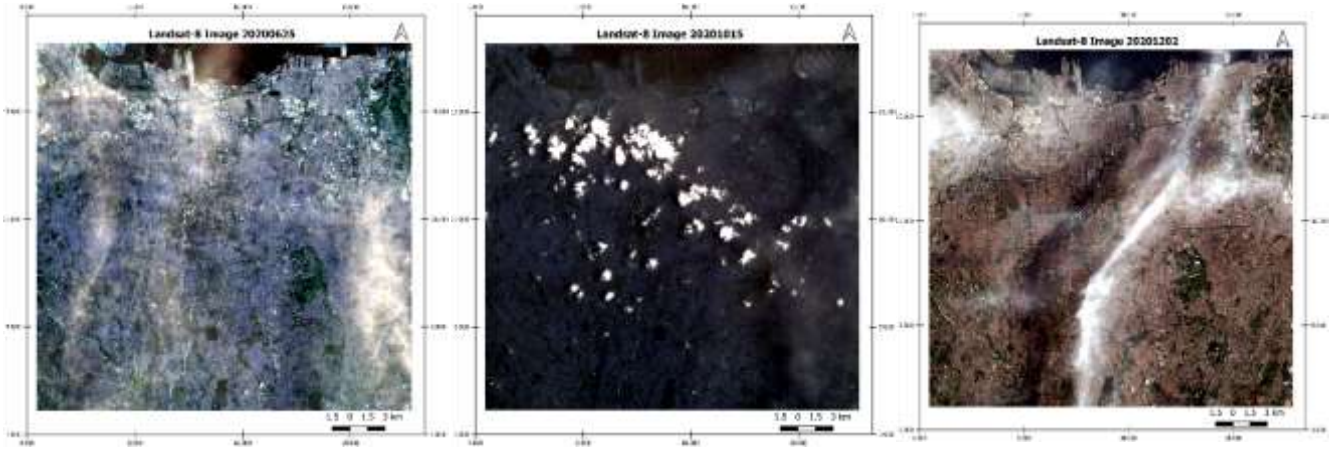


Fig. 6. Landsat 8 Target Image for Cloud removal.

TABLE IV. ESTARFM PARAMETER USED FOR A STUDY AREA

Parameter	Value	Remarks
w	25	# set the half window size. If 25, the window size is 25*2+1=51 fine pixels
num_class	3	# set the estimated number of classes, please set a larger value if blending images with very few bands
DN_min	0	# set the range of DN value of the image,If byte, 0 and 255
DN_max	2	
background	-9999	# set the value of background pixels. 0 means that pixels will be considered as background if one of its bands= 0
patch_long	500	# set the size of each block,if process whole ETM scene, set 500-1000

C. Discussion

At the initial stage of data acquisition, the difficulties in obtaining an image that meets the research needs. The research condition is determined that the correcting image uses 0% to 5% cloud cover percentage. This is because of Jakarta area is quite often covered in clouds. Besides, the resolution of both Sentinel 2 and Landsat 8 imagery in RGB band is 10m, and 30 m and Jakarta area are 662.5 km^2, if the cloud is covered, it will almost cover all the images. In addition, the Landsat 8 16-day temporal resolution is quite challenging to get cloud-free pictures. Even though the spatial resolution is just 10m and is sometimes obscured by clouds, Sentinel 2 imagery is quite helpful in this regard. Still, there are more image options available because the temporal resolution is just five days.

The picture utilized by the ESTARFM model is RGB, but because it uses a GeoTIFF file, the model can see the band-specific data. Landsat-8 and Sentinel-2 spectral bands (bands 2, 3, and 4) are utilized to create synthetic Landsat-like pictures [13]. Because the research region is less diverse, the class labels in Eq. 1 are adjusted to three. Because the pixel vector's dimension was 3 x 1 (band 2, band three, and band 4 in Landsat 8), the criterion in Eq. (1) had to be passed three times to choose a spectrally comparable homogeneous pixel vector [24]. Table IV shows the exact parameters used in this study. As stated in Eq 2, moving windows pick spectrally identical pixels and compute the weight function and conversion coefficient, followed by Eq 3 to 6 [13].

The usage of images with higher temporal and spatial resolution, such as Sentinel 2, to predict the cloud cover area in Landsat 8 is comparable and exhibits spectral similarities. Visually, the prediction image is relatively able to reduce or even expect the cloud-covered site. In the scatterplot distribution, there is a reasonably good correlation. However, on the 25 June 2020 data, the predicted image has a red spot that looks similar to the cloud cover in the reference Sentinel 2 image. This can also be seen in the results of the ESTARFM on the other two data, on August 27th, 2020, and December 2nd, 2020. Likewise, on the data on December 15th, 2020, the cloud noise in the original data has been reduced. However, since the correcting information still contains clouds, the results of the ESTARFM prediction image are also marginally influenced by shadows.

On the other hand, these results show that the ESTARFM model can perform image fusion between Landsat 8 and Sentinel 2. According to the author, there is a possibility that the test will be conducted under the opposite conditions, namely with Sentinel 2, which has a higher spatial resolution and a tighter temporal resolution. As a result of this research, it can be demonstrated that the ESTARFM model can also be used with a Landsat 8 and Sentinel 2 combination.

The ESTARFM parameter setting also determines a good predictive image result. Several test iterations of the model with the parameters used are also sufficient to reconstruct the cloudy data. The parameters that need to be considered in this ESTARFM node are listed in Table IV. Based on the author's trial for the size of interest in Jakarta, the most influential parameters are DN max and CN min, num-class. DN max DN min value is the reflectance value limit that depends on the satellite image data used. Meanwhile, num-class estimates the number of classes that are very influential in combining several image bands so that these parameters need to be explicitly set as required.

PSNR is a frequently used metric to compare the quality of reconstructed pictures to the original image [23]. The PSNR and MSE values are inversely related. This indicates that the larger the PSNR value of the two pictures, the lower the MSE between the compressed image and the slice image. In lossy compression, the compressed image closely resembles the original. Therefore, a higher PSNR value implies more excellent image quality. Meanwhile, there is no difference between the compressed and original images in lossless compression resulting in a PSNR value of infinity [25].

In this study, it can be seen that the PSNR value for each satellite image band is around 26.4 to 27.5 on the acquisition date of June 25th and December 2nd, 2020. Meanwhile, on the acquisition date of October 15th, 2020, the PSNR value ranges from 21.8 to 22.6. Based on this PSNR value, it can be seen that the correlation is that the higher the PSNR value, the results of the ESTARFM prediction image are also closer to the cloud-covered Landsat image. Indeed, in this study, the authors still have difficulty determining the limit of the PSNR value in remote sensing. This is mainly because PSNR is commonly used for image compression as a giant image storage solution. However, based on PSNR writings, the authors assess PSNR above 25 as good for cloud removal cases with ESTARFM [5;12].

TABLE V. VISUAL ASSESSMENT OF OBSERVED AND PREDICTED IMAGE LANDSAT

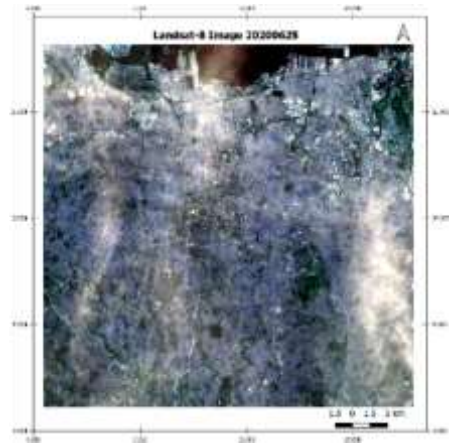
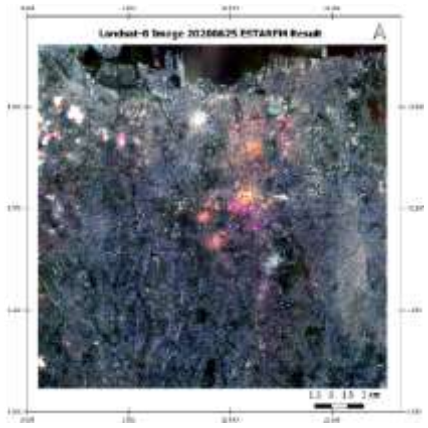
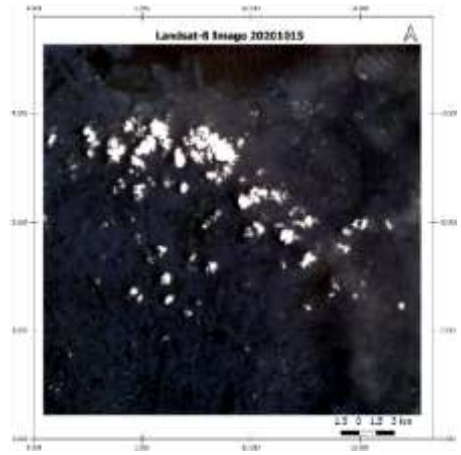
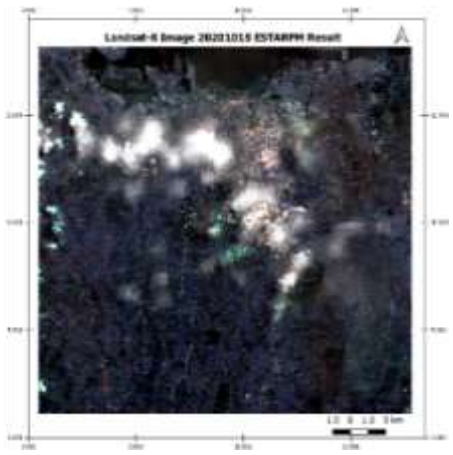
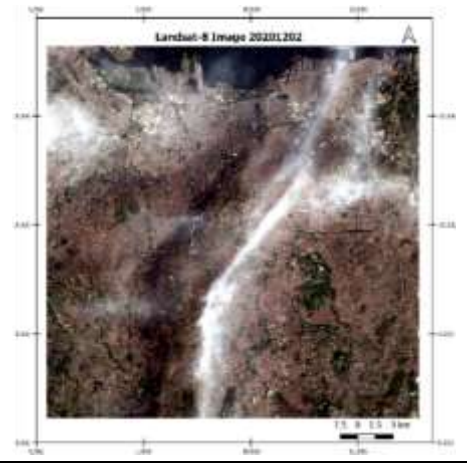
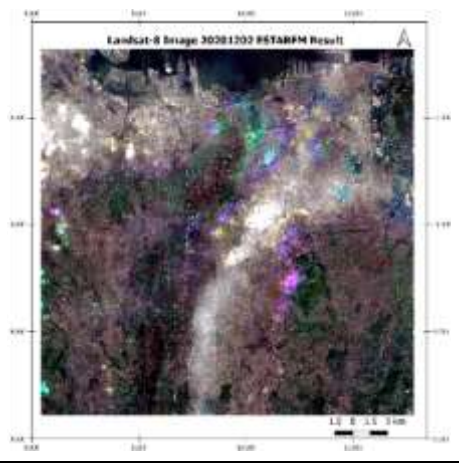
Date	Observed Images	ESTARFM Predicted Image
June 25th, 2020		
October 15th, 2020		
December 2nd, 2020		

TABLE VI. PSNR PERFORMANCE EVALUATION OF DIFFERENT DATE SIMULATION ON RESPECTIVE BANDS

Date	PSNR		
	Band 2 (Blue)	Band 3 (Green)	Band 4 (Red)
June 25, 2020	26.55	27.08	27.42
October 15, 2020	22.61	21.98	21.89
December 2, 2020	27.22	26.42	26.63

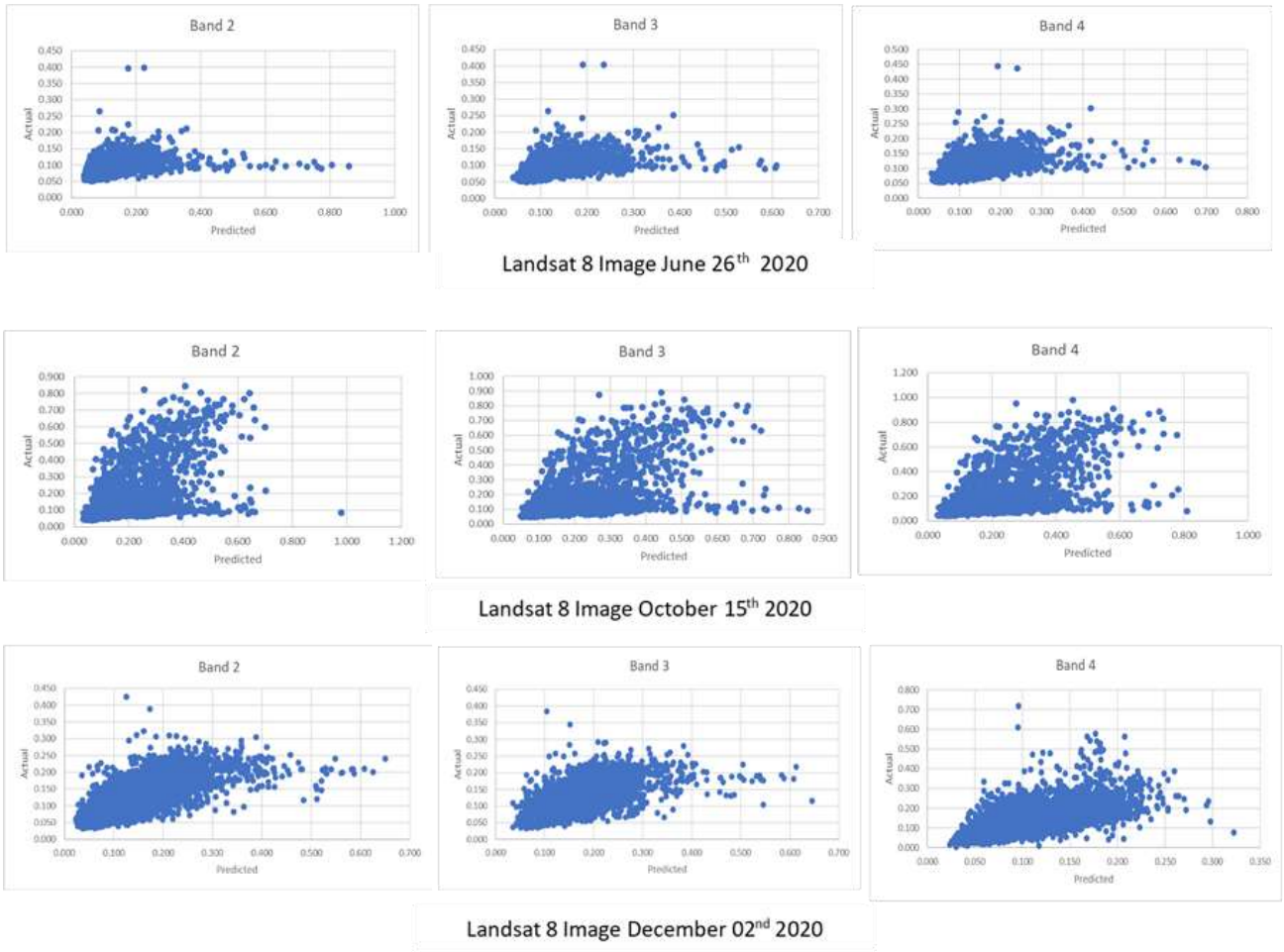


Fig. 7. Actual and Projected Landsat Pictures on Corresponding Dates, with Band-specific Density Scatterplots.

In addition, the scatterplot graph, as seen in Fig. 7, can also be seen the correlation between the reflectance value of the cloud-covered image and the predicted ESTARFM image. Overall, the point distribution is close to the diagonal line, but some wrong reflectance values still move away from the trendline. Each scatterplot consists of 10000 points of reflectance values from each band of observation dates. So that the visual analysis of the reflectance value can be done more easily besides seeing the image directly, which cannot be measured and tends to be qualitative, however, the scatterplot analysis can be done quantitatively. So, it can be said that the combination of Landsat 8 and Sentinel 2 in the ESTARFM model for cloud removal has worked well and can correct some cloud-covered values.

VI. Conclusion

The primary objective of this research was to contribute a PSNR review of ESTARFM as one of many spatiotemporal fusion frameworks using specific parameters for the area of interest Jakarta. Overall, ESTARFM performs well in our research. The weight-function-based fusion model ESTARFM has successfully blended high spatial-temporal resolution of Sentinel-2 data with the low spatial-temporal resolution of Landsat-8. As seen in Table V, Fig. 2, and Table VI, ESTARFM performance is average beyond good. The further research possibilities, authors recommended using 100% cloud-free image or below 2% cloud coverage area as based fusion image to predict and correct cloud-covered area of Sentinel-2 and Landsat-8 imagery. One source of satellite imagery that can be used is Lapan. On the official website https://inderaja-catalog.lapan.go.id/ you can find satellite images of various resolutions, especially for the Indonesian region. In addition, further parameter analysis can also be carried out in other regions in Indonesia.

Acknowledgment

We are thankful to the Faculty of Computer Science, Binus University, Jakarta, for giving us this opportunity and facilities to carry out this study.

References

[1] J. B. Campbell, Introduction to remote sensing, New York: Guilford Press, 1996.

[2] H. Briassoulis, "Analysis of Land Use Change: Theoretical and Modeling Approaches," WVU Research Repository, 2020.

[3] C. Kwan, X. Zhu, F. Gao, B. Chou , D. Perez , J. Li and G. Marchisio, "Assessment of spatiotemporal fusion algorithms for planet and worldview images," Sensors, vol. 18, no. 4, 2018.

[4] H. Shen, X. Li, Q. Cheng, C. Zeng, G. Yang, H. Li and L. Zhang, "Missing information reconstruction of remote sensing data: A technical review," IEEE Geoscience and Remote Sensing Magazine, vol. 3, no. 3, pp. 61-85, 2015.

[5] H. Shen, J. Wu, Q. Cheng , M. Aihemaiti, C. Zhang and Z. Li, "A spatiotemporal fusion based cloud removal method for remote sensing images with land cover changes," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 3, pp. 862-874, 2019.

[6] F. Van der Meer, "Remote-sensing image analysis and geostatistics," International Journal of Remote Sensing, 33(18), pp. 5644-5676, 2012.

[7] C. Zhang, W. Li and D. Travis, "Gaps‐fill of SLC‐off Landsat ETM+ satellite image using a geostatistical approach," International Journal of Remote Sensing, vol. 28, no. 22, pp. 5103-5122, 2007.

[8] P. Scaramuzza and J. Barsi, "Landsat 7 scan line corrector-off gap-filled product development," In Proceeding of Pecora, pp. 23-27, 2005.

[9] P. Rakwatin, W. Takeuchi and Y. Yasuoka, "Restoration of Aqua MODIS band 6 using histogram matching and local least squares fitting," IEEE Transactions on Geoscience and Remote Sensing, 47(2), pp. 613-627, 2008.

[10] Y. Zhang , B. Guindon and J. Cihlar, "An image transform to characterize and compensate for spatial variations in thin cloud contamination of Landsat images," Remote Sensing of Environment, vol. 82, no. 2, pp. 173-187, 2002.

[11] C. Lin, K. Lai, Z. Chen and J. Chen, "Patch-based information reconstruction of cloud-contaminated multitemporal images," IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 1, pp. 163-174, 2014.

[12] C. Lin, P. Tsai, K. Lai and J. Chen, "Cloud removal from multitemporal satellite images using information cloning," IEEE transactions on geoscience and remote sensing, vol. 51, no. 1, pp. 232-241, 2013.

[13] R. Ghosh, P. Gupta, V. Tolpekin and S. Srivastav, "An enhanced spatiotemporal fusion method – Implications for coal fire monitoring using satellite imagery," International Journal of Applied Earth Observation and Geoinformation, vol. 88, no. 102056, pp. 1-15, 2020.

[14] X. Zhu, J. Chen, F. Gao, X. Chen and J. Masek, "An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions," Remote Sensing of Environment, vol. 114, no. 11, pp. 2610-2623, 2010.

[15] J. Xue, Y. Leung and T. Fung, "An unmixing-based Bayesian model for spatio-temporal satellite image fusion in heterogeneous landscapes," Remote Sens. (Basel) 11 (3), p. 324–346, 2019.

[16] C. Luca, "Semi-Automatic Classification Plugin for QGIS," 2020. [Online]. Available: http://dx.doi.org/10.13140/RG.2.2.25480.65286/1.

[17] J. Louis, R. Debacker, B. Pflug, M. Main-Knorn, J. Bieniarz, U. Mueller-Wilm and F. Gascon, "Sentinel-2 sen2cor: L2a processor for users," in Proceedings Living Planet Symposium 2016, 2016.

[18] K. Hazaymeh and Q. Hassan, "Spatiotemporal image-fusion model for enhancing the temporal resolution of Landsat-8 surface reflectance images using MODIS images," Journal of Applied Remote Sensing,, vol. 9, no. 1, 2015.

[19] J. Ma, W. Zhang, A. Marinoni, L. Gao and B. Zhang, "An improved spatial and temporal reflectance unmixing model to synthesize time series of landsat-like images," Remote Sensing, vol. 10, no. 9, 2018.

[20] J. Ma, W. Zhang, A. Marinoni, L. Gao and B. Zhang, "Performance assessment of ESTARFM with different similar-pixel identification schemes," Journal of Applied Remote Sensing, vol. 12, no. 2, 2018.

[21] L. Bo, Q. Youshan, F. Guilan, Y. Xiufang and X. Bin, "Maximal PSNR wavelet bi-linear interpolation iterative algorithm in remote sensing image," ACTA Photonica Sinica, vol. 35, no. 3, pp. 468-480, 2006.

[22] A. Babaee, S. Shahrtash and A. Najafipour, "Comparing the trustworthiness of signal-to-noise ratio and peak signal-to-noise ratio in processing noisy partial discharge signals," IET Science Measurement & Technology, pp. 112-118, 2013.

[23] N. Kristianti, N. Purnawanti and B. Rolando, "Analisis Pengaruh Citra Gelap, Normal, Terang Terhadap Wavelet Orthogonal.," Jurnal Buana Informatika, pp. 93-100, 2018.

[24] F. Chen, L. Tang and Q. Qiu, "Exploitation of CBERS-02B as auxiliary data in recovering the Landsat7 ETM+ SLC-off image," in 18th International Conference on Geoinformatics, IEEE, 2010.

[25] A. Widipaminto, A. Indradjad, D. Monica and R. Rokhmatuloh, "Analisis Metode Kompresi Berdomain Wavelet Pada Citra Satelit Resolusi Sangat Tinggi," Jurnal Penginderaan Jauh dan Pengolahan Data Citra Digital, vol. 16, no. 1, 2020.

An Improved K-anonymization Approach for Preserving Graph Structural Properties

A. Mohammed Hanafy, Sherif Barakat, Amira Rezk
Dept. of Information System, Faculty of Computers and Information
Mansoura University, Mansoura, Egypt

***Abstract*—Privacy risks are an important issue to consider during the release of network data to protect personal information from potential attacks. Network data anonymization is a successful procedure used by researchers to prevent an adversary from revealing the user's identity. Such an attack is called a re-identification attack. However, this is a tricky task where the primary graph structure should be maintained as much as feasible within the anonymization process. Most existing solutions used edge-perturbation methods directly without any concern regarding the structural information of the graph. While that preserving graph structure during the anonymization process requires keeping the most important knowledge/edges in the graph without any modifications. This paper introduces a high utility *K*-degree anonymization method that could utilize edge betweenness centrality (EBC) as a measure to map the edges that have a central role in the graph. Experimental results showed that preserving these edges during the modification process will lead the anonymization algorithm to better preservation for the most important structural properties of the graph. This method also proved its efficiency for preserving community structure as a trade-off between graph utility and privacy.**

***Keywords*—Privacy; social networks; anonymization; edge-perturbation methods**

I. Introduction

Social network sites have become one of the largest sources of personal information. Daily, millions of users can use social applications like Twitter, Facebook, and LinkedIn to communicate with others. The increase in data being collected from different social network sites has attracted many researchers and social network analysts for extracting knowledge from data [1]. Hence, social network data publishing for analysis purposes becomes Inevitable, as the structural data analysis and studying the relations between individuals can serve many fields including marketing and also business. Social data includes a large amount of sensitive information about individuals, so releasing data of social networks in its primary form without anonymizing it could expose data to many attacks [2], [3], which harms the user's privacy. Many types of data privacy-related attacks had been discussed in previous literature [4], [5], which were summarized as follows: identity disclosure, sensitive attribute disclosure, and link disclosure risk. That's why privacy preservation methods must be implemented by specialists before the release of network data to the public.

The re-identification attack causes dangerous violations of social networks which harm user's privacy. An adversary can violate the user's privacy in two ways: (1) either by reaching the target's personal information such as name, edge, and salary, known as profile data, or (2) by utilizing the graph structural information. Recognition of the topological structure of graphs and relations between individuals enables an adversary to utilize his background knowledge to re-identify individuals. Once an adversary recognizes a specific person in the social network, all sensitive information related to him becomes identified. Also, confidential information regarding the belonging of individuals to a particular community becomes disclosed. For example, in the healthcare domain, PatientsLikeMe is a social network site that consists of several communities of patients. Each community represents the patients that suffer from the same illness. To keep track of their health and benefit from patient-reported concerns, members of this site are allowed to exchange private information such as health status and treatments [6]. In such a case, the disclosure of a patient's existence in a particular group will result in revealing all secret information that they share with others and violating their privacy.

The primitive way that people follow to prevent re-identification attacks, for the publishing data of social networks, is to delete a user's identifier attributes and replace them with symbols or synthetic identifiers. This method is known as simple and naïve anonymization. The authors in [7] presented two types of attacks of the naïve-anonymized graph: passive attack and active attack, which means that this simple method of anonymizing graphs is not enough to prevent the re-identification attack. The attacker can exploit his background knowledge concerning only the graph structure to reach the target and breach privacy.

For example, in the above-displayed graph shown in Fig. 1, each vertex/node represents an individual, and the edge connecting between two individuals represents the relation between them. After performing the naïve anonymization on the original graph G, we can get an anonymous version G^* as shown in Fig. 1(B). If an attacker has some background knowledge about Carl and knows that Carl has five friends. Hence, he can re-identify Carl in the anonymously published graph G^* and reach all sensitive information about Carl. Once an attacker got to the information about Carl, this will also increase the probability that this attacker will reach all of Carl's friends. So, such a method can't preserve the user's privacy. Therefore, researchers extended the well-known K-anonymity [8] model, introduced to protect statistical data from the disclosure risk, to develop different privacy models of the graph according to various assumptions of an attacker's

background knowledge. Such as K-Degree [4], K-Neighborhood [9], [10], and K-Automorphism [11] model to prevent different types of structure-based re-identification attacks.

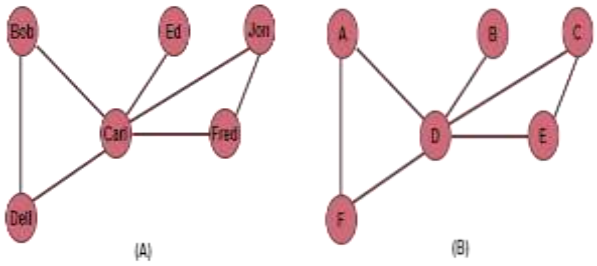


Fig. 1. An Example to show Naïve-anonymization of Graph Data, which (A) is the Primary Graph G and (B) is the Perturbed Version G* of G.

This paper assumes that structural information is the only information available to an attacker that can exploit it to carry out a re-identification attack. Assuming that the attacker is aware of the vertex degree of the target vertex. Researchers introduced many attempts to tackle such a case in previous studies by applying the *K*-Degree anonymity model. This model can distort networks structure by adding or deleting edges so that each vertex of the adjusted version is identical with at least $(K-1)$ vertex concerning vertex degree. However, this approach may cause a large distortion to the local structure of the primary graph. Thus, this distortion will harm the utility of data especially, when the anonymized data is used to meet analytical needs. The main reason behind this large distortion is that most existing anonymization algorithms, which are based on the edge modification approaches, don't take into consideration the concept of edge's relevance proposed in [12], which aim at maximizing data utility through keeping the important edges in the graph without any modifications.

In this paper, we introduce edge betweenness centrality measure [13] to highlight the most valuable edges in the graph and apply the *K*-Degree anonymity model only to edges with no or fewer betweenness values to preserve privacy and maximize the data utility, especially for clustering processes. Since edges with high betweenness values consider the most important knowledge for some popular community detection algorithms to discover community [14].

The remnant of this paper will be structured as follows, Section II discusses the literature review, Section III introduces the proposed Method, Section IV declares the results and evaluation, lastly, Section V highlights conclusions and the directions of the future works.

II. Literature Review

According to the previous works of literature [15] [16] on the anonymization of social networks, different anonymization approaches are categorized into three main groups: Edge modification-based anonymization approaches, clustering-based generalization, and differential privacy approaches [17].

Edge modification-based anonymization [4], [9]: these methods can anonymize the graph structure through modifying edges (adding and/or deleting) until reaching the desired value (*K*-anonymity). While some other methods suggest modifying the edges of the graph randomly.

Clustering-based generalization approaches [18]: these methods cluster nodes that are similar together (groups). Then each group will be generalized into an obscure cluster without any information about a specific individual. Although such methods succeeded in hiding the details about individuals, they fail in preserving the local graph structure of the social network. Because the graph structure is shrunk during the anonymization process. Consequently, these methods will not be eligible for analyzing the graph structure [19].

Differential privacy approaches: such methods seek for preserving user's privacy through imposing restrictions on the data release mechanisms; whereas the differentially private-based algorithms aim at providing statistical information about data without allowing direct access to the whole database. Consequently, such methods prevent a malicious attacker who can query the database from disclosing the target's identity.

In this paper, we focus on previous studies that addressed the anonymization problem through Edge modification-based approaches. Some authors concentrate on preserving the general structural properties of the anonymized network [20]–[22], While others are interested in preserving the community structure in the anonymized version [23], [24].

The authors in [25] compared the results of four algorithms, used for implementing *K*-degree anonymity, in terms of the information loss furthermore the data utility. These algorithms were introduced by different authors. The first one introduced the concept of *K*-degree anonymity in [4]. The second and third algorithms are EAGA and UMGA presented [26], [27] respectively. The last one, introduced in [28], which are based on the vertex addition method. They tested all algorithms using the same configurations. Each one follows its method for minimizing the changes performed on the graph structure. Their results showed that the UMGA scored the best results with all tested networks because it succeeds in minimizing the number of edges modified within the anonymization phase.

The authors in [12] propounded an efficient anonymization approach for creating a *K*-degree anonymized graph. They utilized the neighborhood centrality as a measure for assigning the most significant edges in the graph. They proved that preserving these edges during the anonymization process decreases the amount of information loss. At the same time, their method proved its efficiency in increasing the usefulness of the anonymized graph for evaluating the clustering process. Also, their algorithm achieves the highest results with less information loss compared to other popular anonymization algorithms.

The authors in [29] presented a new method to satisfy *K*-degree anonymity through node addition and edge set modification. Instead of adding nodes randomly, they gave the priority to the nodes with low betweenness centrality values to be modified. Their results proved that their approach could preserve APL, Closeness centrality, as well as nodes degree. But they didn't clarify how their proposed method achieves utility about the preservation of the anonymous graph's community structure.

The authors in [30] introduced a genetic *K*-degree anonymity method in two steps to enhance the preservation of the structural information in anonymized graphs. In the first step, they partitioned vertices of the graph and assigned a label for each vertex to show how many edges needed to be added to achieve the required *K*-degree anonymized sequence. Then, they identified the set of vertices that should be existed in each community. In the second step, within each community, a few edges were added between the vertices to modify the graph using a meta-heuristic algorithm [31].

III. The Proposed Method

In our proposed approach we seek to preserve the most impactful edges during the modification phase which in turn help us to limit the number of the modified edges. We present edge betweenness Centrality (EBC) measure to determine the most essential edges in the graph. Also, keeping these edges in the anonymized network will lead the suggested approach to optimize data utility for clustering analysis.

A. Overview

For undirected and unlabeled graph $G(V,E)$, where V describes the set of vertices, and E defines the edges set in the graph. Let DS defines the degree sequence of graph G, where DS is a term to describe the vector of elements, i.e. $DS = \{ d_{V1}, d_{V2}, \dots, d_{Vn} \}$. each element $d_{vi} \in Ds$ is an integer, whereas d_{vi} is the degree value of vertex vi and n is the number of elements (vertices).

Regarding the graph anonymization, Liu and Terzi introduced two essential definitions in [4] for satisfying the *K*-degree anonymity concept:

1) A degree sequence DS is described as *K*-anonymous when each distinct value $d_{vi} \in Ds$ appears not less than *K* times.

2) A graph $G(V,E)$ is known as a *K*-degree anonymous graph when the degree sequence of the graph G is *K*-anonymized. As shown in Fig. 2.

By considering the previous definitions, we introduce our enhancing approach to anonymize the graph as described in Fig. 3. Our approach goes through two main stages. The first one accepts the original graph and anonymized the degree sequence. After executing this stage and getting the anonymized degree sequence, the second stage starts to realize the anonymized graph G^*. Finally, the utility estimation of the anonymized graph version will be evaluated in the experimental results section by extracting the community structure for both the initial and anonymized version of the graph.

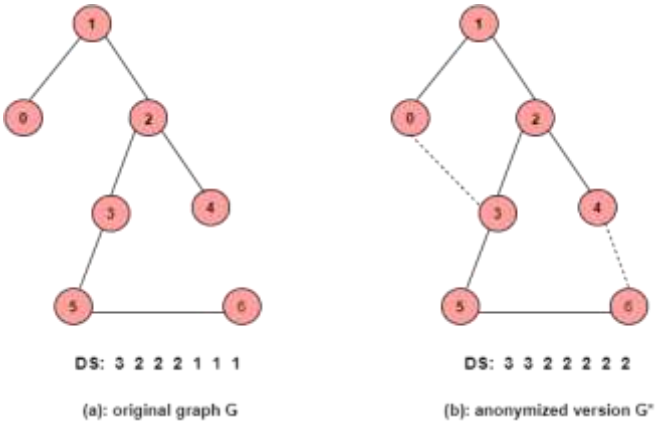


Fig. 2. Show an Example of Achieving 2-Degree Anonymity through Inserting some of Edges. (a): Original Graph G. (b): Anonymized Version G*.

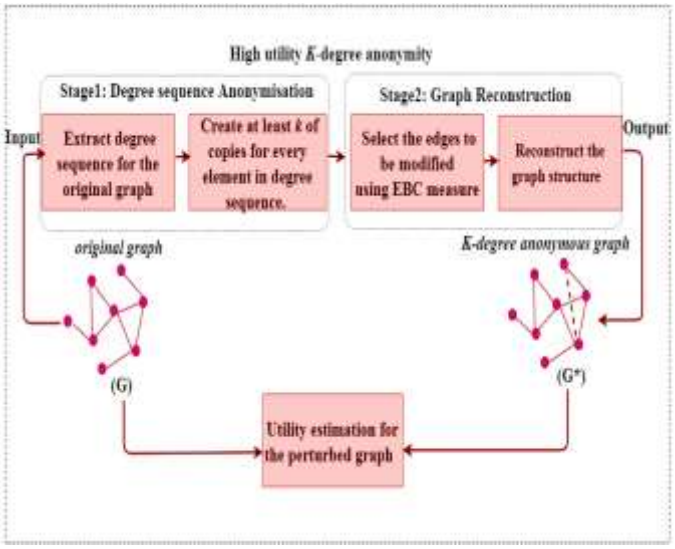


Fig. 3. Overview of Scheme.

B. Stage1: Degree Sequence Anonymization

Taking description (A) into consideration, we must adjust the values of DS to construct groups of at least K copies for each element. To satisfy this definition, we adopt the well-known univariate microaggregation technique proposed in [32] to perturb the degree sequence of the primary graph. The main objective is to get the optimal solution that decreases the distance between the primary degree sequence (DS) and the resulting K-anonymous sequence (DS^*), using the distance function:

$$dist\ (DS, DS^*) = \sum_{i=1}^{n} \left| d_{vi}^* - d_{vi} \right| \quad (1)$$

Our method starts by getting an optimal partitioning of graph vertices which is an order degree sequence that has been divided into several groups, then modifying the values of each group to achieve the required degree sequence that minimizes distance calculated by Eq.1. As stated in [32], given a directed graph $G_{k,n,}$ the optimal partition is defined as a set of groups which match the arcs of the shortest-paths that follows from the source vertex 0 to a vertex n of the graph. Each group that belongs to the optimal partition represents an arc that exists on the shortest path in the graph. The group size is in the range of K and $(2K-1)$ items. Then, to modify the values of each group of the optimal partition, we calculated the differences matrix as computed by [12]. Using this matrix, several solutions existed to satisfy *K*-degree anonymity. Finally, the Greedy method being selected to find the optimal solution among all possible ones using a probability distribution matrix.

C. Stage2: Graph Reconstruction

In this stage, we start to adjust the original graph according to the anonymous degree sequence DS^* resulting from the first stage. In our approach, the scope of modifications made is limited only to the set of edges, while the vertices set don't get any changes. We deploy three types of operations to modify the set of edges of the original version:

- Edge insertion operation is to link and include a new edge between two vertices vi, vj and it is denoted as $edge^{ins}(vi, vj)$.
- Edge deletion operation is to eliminate an existing edge between two vertices, denoted as $edge^{del}(vi, vj)$.
- Edge swap operation is used to switch between two edges, i.e., $e1(vi, vj)$ with $e2(vi, vf)$. It is referred as edge $^{swap}((vi, vj), (vi, vf))$.

Instead of specifying edges to be modified randomly, as with most previously used anonymization methods, we prefer to select the set of auxiliary edges that help to preserve the graph structure for the community analysis purpose. So, we utilize edge betweenness centrality measure for quantifying the most significant edges in the graph. The betweenness centrality of an edge e being estimated by computing the number of times that this edge exists on the shortest paths in each pair of graph vertices. It is computed as follows:

$$BC(e) = \sum_{s \neq t \in V} \sigma_{st}(e)/\sigma_{st,} \quad (2)$$

Where σ_{st} indicates the number of shortest-paths from a vertex $\boldsymbol{s}$ to a vertex $\boldsymbol{t}$ while $\sigma_{st}(e)$ is the number of the paths that go across $\boldsymbol{e}$.

Among all available edges to be adjusted, we choose edges with high betweenness values to be preserved during the modification process. These edges have more importance than others. Only edges with no or low betweenness values are allowed to be modified during the modification process.

D. Summary

Algorithm: High utility $\boldsymbol{k}$-degree anonymity algorithm

Input: A graph $G(V, E)$, and anonymity parameter $\boldsymbol{k}$.
Output: k-degree anonymous graph G^*.

Begin:
// elements sorted in a descending order.
1: DS ← construct degree sequence (G);
2: $G^*(V, E) = G(V, E)$.
3: DS^* ← anonymize degree sequence (G);
// show vertex set that needs to change its degree.
4: *while* G^* is not feasible *do*:
5: $DS_{diff} = (DS^* - DS)$;
6: S_{ops} ← Identify the operation type needed to satisfy the required degree: (edge ^{ins}, edge ^{del}, edge ^{swap});
7: *while* $S_{ops} \neq \{\}$ *do*:
8: EdgeList ← find the candidate edges to be modified;
9: EBC_List ← calculate betweenness centrality value for each edge (EdgeList, $G^*(V, E)$);
10: S_{aux_edges} ← EBC_List. *min_value* ();
11: run (ops, aux_edges, G^*);
// define new set of operations.
12: S_{ops} ← Identify the operation type needed to satisfy the required degree: (edge ^{ins}, edge ^{del}, edge ^{swap});
13: *end while*
14: *end while*
15: return G^*;

IV. Computational Results

In this section, we show the empirical results to assess the performance of our proposed algorithm. We will compare our method to the results of the two well-known approaches for *K*-degree anonymity. We change the value of *K* to vary from 2 to 10. The two methods are the KDA approach presented in [4] and the UMGA-NC approach proposed in [12]. We run all algorithms on the same dataset and the same configuration. Firstly, we show how far the structural properties of the graph can be conserved. Secondly, we measure how well our anonymization approach could preserve the community structure of the original graph.

A. Datasets and Environment

We test all algorithms on three real datasets which are unlabeled and undirected networks: these networks are Polbooks [33], American College football [13], and Jazz Musicians [34]. Table I shows the original properties of the three networks which include Diameter (D), Average path length (APL), Average Closeness (ACLN), Average betweenness (ABTW), and Transitivity (T). All experiments were tested on Google Colab on a PC with a 2.40 GHz i3 processor, 2 GB RAM, and a 228 GB hard disk running with Microsoft Windows 7 Ultimate. All experiments were implemented using python.

TABLE I. TESTED NETWORKS' PROPERTIES

	Polbooks	**American College football**	**Jazz Musicians**
$\|V\|$	105	115	198
$\|E\|$	441	613	2,742
D	7	4	6
$\overline{deg}$	8.40	10.661	27.697
APL	3.078	2.508	2.235
ACLN	0.329	0.399	0.457
ABTW	0.020	0.013	0.006
T	0.348	0.407	0.520
K	1	1	1

B. Assessment Measures

To assess the performance of our proposed approach compared to the others, we test four important measures that are used commonly in social network analysis. The four used measures are:

- *Average path length* (APL) is the average distance in the graph between every pair of vertices as described in Eq.3. Where V is the vertices set in the graph G, $d(u,w)$ is the shortest path length from vertex u to vertex w , and n is the vertices number in G.

$$APL(\mathrm{G}) = \frac{\sum_{u,w\epsilon V} d(u,w)}{\binom{n}{2}} \quad (3)$$

- *Closeness Centrality* (CLN) [35] is the Inverse of average distances to all reachable vertices. We calculate the Closeness of a vertex of u as follows:

$$CLN(u) = \frac{n}{\sum_{w\epsilon V} d(u,w)} \quad (4)$$

- *Betweenness Centrality* (BTW) [35] of a vertex u is specified as in Eq.5. $\sigma_{st}(u)$ indicate to the number of shortest-paths from the vertex $\boldsymbol{s}$ to $\boldsymbol{t}$ while $\sigma_{st}(u)$ is the number of the shortest-paths that go across u.

$$BTW(u) = \sum_{s\neq t\neq u\in V} \frac{\sigma_{st}(u)}{\sigma_{st}} \quad (5)$$

- *Transitivity* (T) is defined as the fraction of all triangles available in graph G. Available triangles are determined by triads number (two edges with a common vertex). We can compute the Transitivity of a graph G as:

$$T(G) = \frac{3(\text{number of triangles})}{\text{number of triads}} \quad (6)$$

To analyze the performance of our approach compared to the other two methods clearly, we evaluate the perturbation produced during the anonymization process of the four metrics listed above. As in Table II, we calculate mean absolute error (MAE) between the original and anonymized version of the tested networks over ten *K* levels as follows:

$$MAE(G, G^*) = \frac{\sum_{i=1}^{n} | g_i - {g_i}^* |}{n} \quad (7)$$

As ${g_i}^*$ is the value of the tested metric, e.g. (APL, $CC, \dots$), of the anonymized graph G^* at a particular level of *k*, g_i is the true value of the tested metric of the original graph G and n is the number of *K* levels.

C. Structural Analysis of the Perturbed Graph

In this section, we show the results of KDA, UMGA-NC, and our algorithm on the three networks listed in Table I. We calculate the four measures described previously for both the original graph and its anonymized version to show how much information is lost during the anonymization process. The actual metrics values of the original graph are constant for all different *K* values. They are represented by horizontal lines.

Fig. 4a, 5a and 6a show the average path length (APL) of the three anonymized networks as parameter *K* varies from 2 to 10. As we can see, the values of our proposed method are more similar to the actual ones than values of KDA, UMGA-NC, which means that lower information loss on APL.

Fig. 4b, 5b, and 6b refer to the average Closeness (ACLN) of the perturbed networks. All figures indicate that changes produced by our anonymization method on the average closeness also kept much closer to the real ones than existed by the two other methods.

Fig. 4c, 5c, and 6c describe the average node betweenness values. From the indicated figures, we note that our method could preserve the node betweenness values to become identical to the original values with varying anonymity parameter *K* in both football and Jazz Musicians networks. As for the Polbooks network, there are quite a few changes in the betweenness values.

Lastly, Fig. 4d, 5d, and 6d present the transitivity results on the three perturbed graphs. The performance of our proposed method comparing to the two other permutation methods isn't clear. We will quantify the performance of three permutation methods on transitivity obviously in Table II.

TABLE II. ERROR INDICATOR ON THE TESTED METRICS OVER 10 *k* LEVELS

Network	*Algorithm*	*APL*	*ACLN*	*ABTW*	*Transitivity*
Polbooks	*KDA*	0.349	0.042	0.003	0.023
	UMGA-NC	0.201	0.023	0.002	**0.014**
	Our Method	**0.134**	**0.015**	**0.001**	0.031
Football	*KDA*	0.017	0.003	**0.000**	0.012
	UMGA-NC	0.005	0.001	**0.000**	**0.005**
	Our Method	**0.002**	**0.000**	**0.000**	0.006
Jazz Musicians	*KDA*	0.064	0.011	0.004	0.020
	UMGA-NC	0.028	0.006	**0.000**	**0.014**
	Our Method	**0.019**	**0.002**	**0.000**	0.018

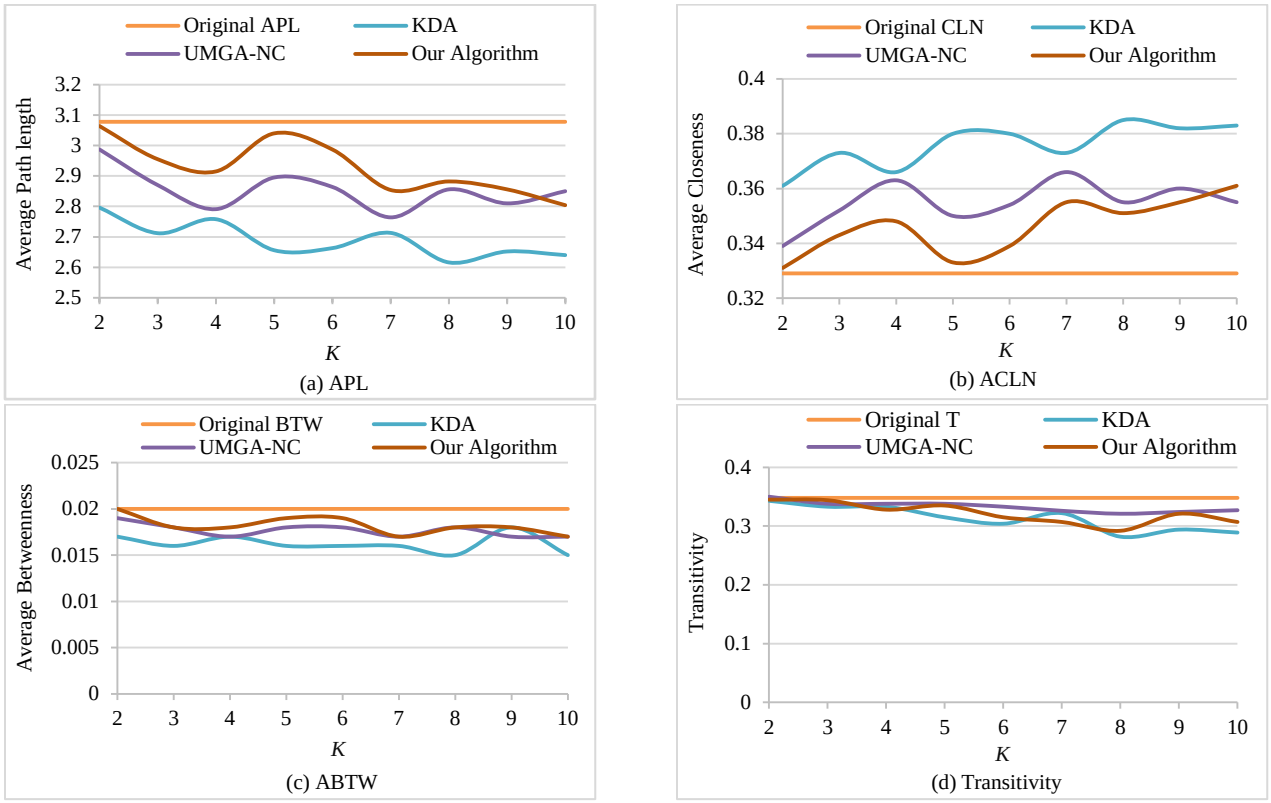


Fig. 4. Utilities of Polbooks Network for different K.

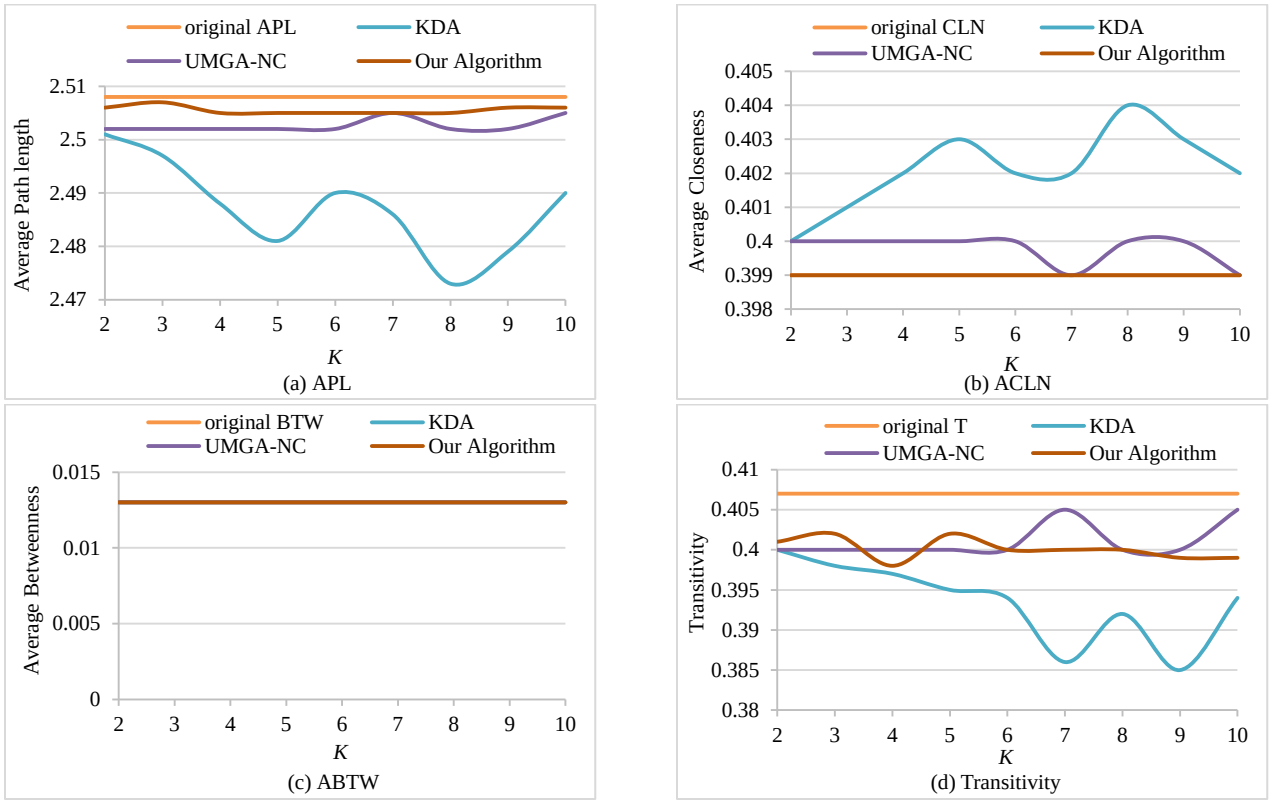


Fig. 5. Utilities of Football Network for different K.

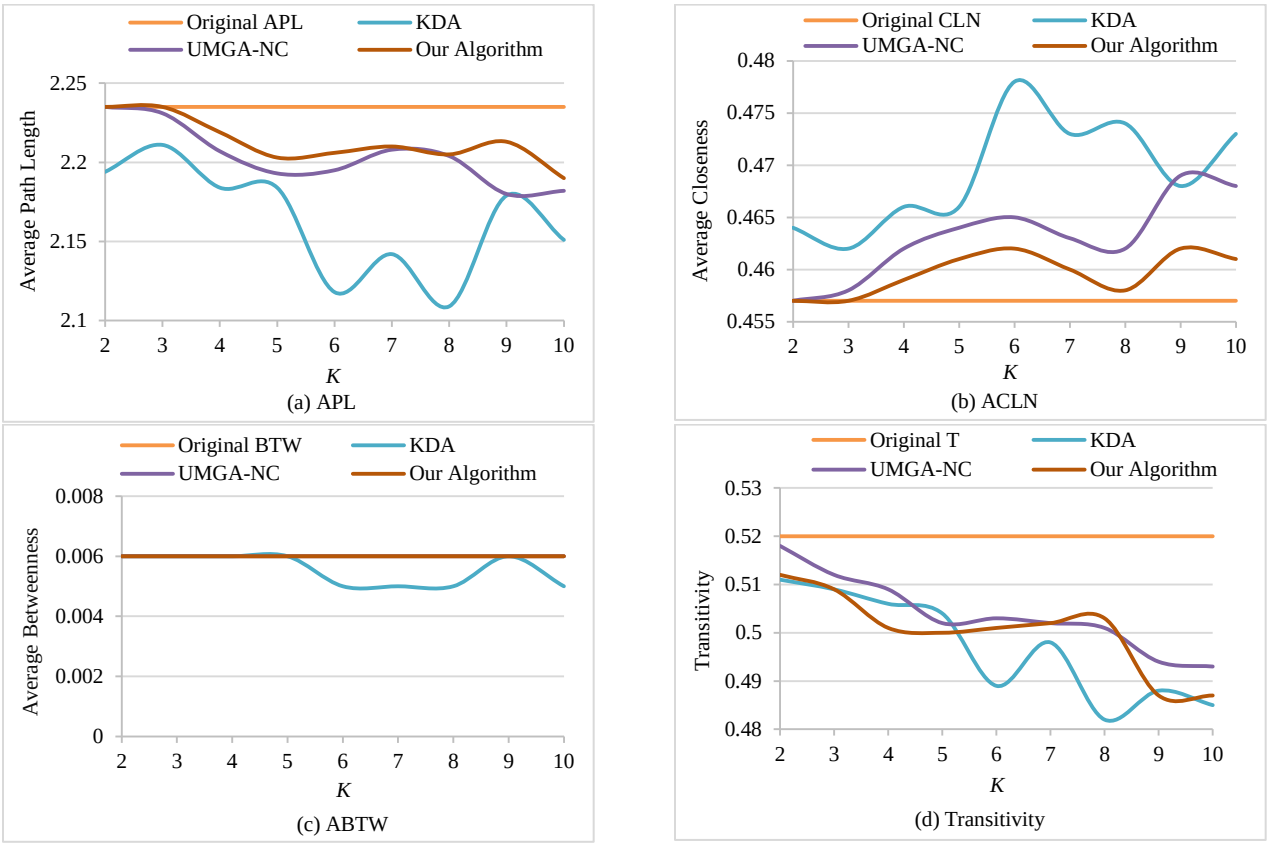


Fig. 6. Utilities of Jazz Musicians Network for different K.

We calculate the amount of error on the four tested metrics over 10 *K* levels as we referred to in Eq.7. As can be seen in Table II, our method gets the best results on the APL, ACLN, and ABTW except for the Transitivity results which are much affected in the anonymous graph. UMGA-NC method ranks first for the Transitivity metric on the three tested networks. Where the values of the mean computed error by the UMGA-NC method are lower than the ones obtained by both KDA and our method. As for KDA, our method achieves better results for Transitivity on both Football and Jazz Musicians.

D. Community Structure Preservation

Community detection algorithms are one of the most significant tasks for the processes of graph mining. This section will appreciate the utility of the perturbed graph for three different community algorithms. The three algorithms are (1) Girvan-Newman algorithm (*GN*) [13], which is is a hierarchical decomposition algorithm where edges deleted in descending order according to their edge betweenness scores. (2) Walktrap (*WT*) introduced in [36] is based on the concept of the random walk where the short random walks are likely to be kept in the same community. (3) Label Propagation (*LP*) proposed in [37], the main notion of the algorithm is to assign each vertex in a graph into a specific community, to which most of its adjacent vertices belong. For more details, see [38].

Using the networkX library, we extract community structure for both the original and the anonymized version of the three previous networks described in Table I. We use the f1-score measure [39] to assess the accuracy of our approach in preserving the actual community structure as described in Eq.8. This measure is used to test the similarity between the predicted communities set of the anonymized graph and the ground truth communities of the original version. We compute the f1-score values of *K*-anonymity for our algorithm and UMGA-NC using the three community algorithms. Then, we estimate the mean error on the f1-score over ten *K* levels. Table III presents the results.

$$f1-score = \frac{2\times recall\times precision}{recall+precision} \quad (8)$$

Whereas: $Precision = \frac{|C_P \cap C_T|}{|C_P|}$ and, $Recall = \frac{|C_P \cap C_T|}{|C_T|}$

Where C_T, is the vertices set that belong to the ground truth communities and C_P, denotes the set of vertices in the predicted communities produced by the community algorithm.

As shown in Table III, our method-EBC could present the lowest error on the tested networks using the three community algorithms. Consequently, a less information loss and better preservation for the community structure compared to UMGA-NC. Comparing the three community algorithms, The Girvan-Newman algorithm (*GN*) performs best on the three networks anonymized by our method. The reason behind this is that the Girvan-Newman algorithm (*GN*) is essentially based on the edge betweenness centrality values to detect communities, and our approach could preserve this metric well during the anonymization process.

TABLE III. MEAN F1-SCORE ERROR OVER 10 *K* LEVELS

	UMGA-NC			***Our Method-EBC***		
Network	***GN***	***WT***	***LP***	***GN***	***WT***	***LP***
Polbooks	0.058	0.073	0.344	0.029	0.030	0.192
Football	0.014	0.044	0.038	0.004	0.004	0.012
Jazz Musicians	0.042	0.442	0.044	0.021	0.309	0.035

V. CONCLUSION

Most of the previous works seek to anonymize graph data, regardless of the role of some edges that have proven their usefulness in analyzing the graph data. In this paper, we focus on optimizing the utility of an anonymized graph by minimizing the changes made to these edges. For this reason, we introduce the edge betweenness measure to identify and preserve the most relevant edges in the graph during the modification operation. Those edges, if modified, will cause large distortion to the local structure of the anonymized graph.

We perform an analysis using many structural metrics and different community algorithms on the graph structure. The final results proved that our method achieves the best performance as less information is lost comparing with other popular anonymization algorithms. Besides that, it can provide better preservation of the community structure compared to other similar methods.

In our future work, we plan to enhance the performance of our proposed approach. We intend to implement our algorithm on big data platforms to utilize graph computation systems such as GraphX on the Apache Spark platform and to test our proposed method on large graphs.

REFERENCES

[1] Nettleton, “Data mining of social networks represented as graphs,” Computer Science Review. 2013.

[2] Y. Mengmeng, Z. H. U. Tianqing, Z. Wanlei, and X. Yang, “Attacks and countermeasures in social network data publishing,” ZTE Communications, vol. 14, no. S0, pp. 2–9, 2019.

[3] C. Watanabe, T. Amagasa, and L. Liu, “PrIvacy risks and countermeasures in publishing and mining social network data,” 2011.

[4] K. Liu and E. Terzi, “Towards identity anonymization on graphs,” 2008.

[5] E. Zheleva and L. Getoor, “Privacy in social networks: A survey,” in Social network data analytics, Springer, pp. 277–306, 2011.

[6] P. Wicks, M. Massagli, J. Frost, C. Brownstein, et al., “Sharing health data for better outcomes on patientslikeme,” Journal of Medical Internet Research, 2010.

[7] L. Backstrom, C. Dwork, and J. Kleinberg, “Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography,” 2007.

[8] L. Sweeney, “k-anonymity: A model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557–570, 2002.

[9] B. Zhou and J. Pei, “Preserving privacy in social networks against neighborhood attacks,” 2008.

[10] B. Zhou and J. Pei, “The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks,” Knowledge and Information Systems, 2011.

[11] L. Zou, L. Chen, and M. Tamer Özsu, “K-automorphism: A general framework for privacy preserving network publication,” Proceedings of the VLDB Endowment, 2009.

[12] J. Casas-Roma, J. Herrera-Joancomartí, and V. Torra, “k-Degree anonymity and edge selection: improving data utility in large networks,” Knowledge and Information Systems, 2017.

[13] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” Proceedings of the National Academy of Sciences of the United States of America, 2002.

[14] P. Bedi and C. Sharma, “Community detection in social networks,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2016.

[15] K. Macwan and S. Patel, “Privacy Preservation Approaches for Social Network Data Publishing,” in Studies in Computational Intelligence, 2021.

[16] B. Ouafae, R. Mariam, L. Oumaima, and L. Abdelouahid, “Data Anonymization in Social Networks,” 2020.

[17] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” Foundations and Trends in Theoretical Computer Science, 2013.

[18] E. Zheleva and L. Getoor, “Preserving the privacy of sensitive relationships in graph data,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2008.

[19] A. Campan, Y. Alufaisan, and T. M. Truta, “Preserving communities in anonymized social networks,” Transactions on Data Privacy, 2015.

[20] E. Sargolzaei, M. J. Khazali, and F. Keikha, “Privacy preserving approach of published social networks data with vertex and edge modification algorithm,” Indian Journal of Science and Technology, 2016.

[21] C. Sun, P. S. Yu, X. Kong, and Y. Fu, “Privacy preserving social network publication against mutual friend attacks,” 2013.

[22] T. M. Truta, A. Campan, and A. L. Ralescu, “Preservation of structural properties in anonymized social networks,” 2012.

[23] J. Vadisala and V. Kumari, “Anonymized Social Networks Community Preservation,” International Journal of Advanced Computer Science and Applications, 2017.

[24] H. Wang, P. Liu, S. Lin, and X. Li, “A local-perturbation anonymizing approach to preserving community structure in released social networks,” 2017.

[25] J. Casas-Roma, J. Herrera-Joancomartí, and V. Torra, “A Summary of $$$$-Degree Anonymous Methods for Privacy-Preserving on Networs,” in Advanced Research in Data Privacy, Springer, pp. 231–250, 2015.

[26] J. Casas-Roma, J. Herrera-Joancomartí, and V. Torra, “Evolutionary algorithm for graph anonymization,” arXiv preprint arXiv:1310.0229, 2013.

[27] J. Casas-Roma, J. Herrera-Joancomartí, and V. Torra, “An algorithm for k-degree anonymity on large networks,” in 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), pp. 671–675, 2013.

[28] S. Chester, B. M. Kapron, G. Ramesh, G. Srivastava, A. Thomo, and S. Venkatesh, “Why Waldo befriended the dummy? k-Anonymization of social networks with pseudo-nodes,” Social Network Analysis and Mining, 2013.

[29] S. Hamzehzadeh and S. M. Mazinani, “ANNM: A New Method for Adding Noise Nodes Which are Used Recently in Anonymization Methods in Social Networks,” Wireless Personal Communications, 2019.

[30] S. Rajabzadeh, P. Shahsafi, and M. Khoramnejadi, “A graph modification approach for k-anonymity in social networks using the genetic algorithm,” Social Network Analysis and Mining, 2020.

[31] V. K. Sihag, “A clustering approach for structural k-anonymity in social networks using genetic algorithm,” 2012.

[32] S. L. Hansen and S. Mukherjee, “A polynomial algorithm for optimal univariate microaggregation,” IEEE Transactions on Knowledge and Data Engineering, 2003.

[33] V. Krebs, “polbooks | Miscellaneous Networks | Network Repository,” 2001. http://networkrepository.com/polbooks.php (accessed Apr. 08, 2021).

[34] P. M. GLEISER and L. Danon, “Community Structure in Jazz,” Advances in Complex Systems, 2003.

[35] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava, Anonymizing social networks. 2007.

[36] P. Pons and M. Latapy, “Computing communities in large networks using random walks,” 2005.

[37] U. N. Raghavan, R. Albert, and S. Kumara, “Near linear time algorithm to detect community structures in large-scale networks,” Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, 2007.

[38] N. R. Smith, P. N. Zivich, L. M. Frerichs, J. Moody, and A. E. Aiello, “A Guide for Choosing Community Detection Algorithms in Social Network Studies: The Question Alignment Approach,” American Journal of Preventive Medicine, 2020.

[39] G. Rossetti, L. Pappalardo, and S. Rinzivillo, “A novel approach to evaluate community detection algorithms on ground truth,” 2016.

Security Enhancement in Software Defined Networking (SDN): A Threat Model

Pradeep Kumar Sharma, Dr. S.S Tyagi
Department of Computer Science and Engineering
Manav Rachna International Institute of Research and Studies, Faridabad, Haryana, India

***Abstract*—Software Defined Networking (SDN) has emerged as a technology which can replace the prevalent vendor based proprietary CLI networking devices. SDN has introduced applications based network control and provided various opportunities and challenges for research and innovation in these networks. Despite many advantages and opportunities in SDN, security is a matter of concern for developers who want to invest in SDN. In this paper we are analyzing the SDN security issues with their countermeasures. We have generalized four use cases threat model that should cover security requirements of SDN. These use cases are: (I) protect controllers from applications, (II) inter-controller protection, (III) protecting data plane or switches from controller, (IV) protecting controllers from malicious switches. We found that these SDN components are inter-related if one is secure another one is already secure. We also compared the SDN and traditional network security in terms of these four use cases and provide the insights for protection mechanism and security enhancements. A framework for the development of a SDN security application has been presented based on ryu controller. We believe that our threat model will help various researchers and developers to understand current security requirements and provide a ready reference to tackle vulnerabilities and threats in this area. Finally, we identify some open research problems and future research directions with a proposed security architecture.**

Keywords*—Software defined networking (SDN); openflow; control plane; data plane; controller; programmability

I. Introduction

Traditional network (TN) devices are very powerful and provide various networking control functions in the form of routers, switches, firewall and load balancer etc. But security is always a big concern due to distributed nature of network containing various devices for various networking functions [1]. A lot of new models are being developed every year with more processing powers and updated software versions by the vendors and customer need to replace the previous hardware for getting new updated software functions. These proprietary devices are very costly and have their own way of configuration through CLI, having some specific commands and different vendors have different commands to communicate with these devices. This may results in configuration errors and various security breaches [2]. The output of these commands is as per human operator in mind and this output cannot be used further to provide programmability. Hence there is no scope for network engineers and researchers who want to scale and automate their network operations as per demands [3]. These hardware dependent systems, tightly coupled with software have failed to evolve the networking world as compare to system administration where software is independent of the hardware. In system administration, operating system is a piece of software which is not tightly coupled with hardware. We are free to install any operating system and applications on any hardware as per the requirement. As a result, system administration is evolving very fast. Today we can install many servers on a single hardware by using hypervisor, which manages several virtual machines with different host operating system. Not even hypervisor, Docker is another solution which provides high level resource utilization [4] as shown in Fig. 1 and 2 respectively.

In virtual machines concept as shown in Fig. 1, we assign dedicated processing resources and operating system to a VM image which is used by a dedicated service but Docker provides containers for hosting the specific services or applications which consumes very little resources as compare to virtual machine as shown in Fig. 2. One Docker engine can contain thousands of containers running various applications specific servers on a single operation system. On the other hand, in network administration we are still working on hardware dependent networking devices which consume a lot of processing power and time on manual configurations. There is a need to redesign the present networking architecture which can full fill the above said requirement with flexibility, programmability and automation.

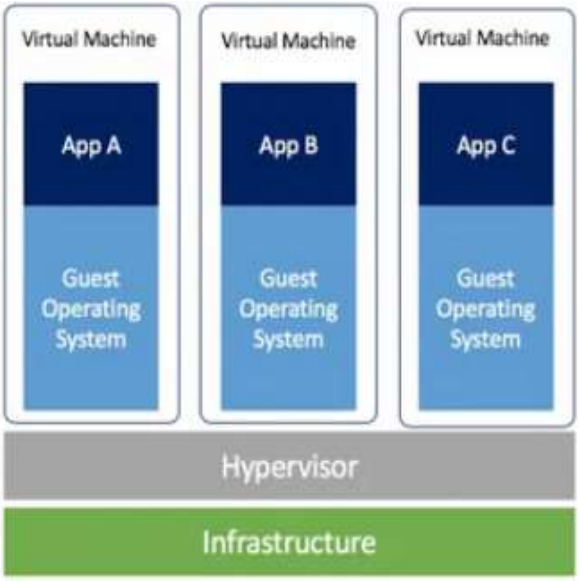


Fig. 1. Virtual Machines Hosted on Hypervisor.

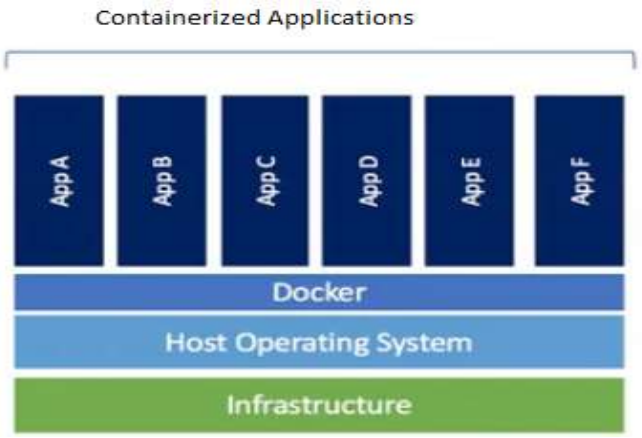


Fig. 2. Containerized Applications on Single OS through Docker.

Software Defined Networking [5] is a new concept which provides an API for configuration and decouples software logic from the devices. These devices work as simple data forwarding devices. The software or logical intelligence has been placed in a centralized controller. . The communication of forwarding devices and controller is established through a southbound API e.g. openflow [3]. All the networking functions like Routing, Security and Network monitoring etc. are done through the applications in application plane. The communication of application plane and controller is coordinated by northbound API e.g. RESTful API [6]. This provides the programmability approach and various applications can be designed a per the network demands. Network engineers can also use third party applications irrespective of hardware based solution for managing their network infrastructure. The idea of SDN is to use vendor specific hardware and we are free to choose software as per network demands irrespective of hardware. This arrangement of network functionality provides various opportunities for research and innovation in these networks. SDN is evolving and it has various advantages or traditional networks like dynamic control, programmability and a complete view of the network. As it is a new technology security solutions in SDN need to redefine and it provides various challenges and opportunities.

The rest of the contents of paper have been presented as under: Section II discusses the related works; a proposed threat model has been depicted in Section III. Comparative analysis of threats in SDN and traditional networks based on threat model has been elaborated in Section IV. Lessons learned and security enhancements by developing a security application have been discussed Section V. Section VI is dedicated to future research directions with a proposed security model. In Section VII we conclude our analysis with open research problems.

II. Related Work

Although there are several papers which provide various studies on SDN security, they do not focus on protecting SDN components from each other as SDN components are interlinked to each other and one component can attack the another component. If one component is malicious it can harm the other SDN components e.g. if application is malicious it can attack the controller and vice versa. Based on this concept we have derived four use cases to analyze the SDN security requirements and their counter measures. Also we have applied our threat model in traditional networks (TN) to analyze how these use cases are tackled in TN. We also provide a comparative analysis to find out the real threats in SDN and their possible resolutions.

In [7] Ali et al undertook a survey of related work in the area of SDN security. They presented programmable networks as an opportunity to improve protection in enterprise network through all the logical control at a centralized place. Real time policy enforcement and flexibility are presented as key tenets for controlling the behavior of network. They divided their study in two parts, one offering the innovative ways for finding the traffic anomalies, reaction to threats, flexibility in policy formation and deployment. Second part of the work provides security mechanism build up using SDN analytics which can be applied to the networks in real time. But they do not discuss the SDN security issues and its comparison with TNs. Dacier et al [8] discussed the current security challenges and showed how the traditional network architecture cannot fulfill the today's network demands. They discussed the various opportunities and challenges for security advancements in SDN. But they did not provide any way or model for SDN threat analysis and their resolution. B. Ahmad et al [9] discussed about Flow Table Entry Attack (FTEA), a kind of DoS attack, when Flow Entry Table gets full it drops the incoming packets or remove the prior flows. They assume that the attacker has access to SDN domain and consumes the controller resources by constantly engaging it to install attacker initiated bogus entries in the FET. However it exhibits only switches attack controller use case. Our work covers the four most important use cases for SDN threat analysis and their countermeasures.

III. Proposed Threat Model

Based on the SDN architecture we have derived a threat model which reflects how the various threats can attack the SDN components. SDN components are interlinked with each other if one component is compromised; it is a threat to another component and even for whole network. Our goal here is to identify the various attacks which can be performed by the attacker on a particular component of SDN. These components are SDN applications (Application Plane), controllers (Control Plane) and networking devices e.g. switches (Data Plane). In Fig. 3 we have shown the block diagram of SDN featuring its components. Based on this architecture we have derived four use cases to analyze the threats.

Threat Model:

There are many ways to exhibit SDN security issues and their resolutions [10][11]. Most of the authors discuss the same with layer based approach but we believe SDN architecture is different aspect from conventional network and we define a new taxonomy to generalize the SDN security issues. We consider a network scenario where there are n no. of controllers $C = \{c1,c2, \ldots\ldots c_n\}$. Each controller ci □ □ C can run at least one application from a set of applications $A^{ci} = \{a1.a2, \ldots\ldots a_n\}$. Each controller has limited resources which make

them vulnerable to denial of service attacks. We have derived four use cases from SDN architecture. Each of the use case has its own importance and security goals. Fig. 4 shows the Threat model for security requirement of SDN. SDN architecture with associated use cases is shown in Fig. 3 and 4 respectively. A semi benign attack is a passive attack which may gather information about network or processes but will not deviate from protocol execution. A malevolent behavior is an active threat which may deviate from protocol rules in order to disrupt the system and attack the other components of the system [12][13]. These four use cases are described as under.

A. Use Case 1: Securing Controller from Applications in Application Plane

In this use case each application in the application plane can be benign, semi benign, or malevolent. These applications may be from different sources i.e. third party apps [14]. The controller proffers an abstraction to application plane so that application can read/edit network state which is generally a degree of network control. If an attacker impersonates application it can gain access to controller i.e. network control and can hamper the network operations [15]. The absence of trust and weak authentication between applications and controllers may lead to spoofing attacks [16][17]. Our goal here is to minimize the attacks on controllers through applications. List of such type of attacks and suggested solutions have been shown in Table I.

B. Use Case 2:Inter Controller Security

In SDN, control is logically centralized. It provides more than one controller for providing scalability and avoiding single point of failure [18]. As a result these controllers share the resources and communicate with each other. It is necessary to review the security of inter controller communication [19]. In this use case we assume one or more controller is semi benign or malevolent. A semi benign controller could be able to access the control data of other controllers, learn resource utilization information and target the integrity of the network. Moreover a malevolent controller can attack to semi benign controller and perform a DoS attack on another controller. Our goal is to protect controller from each other [20]. The possible attack scenario and solutions have been discussed in Table II.

C. Use Case 3: Securing Switches from Controller

In this use case it is assumed at least one controller is semi benign or malevolent. We assume that applications which are used through this controller can be semi benign or malevolent. A semi benign controller can target switches in the data plane. It can attack switch flow table with buffer overflow by sending bogus entry [21]. Our goal here is to eliminate the possibility of controller's ability to target the switch with bogus entry [22]. This case has been shown in Table III with threats and their solutions.

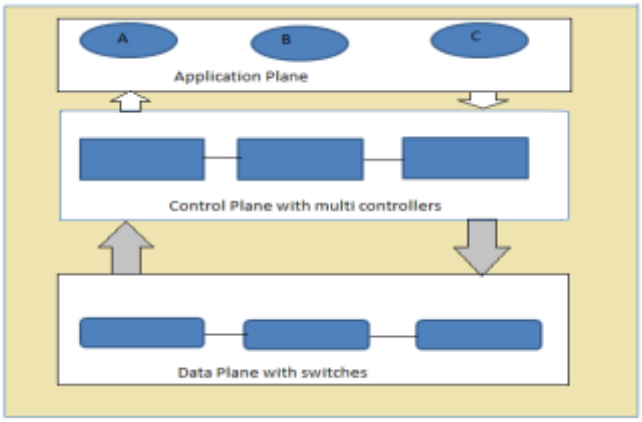


Fig. 3. SDN Architecture.

Usecase	Apps can be			Controllers can be			Switches can be		
	benign	semi benign	malevolent	benign	semi benign	malevolent	benign	semi benign	malevolent
1. Securing Controller from applications	✓	✓	✓	✓	×	×	✓	×	×
2. Inter-controller Security	✓	×	×	✓	✓	✓	✓	×	×
3. Securing switches from controller	✓	✓	✓	✓	✓	✓	✓	×	×
4. Securing controller from switches	✓	×	×	✓	×	×	✓	✓	✓

Fig. 4. Threat Model.

TABLE I. SECURING CONTROLLERS FROM APPLICATIONS (USE CASE 1)

Issues	Possible Attack	Scenario	Possible Solution
Application vulnerabilities	Remote code alteration and execution	An attacker can reprogram the application by using vulnerability and run the malicious code.	Periodic vulnerability scanning for applications
Use of Untrusted applications	Spoofing the messages between controller and applications	Absence of trust between controller and applications may lead to spoofing attack	Apps authentication and authorization must be implemented.
Inappropriate authorization	Unauthorized access to applications	If an application has weak authorization an attacker can gain unauthorized access to application and can attack the controller.	Use of AAA to protect the unauthorized access to application and controller

TABLE II. INTER-CONTROLLER SECURITY (USE CASE 2)

Threat	Possible attack	Possible scenario	Possible Solution
Untrusted Controller	Attack controller within cluster	Untrusted controller software can pose a serious threat to other controllers.	Use the trusted controllers provided by the trusted vendors.
Controllers configuration defects	Unauthorized access and network attacks	Providing unnecessary privileges to an app can result in controller hijacking.	To review the default configuration of controller and to do a proper controller hardening.
Embedded malware	Malware and spyware attacks	Spyware and ransomware gaining control to controller	Monitoring and scanning the network with trusted security applications.
Vulnerabilities within controller runtime	Controller runtime attacks	In case of third party network applications, vulnerabilities at run time can allow applications to modify its default configuration.	Controller's software should be updated periodically with new patches and updated versions.
Poorly separated inter-controller traffic	DOS attack, ARP spoofing	In a multi controller environment If traffic between controllers is poorly separated than it can allow a compromised controller to perform a man in middle or DoS attack.	Controllers in cluster should not be provided unnecessary permissions and should be monitored as per cluster rules

TABLE III. SECURING SWITCHES FROM CONTROLLER (USE CASE 3)

Threat	Possible attack	Possible scenario	Possible Solution
Controller switch communication channel.	Flooding attack on switch flow table.	A compromised controller can flood a switch by bogus entries by sending fake packets to switch.	The best way to secure the controller and switch communication is use of TLS.
Malicious applications	Attack on switches due to malicious application	An infected application may affect the controller and attack the switch through misconfiguration	Use of trusted and stable application for performing the network operations in SDN

D. Use Case 4: Securing Controller from Switches

In this case at least one switch is semi benign or malevolent and it tries to attack the controller [23]. An attacker can send fake message through this compromised switch to controller and tries to exhaust the controller's resources [24]. This condition is called as data leakage where attacker tries to discover the flow rules and forwarding policy information. If an attacker can gain access on packet processing timings and can determine the action related to specific type that are forwarded to controller, attacker can produce the phony flow messages causing to DoS attack [25][26]. Our goal here is to protect controller from switches. If a switch goes out malevolent or semi benign there should be a mechanism to find out the malicious switch in the network. One of the recent researches towards malicious switch detection has been presented in [27]. The authors presented a new algorithm to find a pernicious switch based on control path routing approach. This method chooses two node disjoint control path for every forwarding device in data plane so that a suspicious node can be find out on basis of simple Packet_In messages delivered to control paths. In [28] a novel technique for detecting the link flooding attack has been presented. Authors designed LFA defense system called LFADefender using SDN which contains features like programmability, complete view of network and flow traceability. In LFADefender, authors proposed a LFA target link selection approach and design a LFA congestion monitoring mechanism to effectively detect LFA.

IV. COMPARATIVE ANALYSIS OF SDN THREAT MODEL USE CASES WITH TRADITIONAL NETWORK

Based on the above use cases we have identified the attack scenario of various threats in SDN. Now we will compare the same with traditional network architecture to find out that; are these use cases available there in traditional networks? We will Fig. 5 out if these use cases are available in traditional network how we counter them. Then we will identify the SDN protection mechanisms [29][30] based on this. Traditional network architecture with four routers and two switches has been shown in figure. In Traditional Network (TN) the interface between two routers is called network to network interface (NNI) while the router interface with end user is called user network interface (UNI).

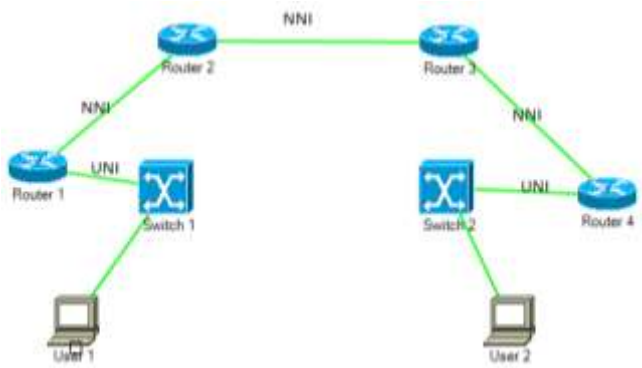


Fig. 5. Traditional Network Architecture.

The fundamental difference between SDN and TNs is control plane [31]-[32]. In TNs network controlling elements are inside the network devices e.g. routers and switches but in SDN it has been decoupled from the devices to a central controller. From the previous use cases in Section II, we can derive that controller security is most important and it can be attacked by applications, by switches in data plane and can even by other controller in multi controller environment [33]. In this section we will find how the network controlling elements are protected in tradition networks. What types of the attacks are faced in TNs and what are the protection mechanisms. We will try to analyze four use cases in TNs

which we implemented in SDN in above section and it will give a clear picture of security problems and challenges in SDN with their possible solutions.

A. Use Case 1: Application Attacks Controller or Controlling Elements

If we talk about the application layer of SDN in terms of traditional network and its applications like mailing, FTP and HTTP etc. then we realize that these are the network functions for which lower layers in TCP/IP model have been designed. We cannot compare these functions with applications in application plane in SDN [34]. Applications in SDN are network controlling elements which have been decoupled from devices like routing, switching, security etc. and decoupled control functions works as applications and perform related network operations in coordination with controller. In TNs applications are the part of network devices and controlling part resides inside the devices. Hence application attacks controller use case 1 does not apply in TNs. However TNs do not provide any programmability to control the behavior of network dynamically like SDN which provide this functionality through applications in northbound API with an alternative to use third party network applications to customize the network as per demands [35].

B. Use Case 2: Controller Attacks other Controller

A controller in SDN performs the network control functions like routing switching etc. In TNs a routing function is performed by routers. A multi controller scenario in SDN can be compared with TNs having multiple routers. In network containing more than one router and various links from one router to another, a routing protocol is used to find the best path from source to destination. There are two types of protocols one distance vector routing protocols Routing Information Protocol (RIP) which find the best path to a remote network by judging distance. Second is link state routing protocols e.g. Open Shortest Path First (OSPF). These routing protocols maintain a routing table and contain the information about the neighbor subnets and links state. The router updates it routing table and advertise the routing information time to time and a best path is selected based on this information. However different routing protocols suffer from different attack methods but the common objective is to pollute routing tables. A routing table poisoning attack is performed to contaminate routing table and network topology information by advertising or infusing a bogus route through announcements. A malevolent router can publish a phony link state advertisement (LSA) with fake link cost to effect rest of the routers routing table calculations. This type of attacks are not difficult to dispatch but has limited influence as adjoining router will publish a right LSA with a new sequence entry which will remove the bogus LSA and it will not be used again for routing table route estimation. However a more powerful attack can also be performed to pollute the functioning routing table. To mitigate routing table poisoning, a routing protocol should only run on NNI and route advertisements from UNI need to be discarded. The origin of message should be checked for authentication to forestall a vindictive router to imitating another router. And routing updates need to be double checked before applying them for route estimation by routing tables. For example a link metric updated by one router need to be double checked with link metric updated by other router on the same link. Such type of defense techniques [36] can also be considered in SDN and the same has been discussed in Section II.

C. Use case 3: Controller Attacks Switches

In traditional network we can refer the L3 devices as control plane. Now we will try to find out, Can a router attack on the functionality of L2 devices? When a router sends a packet to another router, the receiving router performs three operations. First it eliminate the L2 header of packet, second check the routing table for next router in sequence and third bundles L2 header of packet for sending to next node. The next node in the routing table might be a local interface or it is an IP address. This routing process is continued till the next node is a local interface. Upon finding a local interface it will looks up for MAC address in ARP table of that interface. If it is unable to get the MAC the router will run the ARP protocol to get the MAC address associated to respective IP address. Because address resolution protocol (ARP) is used by router to find the MAC address associated with an IP address, an L3 router may suffer with ARP cache poisoning attack [37]. In this attack the attacker associates its own IP address with victim MAC address and receives the traffic intended to for victim node. So an attack can be placed from control devices i.e. from control plane on L2 traffic in traditional networks.

D. Use case 4: Switches Attack Controller

In SDN there is no by default communication when an open flow switch receive a new packet it sent it to the controller using Packet_In message, which includes source and destination address. If destination MAC address is not known by the controller then controller asks the switch to broadcast the packet through Packet_out message. The destination sends the response to the source port and this reply also noted by controller to fulfill the further requests from same source and destination. This process is called host tracking service and is equivalent to L2 MAC learning, in principle, with only difference that MAC learning has been separated from switch and included in controller. In SDN data plane switches communicate with controller for L2 learning process and can attack the same as discussed in Section II. But in TNs L2 learning is implemented inside the switches without any controller. So attack on L2 learning is equivalent to attack on controller from switches. A MAC table is learned from data plane switches including host packets so it is subjected to MAC attacks. An infected host can send a packet with fake MAC address to poison switch's MAC table. MAC spoofing and MAC flooding are two strategies of attacks which affect the L2 learning in traditional network [38]. Threats related to MAC address can be minimize by disallowing the unknown devices to enter the network. This can be done by a switch feature port security. Port security is a technique which allows only known MAC addresses (MAC binding) to be recognized by the network switches). But there is a limit to bind the number of MAC address associated with a switch port. But port security requires a lot of manual configurations which leads to possible overhead and misconfigurations [39].

V. Lessons Learned and Security Enhancements in SDN

We have compared four most important attack use cases in SDN and traditional network. We have seen how the control functions in TNs can be attacked in different use cases as compare to attacks on controller in SDN [40]. Table 4 shows the comparison of threat use cases in SDN and TNs. Now we will discuss the lessons learned on comparing these use cases in terms of threats and their defenses. We will explore how the security can be enhanced [41] in SDN based on our threat model. First we will elaborate each use case then a security application will be developed based on attacks from above use cases.

A. Use Case 1: Securing Controller from Applications

As we have discussed in previous section that network control functions are part of network devices in traditional network hence this case, does not apply on TNs. In SDN network control functions are in the form of applications and have been decoupled from network devices. These applications work for the data plane devices in coordination with controller [42]. As a matter of fact these applications communicate with controller to fulfill the network requirement and an unauthorized application can do a big damage to the controller and even reconfigure the network [43]. In order to counter an unauthorized application access controller and application should maintain a trusted connection and authenticate the identity of entities before exchanging control messages. Both authentication and authorization of applications is to be ensured before establishing a connection. This concern about the untrusted applications authentication and securing the controller has been discussed in [44]. Authors introduced a hierarchical arrangement of controllers. This hierarchical system can minimize the effect of pernicious application as code of the application would run at the middle hierarchy where there will be ample protection. Another work in this direction is FortNox [45]. FortNox is an extension to the open source controller NOX [29]. It is a security enforcement kernel which checks the flow rules for security policy violation in real time. Each openflow application is provided authorization through a role based authentication concept. Three flow rule producer roles are defined; OF Operator, OF Security, and OF Application. In case of any flow rule conflict detected by FortNox, a higher priority rule is accepted. The limitation of FortNox is application identification and priority enforcement. ROSEMARY [46] is the enhancement to controller resilience to malicious applications. It is a high performance network operating system which is robust and secure. It sandboxes the each running instance of application to provide security to control layer from any vulnerability. It also monitors and control the resources consumed by each application. In LegoSDN [47] authors explore about the effect of application failure on controllers reliability. Authors proposed a isolation layer between controller and applications to avoid the consequences of failure of controller due to application failure.

B. Use Case 2: Inter-controller Protection

In SDN to avoid the single point failure a multiple controllers has been suggested. There are two types of controller placement schemes; one is flat controller deployment and another is hierarchical controller deployment. In flat controller concept each controller is assigned a separate sub network. In this solution different operations may not be able to communicate equally with different domains. But in hierarchical mode the local controller is responsible for respective network, and global controller is responsible for local controller. The communication among different controllers is done via global controller. A variety of works has been done towards controller placement problem. In [48] authors proposed an algorithm to find the minimum number of controllers and maximum load on a controller. But this arrangement did not work for the request with variable time. In [49] author proposed an algorithm which divide the network in to different subnets. Every small network contains a controller based on the size of assigned network. It uses a clustering algorithm based on switch density, and divides the network accordingly. When the main link is broken it may use a backup link. But it may result in unnecessary delay. In [50] authors provide a multi controller solution with Byzantine fault tolerant mechanism. When one controller goes down, the other controller takes the charge of network and removes idle link of previous controller. However this solution is good for small network due to performance issues in relatively large networks.

C. Use Case 3: Protecting Switches from Controller

In traditional network, a control element router can attack the switch functionality through the ARP spoofing attack as discussed in Section III. But in SDN controlling element controller has more functionality and a malicious controller can do a lot of damage to the switches of data plane. A compromised controller can attack the switch flow table by generating unnecessary broadcast and overflow the switch flow table. So protecting the controller to become malicious is the main defense for data plane switches. In [26] authors proposed a solution for detecting the malevolent SDN device in the network. They implemented a backup controller and collect the state information and updates from primary controller and switches. They detect the malicious devices by recognizing the unexpected and inconsistent behavior of primary controller, backup controller and SDN switches.

TABLE IV. Comparison of Attacks use Cases

Use Cases	SDN	Traditional Network
Applications attack controller	A malicious application can attack controller.	Not applicable
Inter controller attack	A compromised controller can attack the other controller in a multi controller environment.	A router can attack the other router and can pollute its routing table by fake LSA.
Controller attacks switches	A malicious controller can attack the data plane switches and launch flood attack.	A controlling element router can attack the L2 switches by ARP cache poisoning attack.
Switches in data plane attacks on the controller	A malevolent switch can flood the controller by fake flows may results in DoS attack.	In L2 network ,the control function MAC learning, can be targeted by MAC spoofing and MAC Flooding attacks.

D. Use Case 4: Protecting Controller from Switches

The main protocol which provides the interface for communication between data plane switches and controller is open flow. In respect to southbound interface communication, the open flow switch specifications discuss necessity of TLS with mutual authentication between controller and switches [51]. In [52] the lack of TLS adoption is real world deployments has been discussed. This is very important consideration when switches, controller and application environment is deployed in trust domains. However it is to be noted here that there is definite weakness introduced by separating the control and data plane in SDN [53]. Various solutions to avoid the DoS attack have been proposed. AVANT-GUARD [48] provides the protection to the controller from switches by limiting the number of flow requests sent to the controller by using a connection migration tool. This migration tool removes failed TCP sessions at the data plane prior to any notification to controller. This prevents the occurrences of DoS attack by sending only those flow requests to the controller which completes the TCP handshake.

E. Security Enhancements

By the comparisons and discussion in the last two sections it can be stated that there is a need to develop a security mechanism to counter the security issues of SDN. As discussed that the controlling functions in the SDN are performed by the applications in application plane. For implementing the security functions there is a need to design the security application in SDN. In this section we develop a security application as a part of SDN software. In traditional network if we want to implement security functions then we need to use a hardware device e.g. firewall for the same. But this is advancement in SDN that network controlling functions like security, routing, and monitoring etc., are in the form of applications. For Design and implementation, we will use mininet as network emulator and Ryu as a controller. First we will focus basic steps and algorithm for designing an application as per controller and data plane communication.

Python language is used to develop the network applications based on Ryu controller. Ryu is a components based controller which has various modules for application design and control. In ryu controller setup at home/ubuntu/ryu it has various folders; app, base and ofproto. App folder can contain various applications like firewall, router and load balancer. Base folder contains App_manager which helps to run the different applications and prepares framework and datapath for running the application. Ofproto deals with openflow version related queries and matching capabilities. For designing a SDN application we need to collect and understand the initial requirements and booting process of SDN network framework.

- In first step switch boots up and contact the controller for openflow version related queries and check its capabilities.
- The controller installs Packet In function and table miss function and prepares itself for queries from switch.
- When receiving Packet In, Controller learns the source MAC and mention the MAC and port information in flow table. It checks for destination MAC address if it is available in flow tables, it uses Packet Out function on the port and installs the flow and stores the same for future uses.
- If destination MAC address is not available in flow table i.e. a table miss then controller uses packet out function to broadcast the packet to all ports.

By using the ryu controller framework we can design and deploy customized security applications. With programmability approach in SDN we can have our own security application in ryu app folder and program it as per network demands and configure it through standard API. Traditional security solutions, the vendor specific e.g. fortigate and Cisco, they have their own proprietary code and configuration methods which are fixed and cannot be customized as per demands. Fig. 6 shows how the security app can work in coordination with controller. When Host A wants to communicate to Host B it sends a packet to switch. Switch check for a matching entry in its flow table but when a matching entry is not found in flow table then packet is forwarded to controller. Controller sends the packet to security application for policy check. First it parses the packet and check if it matches to policy specified in firewall. As firewall has a policy to block traffic from A to B (A-->B: Block). The application enforces a rule through controller to drop the packet and controller install a flow rule in switch flow table to drop all the incoming traffic from Host A to Host B. This is how we can block and allow flow in openflow through a security application. It means through this app a switch can work like a firewall i.e. technology allows us to decide the functions of a switch. As a result additional security devices are not required in SDN as security services can be enabled within the devices. In traditional network another problem is placement of firewall for optimized coverage of security services. But it has been nullified as any device in the network can be turned into a security device.

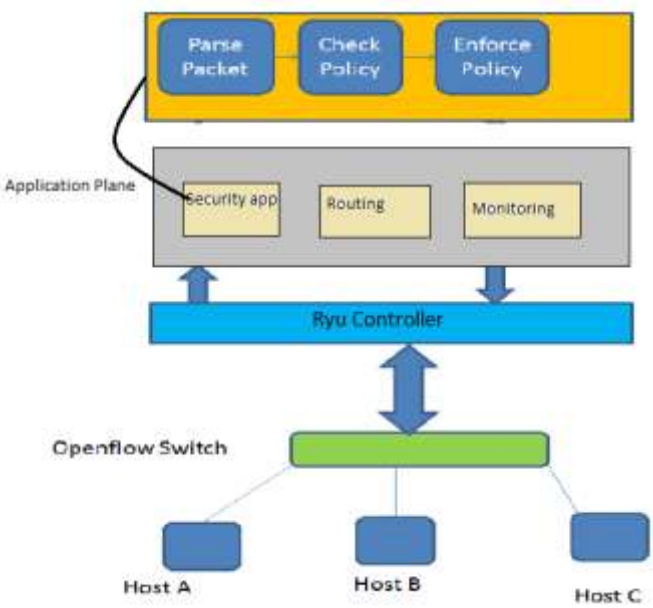


Fig. 6. Implementing Security Application in SDN.

VI. Future Research Directions and Proposed Security Architecture

Based on use cases presented in the paper and comparative analysis with traditional network it may be admitted that SDN has introduced some new security issues and challenges that are not available in TNs. But programmable networks include a level of adaptability and dynamic control that has enhanced the network management and flexibility at scale. SDN has emerged as a technology which may be taken as a replacement of vendor based proprietary CLI networking devices where there is no scope of programmability and automation due to tightly coupled software and hardware concept of networking devices. So it is to be noted that SDN is going to be stay here and some more potential research proposals are required to address the challenges of SDN security issues. Implementation of mandatory TLS functionality a controller and deplane communication channel can solve a lot of problems. Research proposals like AVANT-GUARD [48] has provided a good work by limiting the rate of number of requests sent to controller which improve the controller performance. Implementing some intelligence function to data plane switches may be considered to minimize the controller load. Such type of proposals is under discussion with research community in the form of stateful data planes [54].

Another important area which needs the momentum in research proposals is application-controller interface [55]. Without the presence of a standard open north bound API, it is not possible to design and deploy SDN in enterprise network [56]. The security enhancements explored in Section IV are of no means if application-control interface is vulnerable. This can also be evaluated from Table I of our threat model analysis that the switches in data plane can be attacked if either the applications or controller are malicious. In contrast we have discussed the various innovative proposals which analyses the protection requirements of north bound API. However this use case (securing controller from application) exhibits a lot of vulnerability to various attacks as discussed in Section II. As a results further research in this area are necessary and need to be encouraged for finding a better northbound API. However use of RESTful API [6] is also a good work and this may be extended further. A multi controller solution for addressing the scalability issue of controller has been provisioned in openflow 1.3. Various controllers need to communicate with controller in other domain for performing various operations to fulfill the network requirements [57]. A secure and real time communication of controllers is an open research problem. However a number of solutions have been discussed in section Vth in view of further research directions in this area [58]. A framework for network security application development has been presented based on ryu controller and mininet. This work can also be further explored by adding more security functions if we have a new idea and algorithm as per demands.

A. Proposed Security Architecture

By threat model analysis it can be pointed out that SDN security problem is not a problem of single SDN component, it is scattered in all components of architecture and these components are inter linked with each other and form a system. So there is a need to design a security solution as per system perspective rather than security for individual component. Based on our analysis security architecture for SDN has been proposed in Fig. 7.

Control-application interface is protected with AAA security at application plane. We believe each application should be developed as a module of controller so that it can easily follow the security standard of northbound interface designed to secure the communication. Even third party applications should follow and support the security policy standard at application-controller interface. A multi controller solution with hierarchical control is provided to avoid the single point failure of controller and resource sharing. A backup controller has also been proposed for global controller fault tolerance. At southbound API the communication of controller and data plane switches should be secured with mandatory TLS security function. For minimizing the load of controller some state level intelligence is suggested in data plane switches i.e. stateful data plane [53]. However the management of states and packet level forwarding decisions are taken from controller. Finally it is to be stated here controller security is the prime tenet to secure overall SDN platform and this is ultimately depends of secure applications environment at northbound API. Development of a standard northbound API is still an open research problem. Contribution haven been made in the form of RESTful API but a more research proposals are required in this direction to form a more secure SDN network.

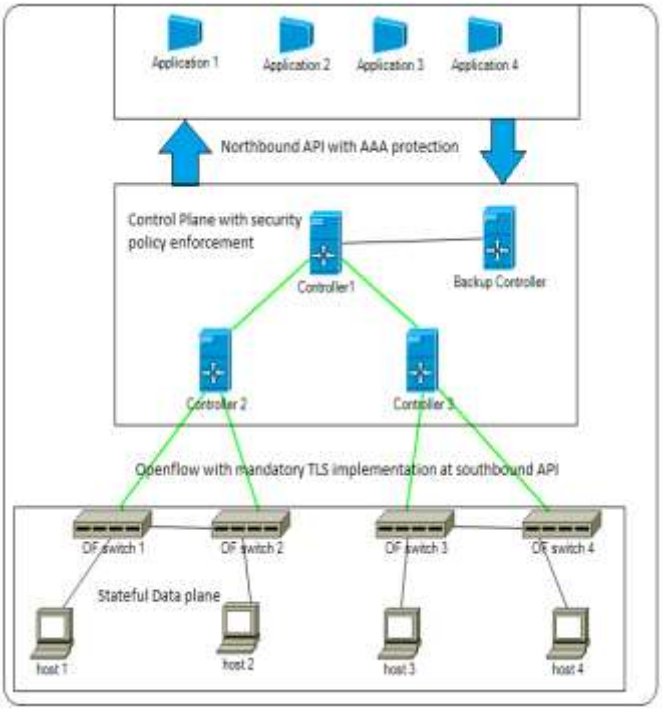


Fig. 7. A Proposed Security Model for SDN.

VII. CONCLUSION

To identify the SDN security issues we developed four use cases and discussed the several attack parameters with their counter measures in tabular format. After identifying the security issues we applied the same use cases in traditional network for a comparative study of risk and security technology in both the networks. After comparative study it can be concluded that SDN has introduced new attack surfaces which is not available in traditional networks. In contrast SDN provides more flexibility, automation and control over the network, traditional networks disappoint there. However security solutions to address the SDN security issues have been presented which includes protection from malicious application, inter-controller protection, protection of data plane and protecting controller from DoS attacks by data plane switches. Based on analysis a framework for development of SDN security application has been presented with ryu controller and mininet network emulator. Insights for security enhancement have been provided by presenting a proposed security model based on recent research and threat model analysis. Moreover research in SDN security is still in beginning stage and there is lot more to do with. By designing novel security techniques and extending the previous research work for solving known problems, we can find the better SDN networks which will be much more secure than traditional networks.

REFERENCES

[1] M. Casado et al., "SANE: A protection architecture for enterprise networks," in Proc. USENIX Security Symp., 2006, p. 10.

[2] M. Casado et al., "Ethane: Taking control of the enterprise," in ACM SIGCOMM Comput. Commun. Rev., vol. 37, no. 4, pp. 1–12, Oct. 2007.

[3] N. McKeown et al., "OpenFlow: Enabling innovation in campus networks," ACM SIGCOMM Comput. Commun. Rev., vol. 38, no. 2, pp. 69–74, Apr. 2008.

[4] R. R. Yadav, E. T. G. Sousa and G. R. A. Callou, "Performance Comparison between Virtual Machines and Docker Containers", IEEE Latin America Transactions, VOL. 16, NO. 8, AUG. 2018, pp. 2282-2288.

[5] S. Jain et al., "B4: Experience with a globally-deployed software defined WAN," in Proc. ACM SIGCOMM Conf., 2013, pp. 3–14.

[6] Li Li, Wu Chou, Wei Zhou and Min Luo, " Design Patterns and Extensibility of REST API for Networking Applications", IEEE, TNSM, 2015, 00814.

[7] S. Taha Ali et. al "A Survey of Securing Networks using SDN", IEEE transactions on reliability, Vol 64, No. 3, 2015.

[8] Marc C. Dacier et al, "Security Challenges and Opportunities of Software Defined Networking", in *IEEE Computer and Reliabilities Societies*, 2017, pp.96-100.

[9] B. Ahmad et al. "Fingerprinting SDN policy parameters : An Empirical Study", IEEE Access, Volume 8, 2020.

[10] D. Li, X. Hong, and J. Bowman, "Evaluation of security vulnerabilities by using ProtoGENI as a launchpad," in Proc. IEEEGLOBECOM, 2011, pp. 1–6.

[11] S. Shin and G. Gu, "Attacking software-defined networks: The first feasibility study," in Proc. 2nd ACM SIGCOMM Workshop Hot Topics Softw. Defined Netw., 2013, pp. 165–166.

[12] L. Schehlmann, S. Abt, and H. Baier, "Blessing or curse? Revisiting security aspects of software-defined networking," in Proc. 10th Int. CNSM, 2014, pp. 382–387.

[13] S. Sezer et al., "Are we ready for SDN? Implementation challenges for software-defined networks," IEEE Commun. Mag., vol. 51, no. 7, pp. 36–43, Jul. 2013.

[14] W. Han, H. Hu, and G.-J. Ahn, "LPM: Layered policy management for software-defined networks," Data and Applications Security and Privacy XXVIII. Berlin, Germany: Springer-Verlag, 2014, pp. 356–363.

[15] X. Wen, Y. Chen, C. Hu, C. Shi, and Y. Wang, "Towards a secure controller platform for OpenFlow applications," in Proc. 2nd ACM SIGCOMM Workshop Hot Topics Softw. Defined Netw., 2013, pp. 171–172.

[16] S. Scott-Hayward, C. Kane, and S. Sezer, "OperationCheckpoint: SDN application control," in Proc. 22nd IEEE ICNP, 2014, pp. 618–623.

[17] P. Porras, S. Cheung, M. Fong, K. Skinner, and V. Yegneswaran, Securing the software-defined network control layer," in Proc. NDSS, San Diego, CA, USA, Feb. 2015, pp. 1–15.

[18] P. Berde et al., "ONOS: Towards an open, distributed SDN OS," in Proc.3rd Workshop Hot Topics Softw. Defined Netw., 2014, pp. 1–6.

[19] M. M. O. Othman and K. Okamura, "Securing distributed control of software defined networks," Int. J. Comput. Sci. Netw. Security, vol. 13, no. 9, pp. 5–14, Sep. 2013.

[20] F. Botelho, A. Bessani, F. M. Ramos, and P. Ferreira, "On the design of practical fault-tolerant SDN controllers," in Proc. 3rd EWSDN, 2014, pp. 73–78.

[21] H. Mai et al., "Debugging the data plane with anteater," ACM SIGCOMM Comput. Commun. Rev., vol. 41, no. 4, pp. 290–301, Aug. 2011.

[22] Ahmad, B. et al., "Fingerprinting SDN policy parameters : An Empirical Study", IEEE Access, Volume 8, 2020.

[23] C. Jeong, T. Ha, J. Narantuya, H. Lim, and J. Kim, "Scalable network intrusion detection on virtual SDN environment," in Proc. IEEE 3rd Int. Conf. CloudNet, 2014, pp. 264–265.

[24] S. A. Mehdi, J. Khalid, and S. A. Khayam, "Revisiting traffic anomaly detection using software defined networking," in Recent Advances in Intrusion Detection. Berlin, Germany: Springer-Verlag, 2011, pp. 161–180.

[25] R. Braga, E. Mota, and A. Passito, "Lightweight DDoS flooding attack detection using NOX/OpenFlow," in Proc. IEEE 35th Conf. LCN, 2010, pp. 408–415.

[26] J. Suh et al., "Implementation of content-oriented networking architecture (CONA): A focus on DDoS countermeasure," in Proc. European NetFPGA Developers Workshop, Cambridge, U.K., 2010, pp. 1–6.

[27] Purnima Murali Mohan et. al., "Towards resilient in-band control path routing with malicious switch detection in SDN", IEEE COMSNETS, 2018, PP.9-16.

[28] Haifeng Zhou et. al., "SDN-RDCD: A Real-Time and Reliable Method for Detecting Compromised SDN Devices", IEEE/ACM transactions on networking, vol. 26, no. 5, october 2018 pp. 2048-2061

[29] D. Kreutz, F. Ramos, and P. Verissimo, "Towards secure and dependable software-defined networks," in Proc. 2nd ACM SIGCOMM Workshop Hot Topics Softw. Defined Netw., 2013, pp. 55–60.

[30] S. Shin et al., "FRESCO: Modular composable security services for software-defined networks," in Proc. Netw. Distrib. Security Symp., San Diego, CA, USA, 2013, pp. 1–16.

[31] N. Gude et al., "NOX: Towards an operating system for networks," ACM SIGCOMM Comput. Commun. Rev., vol. 38, no. 3, pp. 105–110, Jul. 2008.

[32] D. Erickson, "The beacon OpenFlow controller," in Proc. 2nd ACM SIGCOMM Workshop Hot Topics Softw. Defined Netw., 2013, pp. 13–18.

[33] A. Guha, M. Reitblatt, and N. Foster, "Machine-verified network controllers," ACM SIGPLAN Notices, vol. 48, no. 6, pp. 483–494, Jun. 2013.

[34] S. H. Yeganeh and Y. Ganjali, "Kandoo: A framework for efficient and scalable offloading of control applications," in Proc. 1st Workshop Hot Topics Softw. Defined Netw., 2012, pp. 19–24.

[35] N. Foster et al., "Frenetic: A network programming language," ACM SIGPLAN Notices, vol. 46, no. 9, pp. 279–291, Sep. 2011.

[36] T. Koponen et al., "Onix: A distributed control platform for large-scale production networks," in Proc. OSDI, 2010, vol. 10, pp. 1–6.

[37] Seung Yeob Nam, Dongwon Kim and Jeongeun Kim. “Enhanced ARP: Preventing ARP Poisoning-Based Man-in-the-Middle Attacks” IEEE Communications Letters, Vol. 14, No. 2, February 2010, pp. 187-189.

[38] Songyi Liu, “MAC Spoofing Attack Detection Based on Physical Layer Characteristics in Wireless Networks” IEEE, ICCEM, 2015.

[39] Timo Kiravuo, Mikko S¨arel¨a, and Jukka Manner, “A Survey of Ethernet LAN Security” IEEE Communications Surveys & Tutorials, Vol. 15, No. 3, 2013,pp. 1477-1491.

[40] A. Zaalouk, R. Khondoker, R. Marx, and K. Bayarou, “OrchSec: An orchestrator-based architecture for enhancing network-security using network monitoring and SDN control functions,” in Proc. IEEE NOMS, 2014, pp. 1–9.

[41] Pradeep Kumar Sharma and S.S Tyagi “Improving Security through Software Defined Networking (SDN): An SDN based Model”, IJRTE, vol. 8, issue 4, 2019, pp. 295-300.

[42] Marcelo Ruaro , Luciano Lores Caimi , and Fernando Gehm Moraes, “SDN-Based Secure Application Admission and Execution for Many-Cores”, IEEE Access, volume 8, 2020, pp. 177296- 177306.

[43] D. Kreutz et al., “Software-defined networking: A comprehensive survey,” arXiv preprint arXiv:1406.0440, 2014.

[44] D. Yu, A. W. Moore, C. Hall, and R. Anderson, “Authentication for resilience: The case of SDN,” in ser. Security Protocols XXI. Berlin, Germany: Springer-Verlag, 2013, pp. 39–44.

[45] P. Porras et al., “A security enforcement kernel for OpenFlow networks,” in Proc. 1st Workshop Hot Topics Softw. Defined Netw., 2012, pp. 121–126.

[46] S. Shin et al., “Rosemary: A robust, secure, and high-performance network operating system,” in Proc. ACM SIGSAC Conf. Comput. Commun. Security, 2014, pp. 78–89.

[47] B. Chandrasekaran and T. Benson, “Tolerating SDN application failures with LegoSDN,” in Proc. 13th ACM Workshop Hot Topics Netw., 2014, p. 22.

[48] G. Yao, J. Bi, Y. Li, et al., “On the Capacitated Controller Placement Problem in Software Defined Networks”, IEEE Communications Letters,vol.18, no.8, 2014, pp. 1339-1342.

[49] J. Liao, H. Sun, J. Wang, et al., “Density cluster based approach for controller placement problem in large-scale software defined networkings”, Computer Networks, vol.112, 2017, pp. 24-35.

[50] H. Li, P. Li, S. Guo, et al., “Byzantine-resilient secure software-defined networks with multiple controllers”, Proc. IEEE International Conferenceon Communications, 2014, pp. 695-700.

[51] OpenFlow Switch Specification Version 1.4, Open Network.

[52] K. Benton, L. J. Camp, and C. Small, “OpenFlow vulnerability assessment,” in Proc. 2nd ACM SIGCOMM Workshop Hot Topics Softw. Defined Netw., 2013, pp. 151–152.

[53] Josy Elsa Varghese and Balachandra uniyal, “An efficient IDS framework for DDOS attacks in SDN environment”, IEEE Access, 2021, pp. 69680-69699.

[54] Tooska Dargahi et. al., “A Survey on the Security of Stateful SDN Data Planes” IEEE Communications Surveys & Tutorials, Vol. 19, No. 3, 2017, PP. 1701-1724.

[55] A. A. Z. SOARES et. al., “3AS: Authentication, authorization, and accountability for sdn-based smart grids ”, IEEE Access, volume 9, 2021, pp. 88621-88640

[56] Kevin Barros Costa et al., “Enhancing Orchestration and Infrastructure Programmability in SDN with NOTORIETY”, IEEE Access, Volume 8, 2020, pp. 195487-195502.

[57] Basem Almadani , Abdurrahman Beg and Ashraf Mahmoud, “DSF: A Distributed SDN Control Plane Framework for the East/West Interface” IEEE Access, Volume 9, 2021, pp. 26735-26754.

[58] Ahmed Sallam , Ahmed Refaey,and Abdallah Shami, “On the Security of SDN: A Completed Secure and Scalable Framework using the Software-Defined Perimeter”, IEEE Access, volume 7, 2019. pp. 146577-146587.

A Comprehensive Framework for Big Data Analytics in Education

Ganeshayya Shidaganti[1], Prakash S[2]
Research Scholar, Visvesvaraya Technological University, Belagavi and Assistant Professor, Department of C.S.E[1]
M.S. Ramaiah Institute of Technology (Affiliated to VTU, Belagavi), Bengaluru, Karnataka, India[1]
Professor, Department of C.S.E, East Point College of Engineering and Technology, Bengaluru, India[2]

***Abstract*—With the adoption of cloud services for hosting knowledge delivery system in educational domain, there is a surplus quantity of education data being generated every day by current learning management system. Such data are associated with certain typical complexities that impose significant challenges for existing database management and analytics. Review of existing approaches towards educational data highlights that they do not offer full-fledged solution towards analytics and still there is an open-end problem. Therefore, the proposed system introduces a comprehensive framework which offers integrated operation of transformation, data quality, and predictive analytics. The emphasis is more towards achieving distributed analytical operation towards educational data in cloud. Implemented using analytical research methodology, the proposed system shows better analytical performance with respect to frequently used educational data analytical approaches.**

Keywords*—Big data; data analytics; educational big data; predictive analytics; text mining; machine learning; education technology

I. INTRODUCTION

There has been significantly increase in adoption of technique in the area of education system in recent time. It is because the process of knowledge delivery system is required to be carried out in order to reach number of users. However, this mechanism of online educational knowledge delivery system will demand heavier platform [1]. In this aspect, cloud computing offers a better option for hosting educational delivery system for online learning management [2]. However, such online learning management scheme is not only for delivering the live classes but it is also about understanding the entire process of service delivery from quality perspective [3]. Different information in the form of archives of study materials, online forums, information about students, instructors, payment and service syncing, etc. are required to be stored. With proliferation of mobile internet, adoptions of such educational services are tremendously increasing [4]. This results in generation of massive amount of data in the form of stream. Some data are generated voluntarily while many other data are generated involuntarily. The voluntary data will consist of study material, online notes, and respective information about students /instructor. The involuntary data consists of data that is generated autonomously e.g., positional data, trace data, behavioral specific data, etc.

There is problem associated with such form of educational data viz. i) the data generated is so massive that it cannot be supported by small scale deployment. These large data cannot be stored directly to the storage unit of cloud efficiently in distributed fashion, ii) the next issue is into form of the data that is being subjected to mining. Normally, the data is either semi-structured or unstructured which makes them ineligible for storing it into conventional storage units, iii) another bigger issue is the distributed deployment of educational data. If the cloud-based educational services are running from different geographical regions, than there will be different origination point of such educational big data. All these data are required to be aggregated in distributed manner before deploying them to the analytical process. In case of incomplete or imperfect aggregation, the mining process will be carried out over impartial data resulting in outliers. Hence, mining will be not effective in such way, iv) Finally, the problem is in applying machine learning in order to carry out predictive value of it as the form of knowledge extraction. A data is of no use until and unless it is not predictive. Therefore, it is not simpler mechanism to perform storage, processing, and analyzing existing educational big data.

At present, there has been various works being carried out towards analyzing such education data. Existing studies on big data is found to use sophisticated tools and framework [5-8]. However, such complex adoption cannot be taken into consideration in real cases. Apart from this, existing approaches of educational big data [9] lacks various conceptualization like heterogeneity in data, multiple and large-scale origination of educational data, streaming of educational data, etc. Irrespective of presence of multiple problems, only few problems are addressed symptomatically in existing approach. Therefore, this paper presents discussion of the unified approach where multiple problems associated with the educational big data is addressed.

The organization of the paper is as follows: Section 2 discusses about the existing studies while Section 3 discusses about the research problems, Section 4 discusses about the adopted research methodologies, while Section 5 discusses about the algorithm implementation. Results are discussed in Section 6 while conclusion is discussed in Section 7.

II. Related Work

There have been various forms of work carried out towards educational big data in recent times [10-11].

Uses of big data approach in educational domain were witnessed to explore the popular topic of study [12] Inclusion of deep learning concept has assisted in constructing such framework that could further facilitate in extraction of key-words. The work carried out by [13] have developed a model using structural equation modeling. The framework is designed using conventional technical adoption model toward educational domain targeting to find the comfortability of end user. Existing studies highlights that combined usage of warehouse, business analytics, and enterprise architecture can be used for improving the analytical operation [14]. Apart from this, the researchers have also offered an importance towards clustering process of educational data which consist of three stages viz. pre-processing, standardization, and modeling [15]. Clustering approach toward online learning can be personalized for improving the knowledge delivery process [16]. Along with clustering, classification-based approaches are also used for improving analytical processes over educational data [17]. All such advanced mechanism gives rise to evolution of smart campus; however, there are still issues associated with integrating such devices. Development of smart campus can be carried out over a platform of effective data fusion [18]. Inclusion of ubiquitous approach in the form of framework for facilitating distance learning is carried out by [19]. In order to offer user friendly experience, visualization of data can offer faster access to the knowledge. There are various visualization tools at present that can carry out this task [20-21] have presented another visual analytical scheme for online courses. This scheme is deployed for offering visual representation of the learning groups. Machine learning has been consistently found to be adopted in existing approaches towards incorporating smart features in framework building. Such framework is witnessed to assess the competencies of student [22]. Existing approach has also witnessed the usage of collaborative filtering process using predictive method [23]. This model predicts scores about the courses. The work of [24] has implemented a unique approach where the study contents are emphasized during the knowledge delivery process. Nearly similar predictive approach is presented in the work of [25] where the idea is to perform identification of the students that are in the level of risk. The work carried out by [26] has discussed that adoption of mining approach as well as analytical approach can be mechanized. This work has presented a comprehensive representation scheme of the involuntary regulation of the behavior of student. Apart from this, there are various other schemes towards big data analytics e.g., tensor-based scheme [27], compression based on context [28], pattern analysis [29], deep learning [30], clustering technique [31]. Existing system has also witnessed an extensive usage of Hadoop framework. Some of the existing approaches are discussed by [32-35]. Adoption of machine learning is another frequently used approaches in educational analytics viz. [36-39] finally, text mining is another frequently used approach for educational analytics e.g. [40-44]. Ifenthaler [45-46] have carried out study towards proving that all the upcoming forms of education system will be requiring advanced forms of analytics. The similar forms of proposition towards adoption of learning-based analytics during pandemic is also studied by Beerwinkle [47]. Further, the recent work of Lee et al. [48] have discussed about various innovative practices of using analytics over educational data in order to cater up both technological and pedagogical demands of students. The next section discusses about the research problems.

III. Research Problem

After reviewing the existing approaches of analytical operation associated with educational domain, it has been observed that that present schemes have certain loopholes viz.

Simplified Transformation Scheme: Majority of the transformation scheme is meant for making the data suitable for structurization and mining where complex operation is involved. Moreover enough emphasize is not offered on transformation operation. Educational big data is highly heterogeneous in its form and it requires cost effective transformation scheme. Data that can actually be carried out on educational data.

Presence of Artifacts in Transformed Data: Usually storage and analysis is carried out in explicit manner in two different places. A transformed data is forwarded to special block of operation that is meant to be carrying out analytical operation. Owing to various impediments in communication medium, there are fair chances of inclusion of artifacts in transformed data. Existing system has no scheme to solve this problem.

Cost Ineffective Predictive Schemes: Existing schemes uses various predictive techniques where majority of this techniques are highly iterative and depends upon the trained data. Higher the trained data, higher is accuracy in prediction.

Therefore, all the above points are required to be addressed in order to mechanism a truly distributed analytical modeling of educational data. Until and unless, these research problems are not addressed, it is challenging to evolve up novel solution.

IV. Research Methodology

This part of the proposed study focuses on achieving a high-quality data in order to make it suitable to apply analytics for better value extraction. Following are the details of proposed implementation:

The proposed system addresses an explicit problem of analysis the complex form of educational big data. With the generation of data from multiple sources, there is higher feasibility of inclusion of errors in the form of noise. These errors could be human-based, machine-based, as well as network-based. Hence, the presence of errors will generate significant outliers, which is detrimental to carry out analysis. At the same time, solution to eliminate or reduce errors cannot be carried out locally as it will be not be cost effective and moreover, it cannot offer instantaneous query processing capability. The complete implementation of the proposed study was carried out considering an explicit case study. In order to solve the above-mentioned problem, the proposed study considers a case study. Referring to Fig. 1, the study considers m number of data node (d_n) to represent repository of

educational data in different geographic location with an inclusion of errors α in each data nodes of k types. For simplicity, the study considers k<<m, which will mean that number of error types are considerably lower than number of data nodes. As the computation for error elimination cannot be carried out in local level, so the study considers the presence of a memory stream which can connect to indexes of all the data nodes and an external cluster hosted in cloud is considered as the prime location where the data aggregation is carried out. By data aggregation, it will mean consolidation of all the individual data along with the level of errors maintained within each data nodes. The proposed algorithm for error elimination is then applied over the cluster in order to finally generate an error free data (dbef).

Fig. 1 highlights an adopted solution strategy where the prime agenda of the proposed system is to make the incoming stream of educational big data effectively structured and prepared for high end and cost-effective analytical operation. For this purpose, the first preference is offered towards rectifying the structuredness of the raw data by implicating a simple and novel data transformation scheme. After the data transformation scheme is implemented, the next focus of the proposed system will be towards identifying the presence of artifacts in that transformed data when they are forwarded from various data nodes via memory streams. A superlative indexing scheme is implemented in the proposed system which indexes all forms of data especially when the data is further classified into two forms. One form of data is permanently saved while other form is stored in volatile memory system and the complete algorithm is implemented over the volatile memory system. Thereby, a significant saving of storage units is emphasized in proposed system. The proposed system also assists in identification of the position in the cell over the temporary storage units with respect to error prone data. Such data are not only identified but also substituted by statistically computed value. This mechanism assists in maintaining higher degree of data purity. Once the quality data is obtained, the next process is to carry out predictive analysis using a novel deep learning mechanism. The overall scheme of the proposed system is to offer the complete educational mined data to be used in the form of cloud-based services. So that a new avenue of analytical application can be used for automating the knowledge delivery system over educational domain in various perspective.

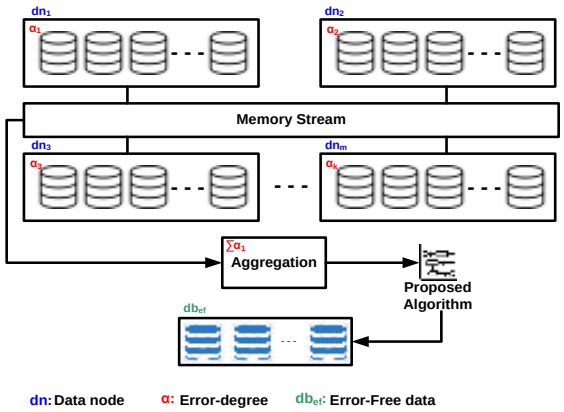


Fig. 1. Adopted Solution Strategy.

V. Algorithm Implementation

This proposed algorithm uses a series of three essential operations in order to facilitate development of a comprehensive analytical framework. The complete implementation is carried out with respect to three sequential phases of operation i.e., i) Data Transformation Phase, ii) Data Quality Incorporation Phase, and iii) Data Predictive Analysis Phase. It should be noted that all the above-mentioned approaches are applied over an educational big data. The discussion of this implementation phases is carried out as follow:

A. Data Transformation Phase

This is the preliminary phase of implementation where the emphasis is offered towards processing followed by an effective transformation of the data. The study considers data transformation as one of the essential operations which makes the incoming stream of data more suitable to be subjected to analytical operation. For this purpose, the study considers one unit of educational data in following form as shown in Fig. 2.

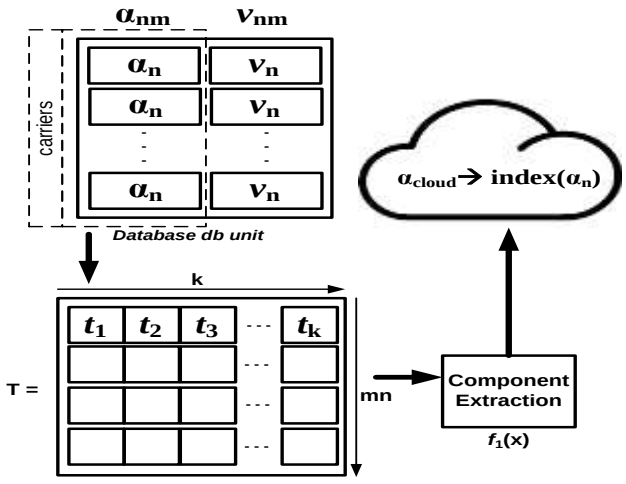


Fig. 2. Structure of Educational Data.

Fig. 2 highlights the structure of one database unit from the massive stream of educational big data. The study consider that each database unit has two essential parts i.e., carriers and value. The carrier is static information type of the educational data it targets to carry different explicit information called as components in each database units. The study considers αnm to represent components where n and m represents number of component and number of carriers in one database unit respectively. In case of educational big data, the constraint in this regards is,

$$m>>n \quad (1)$$

There are various possibilities of the number of component n, but during analysis, the proposed study can select a fixed integer value to obtain measurable outcome. Considering the educational data with respect to text mining approach, the study constructs a matrix T that retains all the textual contents in the complete database which can now be subjected to next phase of operation. The proposed algorithm for the process of data transformation is as follows:

Algorithm for Data Transformation

Input: c (carriers), α (component)

Output: O1 (transformed mined data)

Start

1. For i=1: c c is carriers
2. $\alpha f_1(T)$
3. dataβ {stream(v_{nm}) (α_{cloud})}
4. posg(data)
5. φ (γ)data
6. generates O_1

End

The description of the algorithmic steps are as follows: The algorithm takes the input of c(carriers) and α(component) for all the database units (Line-1). An explicit function f1(x) is applied over the matrix T which retains all the textual contents with a dimension size of (k x mn). Applying this function results in extraction of all the essential components which α which are then retain in cloud storage system permanently (Line-2).

Each individual component is further indexed within the cloud storage which ensures the rejection of storage of similar components as well as it can be called upon for looking for ownership of all individual values during query processing. This results in further storage optimization while values are never forwarded in this stage. The next part of the algorithm is to carry out data organization as follows:

$$\alpha\text{cloudindex}(\alpha n) \quad (2)$$

The expression (2) exhibits that all the incoming streams of educational data value vnm are now assessed for their owner components which now resides in cloud αcloud considered for all database db. (Line-3). The expression (3) exhibits that proposed system applies a function β(x) where web-script tags are applied on both components in cloud αcloud and its respective value vnm. This operation results in well-formatted data that could be supported on any client application over cloud interface in semi-structured manner. The algorithm also extracts the position information of the individual data which directly assists in lowering the search time during the query processing (Line-4). A function g(x) is designed for extracting all the positional information from the data and stored it in pos matrix. Finally, the knowledge extraction is carried out where γ syntactical rule set is used as following:

$$\gamma = \{r1, r2 \ldots rl\} \quad (3)$$

In the above expression, the variables r is l number of rule set which offers semantic information of the chosen text from the value of data. The algorithm constructs a function φ(x) which computes the syntactical correlation between rl and all the incoming data (Line-5). This operation results in the generation of knowledge as the outcome O_1 (Line-6). This outcome can be now stored back in the cloud storage system. The algorithm therefore offers a good balance between storage optimization (by storing only the mined data and non-repeating components) and data transformation operation. The study outcome is now assessed for next level of challenge with solution.

B. Data Quality Incorporation Phase

This module of operation is executed after the first algorithm is successfully executed that results in transformed data O1. It should be noted that execution of first algorithm is accompanied by storage of component information αnm in the indexed cloud storage permanently. However, the evaluated mined data O1 is ready to be stored distributed manner in cloud. In the process of distributed transmission of the aggregated transformed data O1, there are various possibilities of inclusion of further artifacts associated with network-based transmission. This phenomenon could significantly affect the quality of data resulting in inclusion of storage of artifact-incorporated transformed data. Hence, this part of the algorithm is mainly focused on rectifying the artifacts and substitutes the artifact-based transformed data into more quality data O2. The process flow of this part of implementation is as follows:

Fig. 3 highlights the process implementation towards identification and removal of the artifacts in order to retain higher degree of data quality. By data quality, it means that complete structure of the distributed database db should be fulfilled. Presences of any missing value of noisy values are easier to find but difficult to be rectified in distributed manner. For this purpose, the proposed system carries out following steps of operation in the form of algorithm: The algorithm initially takes the input of all the transformed data released from prior algorithm and computes its size (Line-1). The next part of the implementation is to split the complete text-based data T (considering both components α and their respective values v) into smaller components T1, T2, . . . Th, where h is number of splits of the data carried out on the basis of total number of available storage slots H in cloud (Line-2). It will eventually mean that the proposed algorithm offers a better form of elastic cloud usage where the on-demand scaling process of the data splitting operation is carried out. This operation is one significant step towards i) storage optimization as well as ii) faster query processing in distributed manner over cloud environment. By splitting the aggregated data into different segments, the network overhead towards processing the mined data is potentially controlled. Apart from this, it also offers a significant level of mined data availability which is also a part of solution towards dense state of traffic.

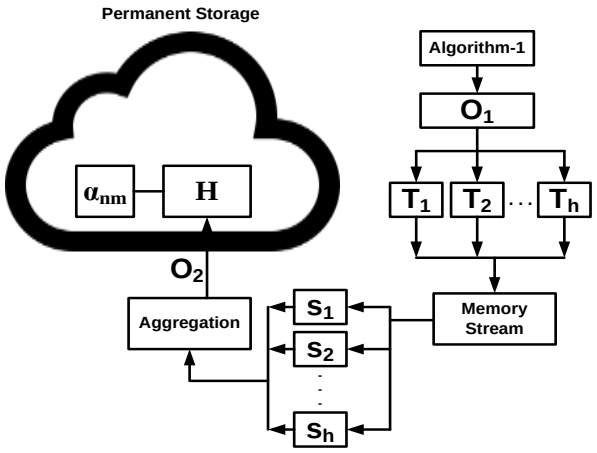


Fig. 3. Process of Data Quality Incorporation.

Algorithm for Data Quality

Input: O_1 (transformed data)

Output: O_2 (quality data)

Start

1. For i=1: size (O_1)
2. construct T= {T_1, T_2…., T_h} h⩽H
3. create s= {s_1, s_2……s_u }|u = h
4. For j=1: num (α_{nm})
5. search (v_{nm})err
6. $(v_{nm})_{corr} f_2((v_{nm})err, (v_{nm}+1))$
7. $O_2(v_{nm})_{corr}$

End

All the split data (T) is now forwarded to the memory stream which is an explicit buffer maintained over the cloud cluster or edge server where adaptive queue management is carried out. The next part of the implementation is to perform allocation of this split data with distinct streams. The algorithm creates a matrix s which has u number of streams where each stream will carry different split of data. In order to perform optimization of the network resources, the algorithm considers the equivalent value of both u and h (Line-3). This consideration prevents the system memory stream in edge server to create unnecessary streams of data. This will eventually mean that outgoing traffic of s doesn't offer any form of data overhead over the cloud interface prior to storage. The algorithm then looks for all the number num of value vnm (Line-4) from this stream of data to find if there is presence of any values with artifact (Line-5). The study considers that there are no artifacts associated with components as all the components are stored permanently over cloud. Hence, lesser chances of error prone err component information and all the errors will be related to values vnm itself.

The prime contribution of this algorithm is that it constructs a function f2(x) which performs statistical calculation in following manner. The columns of all the respective values are considered which have presence of artifact data followed by computation of statistical value (mean) of it. This extracted statistical value is now compared with all the columns of remaining streams of data (Line-6). Only the highest correlated data is extracted and substituted in the target value with prior artifacts (Line-7). It will mean that the final steps of this algorithm result in substitution of statistically computed data in the cell which has priory data with artifacts. Assuming that the proposed system works on defined domain of heterogeneous data, there are no chances of computed data with higher fluctuation or deviation. Therefore, the computation of mean value offers faster and reliable substitution of computed data. This operation ensures that any incoming data should never have any artifacts and in case there are any artifacts than they are going to be searched upon by this algorithm and substituted with accurate value. Hence, the process eliminates any presence of data uncertainty issue and offer inclusion of higher quality of data in cost effective manner.

C. Data Predictive Analysis Phase

This is the final part of execution which applies machine learning mechanism over the quality data O2 obtained from prior algorithm implementation. The part of the implementation considers the similar distributed scenario where the quality data O2 is assumed to be generated in distributed manner. The process flow of the proposed system is shown in Fig. 4.

After the data is considered to be distributed i.e., O21, O22. . . O2H, it is subjected to dual sequential operation of indexing and then sorting. This process is slightly different in contrast to conventional deep learning approach where there are possibilities of various numbers of features. The essential steps of processing of proposed algorithm are as follows:

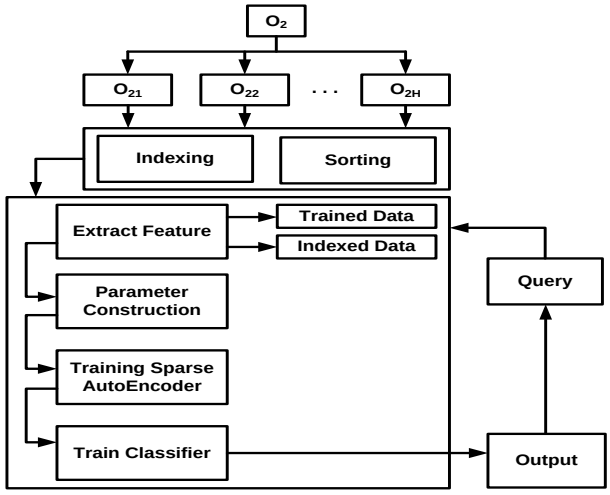


Fig. 4. Process of Predictive Analysis.

The proposed system takes the input of the quality data O2 from the prior algorithm implementation and computes its size (Line-1). The process of data aggregation is further carried out using a function agg applied over quality data O2 (Line-2). The proposed system considers very simple form of feature feat, which are indexed data μ and sorted data τ (Line-3). This step is uniquely carried out in proposed system. The next step is to carry out construction of various fundamental parameters e.g., size of input data, number of classes which are length of uniquely trained data, weight decay parameter and anticipated activation of hidden layers, weight of sparsity penalty term, and maximum number of iterations. The proposed system makes use of the auto-encoding system, which are subjected to training depending upon the fact of replication of input with the outcome data. Upon encoding the data, the transformation is carried out over the obtained set of features which is actually chosen by the auto-encoders itself. This is further used for the decoding purpose in consecutive process. Finally, the error is computed from the difference obtained from decoded data and input data, while this information is further used for minimizing the rate of error resulting in predictive outcome. In this entire process, the proposed system also adopts the usage of the optimization process towards solving non- linear unconstrained problems using iterative process. The essential steps of the proposed system are as follows:

Algorithm for Prediction

Input: O_2 (quality data)

Output: O_3 (predicted value)

Start

1. For i=1: size (O_2)
2. a=agg (O2) = {O_{21}, O_{22}, O_{23}... O_{2H}}
3. [μ τ] [index (a), sort (a)]
4. feat [μ τ]
5. trf_3(feat) ^{sc}
6. O_3tr

End

The next part of the algorithm implementation is about applying a sparse auto encoding mechanism which has many numbers of neuron present in hidden layers in contrast to its respective input later. However, there is a good possibility of allocating of single neuron for one input data. This problem is overcome, the proposed system performs frequent switching of neurons over varied ranges of iteration. This mechanism allows the precise encoding of the features leading to better form of predictive analysis. The proposed system furthers carry out a unique mechanism of training. It trains the initial layer of the sparse auto-encoder with numerical optimization mechanism such that presence of any instances related to the non-smooth optimization can be considered. This optimization technique ensures the presence of zero gradients for all the required condition in order to achieve optimized performance. The next sparse auto encoder is feed forward mechanism considering the trained data, size of input layer, hidden layer. This mechanism is nearly equivalent to the single layer perception where the number of layers for both input and output are same but it can have varied number of hidden layers. The prime agenda in this part of operation is to reduce the significant difference between the input and output. One of the significant contributions of this algorithm is that it can carry out unsupervised form of learning mechanism without any form of dependency towards indexed data in its input layer in order to carry out learning operation. This characteristic of the proposed operation ensures that even if the incoming stream of the data is not indexed appropriately, then also the proposed system is able to perform the encoding operation. The final part of the implementation is followed by performing classification of the trained data that offers the complete probability of all the classes of indexed data. This form of the classification is essentially a binary type of the statistical regression that offers a significant precision in the process of classification. One of the interesting points of proposed algorithm is the generation of elite feature in each cycle of training which significantly improves the accuracy level. It also offers higher scope of utilization of the unstructured educational data and yet maintaining better for of predictive performance. By its unique mechanism of training, the algorithm also reduces the cost of training operation as better results are obtained in reduced number of training cycle. Apart from this, the accuracy doesn't get affected in presence of indexed data; however, it is used for further improving the classification performance.

VI. Result Analysis

This section discusses about the outcome obtained after implementing the proposed logic in the prior section. The proposed study implements a comprehensive mechanism which aggregates, organizes, transforms, processes, and performs predictive operation over educational big data. This section discusses about the strategy adopted for analysis, the dataset considered for the result analysis, test case used, and discussion of the accomplished results.

A. Analysis Strategy

A closer look at the proposed system shows that it carries out three sets of sequential operation in order to carry out comprehensive analytical operation. However, in order to ascertain the effectiveness of the proposed system, there is a need of using certain strategy to carry out analysis. The proposed system makes use of three essential strategies in order to measure the effectiveness of the proposed system. The primary strategy of the analysis is to assess all the performance parameters with respect to size of the data. There are multiple reasons behind this adopted strategy. The first reason is associated with the scalability factor of the proposed system. By scalability, it will mean that proposed system will successfully deliver similar form of optimal services towards data analysis. Normally, the capacity of the cloud servers is finite in spite of using distributed environment, therefore, increasing traffic will have possibility towards overloading the task towards this cloud server that could affect the database management for heavier and uncertain traffic condition. The second reason for assuming size of data is because in educational domain, the size of the data is always exponentially increasing in shortest span as well as in various forms. Therefore, analyzing size of data contributes towards effective review of scalability factor of proposed system. The secondary strategy of the proposed analysis is carried out for comparative analysis of existing approach that is frequently used for classification purpose while boosting the analytical operation. By comparing with the existing approach of analytics, the accomplished outcome can be generalized for the effectiveness towards the distributed data mining operation associated with educational domain. The tertiary strategy of analysis is to perform selection of multiple forms of performance parameter over the same test environment along with other existing approaches. The proposed system uses data transformation time and data transformation accuracy as the performance parameters for assessing first algorithm. In order to assess the second algorithm, the proposed system uses data fusion time and data fusion quality while data prediction time and data prediction accuracy is used for assessing the last algorithm. Therefore, a comprehensive set of performance parameters are used for this purpose. The study outcome is assessed and compared with respect open-source distributed framework Hadoop, machine learning, and conventional text mining approach. All the assessment has been carried out considering similar environmental variable in order to obtain an unbiased outcome of the proposed system. The efficiency of mining operation is assessed using different methods in proposed system considering educational dataset.

B. Dataset Considered

The discussion of dataset is quite important in proposed system. In order to testify the effectiveness of the proposed system, the input data should have certain criteria to fulfill viz. i) the input data should be associated with educational domain, ii) the input data should be stream of educational data originated from distributed systems in institution, iii) the input data should have the characteristics of massive size, inclusion of artifacts, lesser valuable data. However, availability of such form of data is quite difficult. Therefore, the proposed system initially reviews the publically available big data in order to understand the form and specification of big data. It is found that existing big dataset has a finite smaller size and it doesn't possess the 3rd criteria discussed about the suitable input file. Therefore, the proposed system chooses to construct a synthetic data considering the domain of educational system. The dataset considered for the analysis of the proposed system has an overall size of 100 GB with 7,500 plain text files consisting of specific performance information about educational delivery system.

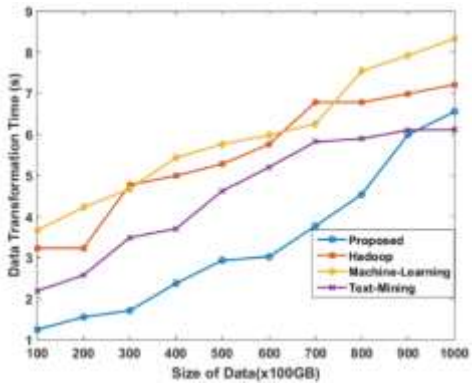
Table I highlights one set of information among many existing within one plain text file. It can be seen that there are 9 categories involved for one course undertaken and a single plain text file can contain many more such course information within it. The study assumes that Course ID is a unique number which is allocated by the system for every course undertaken by the scholar. Hence it is non-repetitive in nature in overall dataset considered for assessment. Course Title represents the name of the course undertaken by the scholar and it is also fixed and non-iterative in nature. It should be noted that every Course Title has uniquely allocated Course ID which is non-iterative. Course Type represents the do- main of the course viz. i) social, ii) political, iii) engineering, iv) medical, v) literature, etc. Date represents the start point of course while Location represents the geographical position of the knowledge delivery point of the scholar. This information can be easily retrieved from IP address of scholar. The category of Course Status is either active or inactive. The flag active will represent ongoing course and inactive will represent already delivered course. The category Total Episode will represent total number of online sessions allocated for specific course. Scholar Name is name of the client who has privilege access to this cloud-based knowledge delivery application and receives training from instructor while the category Scholar Feedback is the response given by the user for either the ongoing courses or the completed course.

Scripted in MATLAB using normal windows machine, the proposed system is analyzed for all the three discrete and sequential algorithms in order to measure their level of effectiveness. The proposed system is basically an analytical model and MATLAB is one of the best tools for this. The prime reason to use MATLAB is that it offers simpler operation for carrying out extensive evaluation where the focus can be purely on the analysis without any concern of tools, writing bigger scripts, or dependency of extensive code. The first algorithm is assessed with respect to data transformation time and accuracy (Fig. 5). The outcome shown in Fig. 5 (a) highlights that although transformation time increases with increase in data, yet proposed system offers reduced data transformation time. The similar trend is also observed in Fig. 5 (b) where the proposed system is found to offer increased data transformation accuracy compared to existing approaches. The prime reason behind this is that proposed data transformation algorithm is carried out in a smaller number of non-iterative steps without offering any dependencies of third parties causing faster processing time. Apart from this, the proposed system makes use of semantics for the purpose of extracting the essential elements of data during mining process causing higher accuracy. Such semantics are further user-defined and can be constructed depending upon the demands making the proposed system free from any lexical database system. On the other hand, adoption of Hadoop has higher dependencies on system requirements and construction of higher number of tall arrays in order to deal with increasing dimension of dataset. Hence, transformation time evidently increases while accuracy decreases in such case. In the case of machine learning approach, the approach is extensively iterative in its operation as well as there is a dependency on training dataset. Finally, all the text mining approaches are assessed to find out that they consume more time to perform transformation of educational data. Apart from this, there is also an increasing dependency on lexical dataset in order to extract the logical meaning of the elements within corpus.

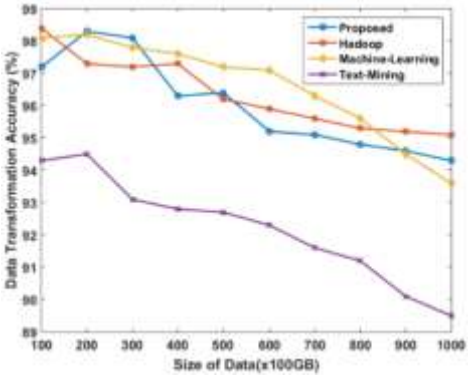
The effectiveness of the second algorithm is assessed and outcomes are shown in Fig. 6. The proposed system offers highly reduced data fusion time in comparison to existing approaches (Fig. 6 (a)). The prime justification for this lowered fusion time is that proposed system performs the queuing of the incoming traffic in its data node which is capable of carrying out distributed stream management. The likability of the data and traffic are highly maintained owing to an effective indexing policy executed in this part of implementation. This causes faster aggregation of data as although the data nodes are distributed but they have a good linkage causing an effective redundancy management too. At the same time, the proposed system also offers increasing data quality without using any third party. Hadoop has increasing dependencies on constructing arrays in order to save increasing data. It uses too many mechanisms for compacting the data which consumes time for performing data fusion. Machine learning has inclusion of training while text mining approaches has too much simplified and often fails to understand the relationship among the data especially if the database is of massive and uncertain scale.

TABLE I. DATA SPECIFICATION

#	Categories	Character Range	Data-Type
1	Course ID	1-5	Number
2	Course Title	1-15	String
3	Course Type	1-15	String
4	Date	10	Number
5	Location	1-15	String
6	Course Status	6/8	String
7	Total Episodes	1-3	Number
8	Scholar Name	1-20	String
9	Scholar Feedback	1-200	String

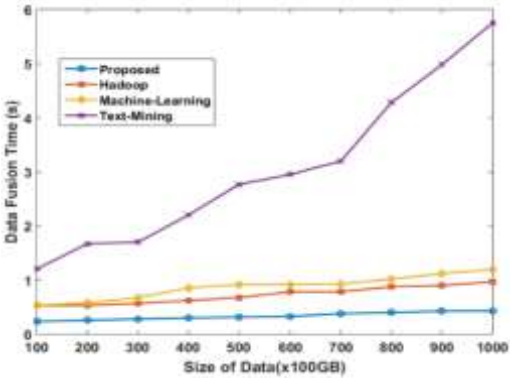


(a) Data Transformation Time.

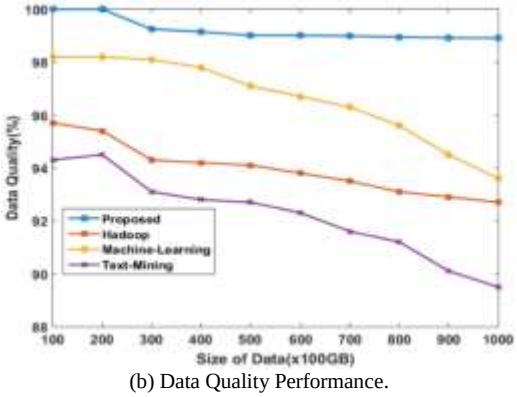


(b) Data Transformation Accuracy.

Fig. 5. Outcome for Algorithm of Data Transformation.



(a) Data Fusion Time.

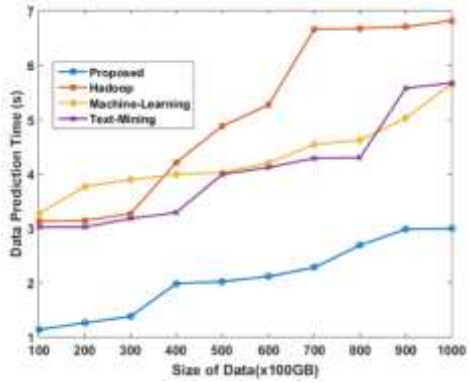


(b) Data Quality Performance.

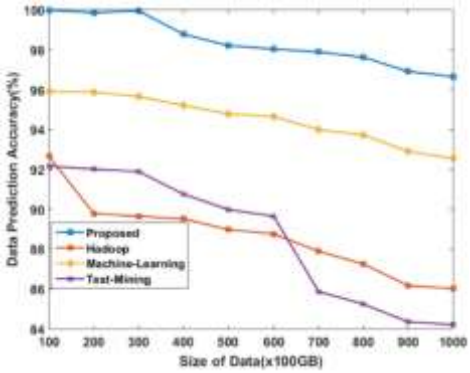
Fig. 6. Outcome for Algorithm of Data Quality.

The proposed system makes use of simplified statistical measures in order to offer an improved data quality in much reduced number of steps, which is not found in any existing approaches (Fig. 6 (b)). Hadoop has the better capability of storing and managing large and distributed data; however, they don't offer any form of identification and substitution of artifact data by itself. This causes slight increase in data fusion time and it takes assistance of zookeeper and other metadata management present in its architecture. This is also the reason of reduced data quality in Hadoop. Machine learning offer nearly equal time of operation for data fusion just like Hadoop, but its performance towards data quality depends upon its epoch level. Machine learning is capable of offering higher accuracy but the accuracy for higher size of data in present state is found to be reduced in it. The reason is simplified as the activation function in it is incapable of offering higher accuracy. On the other hand, text mining approach does not offer any form of substitution operation for the artifact elements in the corpus. Therefore, the proposed system is capable of offering a better form of data quality along with replacement of the error-prone data in faster way.

In order to assess the predictive analysis of the proposed system, the processing time as well as accuracy is the prominent indicator of its performance. The outcome shown in Fig. 7 eventually highlights that proposed system offers better predictive performance in contrast to existing approaches. It should be noted that the outcome of the predictive analysis is also a mined data and this data is essential for the data analyst (or stakeholder of the data). The first prominent reason for data prediction time is an inclusion of lesser number of iterations toward reaching the minimum gradient (Fig. 7 (a)). As the proposed system offers an exclusive extraction of sequential mined data with progression of each algorithms, therefore, a greater number of information is obtained in this process which reduces the decision-making time for the proposed system in order to make prediction. On the other hand, availability of more precise and larger set of filtered information also results in higher accuracy in proposed system (Fig. 7 (b))). On the other hand, Hadoop doesn't have extensive capability to carry out data prediction and hence it offers extensive time processing (Fig. 7 (a)). Hadoop doesn't address the problems associated with data uncertainty although it can offer better data transformation scheme for homogeneous data. This adversely affects the accuracy score of Hadoop. The machine learning scheme is found to offer increased consumption time owing to an inclusion of training however, its accuracy is the next better score after the proposed system. Text mining approach offers simplified mechanism; however, knowledge extraction process is quite length process in it resulting in higher involvement of prediction time. This also results in reduced accuracy.



(a) Data Prediction Time



(b) Data Prediction Accuracy.

Fig. 7. Outcome of Algorithm for Predictive Analysis.

VII. CONCLUSION

This paper has presented a framework that is meant to carry out comprehensive operation that leads towards an effective analytical operation over educational data. The proposed system has emphasized over data transformation, data quality incorporation, and predictive analytics in educational data. Scripted in MATLAB, the study highlights that it is capable of better analytical operation in contrast to text mining approach, machine learning approach, and Hadoop, which are the most used techniques in data analytics over educational domain. Our future work will towards optimizing the performance more by exploring more approaches towards its implication on real-time applications.

REFERENCES

[1] Al-Maroof, R. S., Alfaisal, A. M., & Salloum, S. A. (2021). Google glass adoption in the educational environment: A case study in the Gulf area. Education and Information Technologies, 26(3), 2477-2500.

[2] Sharma, D., & Kumar, V. (2021). A framework for collaborative and convenient learning on cloud computing platforms. In Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing (pp. 629-650). IGI Global.

[3] Ammenwerth, E., Hackl, W. O., Hoerbst, A., & Felderer, M. (2021). Indicators for cooperative, online-based learning and their role in quality management of online learning. In Research Anthology on Developing Effective Online Learning Courses (pp. 1709-1724). IGI Global.

[4] Huang, R., Tlili, A., Chang, T. W., Zhang, X., Nascimbeni, F., & Burgos, D. (2020). Disrupted classes, undisrupted learning during COVID-19 outbreak in China: application of open educational practices and resources. Smart Learning Environments, 7(1), 1-15.

[5] Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2020). Big data in education: a state of the art, limitations, and future research directions. International Journal of Educational Technology in Higher Education, 17(1), 1-23.

[6] Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., ... & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. Review of Research in Education, 44(1), 130-160.

[7] Williamson, B. (2017). Big data in education: The digital future of learning, policy and practice. Sage.

[8] Huda, M., Maseleno, A., Atmotiyoso, P., Siregar, M., Ahmad, R., Jasmi, K., & Muhamad, N. (2018). Big data emerging technology: insights into innovative environment for online learning resources. International Journal of Emerging Technologies in Learning (iJET), 13(1), 23-36.

[9] Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., ... & Hirokawa, S. (2017). Learning analytics for e-book-based educational big data in higher education. In Smart sensors at the IoT frontier (pp. 327-350). Springer, Cham.

[10] Quadir, B., Chen, N. S., & Isaias, P. (2020). Analyzing the educational goals, problems and techniques used in educational big data research from 2010 to 2018. Interactive Learning Environments, 1-17.

[11] Zhang, W., & Qin, S. (2018, March). A brief analysis of the key technologies and applications of educational data mining on online learning platform. In 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA) (pp. 83-86). IEEE.

[12] Wang, B., Yang, B., Shan, S., & Chen, H. (2019). Detecting hot topics from academic big data. IEEE Access, 7, 185916-185927.

[13] Al-Rahmi, W. M., Yahaya, N., Aldraiweesh, A. A., Alturki, U., Alamri, M. M., Saud, M. S. B., ... & Alhamed, O. A. (2019). Big data adoption and knowledge management sharing: An empirical investigation on their adoption and sustainability as a purpose of education. IEEE Access, 7, 47245-47258.

[14] Moscoso-Zea, O., Castro, J., Paredes-Gualtor, J., & Luján-Mora, S. (2019). A hybrid infrastructure of enterprise architecture and business intelligence & analytics for knowledge management in education. IEEE Access, 7, 38778-38788.

[15] Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. Ieee Access, 5, 15991-16005.

[16] Kausar, S., Huahu, X., Hussain, I., Wenhao, Z., & Zahid, M. (2018). Integration of data mining clustering approach in the personalized E-learning system. IEEE Access, 6, 72724-72734.

[17] Yu, L., Wu, X., & Yang, Y. (2019). An online education data classification model based on Tr_MAdaBoost algorithm. Chinese Journal of Electronics, 28(1), 21-28.

[18] Yang, A. M., Li, S. S., Ren, C. H., Liu, H. X., Han, Y., & Liu, L. (2018). Situational awareness system in the smart campus. Ieee Access, 6, 63976-63986.

[19] Mehmood, R., Alam, F., Albogami, N. N., Katib, I., Albeshri, A., & Altowaijri, S. M. (2017). UTiLearn: a personalised ubiquitous teaching and learning system for smart societies. IEEE Access, 5, 2615-2635.

[20] Liu, J., Tang, T., Wang, W., Xu, B., Kong, X., & Xia, F. (2018). A survey of scholarly data visualization. Ieee Access, 6, 19205-19221.

[21] Chen, Q., Yue, X., Plantaz, X., Chen, Y., Shi, C., Pong, T. C., & Qu, H. (2018). Viseq: Visual analytics of learning sequence in massive open online courses. IEEE transactions on visualization and computer graphics, 26(3), 1622-1636.

[22] Chou, C. Y., Tseng, S. F., Chih, W. C., Chen, Z. H., Chao, P. Y., Lai, K. R., ... & Lin, Y. L. (2015). Open student models of core competencies at the curriculum level: Using learning analytics for student reflection. IEEE Transactions on Emerging Topics in Computing, 5(1), 32-44.

[23] Huang, L., Wang, C. D., Chao, H. Y., Lai, J. H., & Philip, S. Y. (2019). A score prediction approach for optional course recommendation via cross-user-domain collaborative filtering. IEEE Access, 7, 19550-19563.

[24] Xie, T., Zheng, Q., Zhang, W., & Qu, H. (2017). Modeling and predicting the active video-viewing time in a large-scale E-learning system. IEEE Access, 5, 11490-11504.

[25] Hung, J. L., Shelton, B. E., Yang, J., & Du, X. (2019). Improving predictive modeling for at-risk student identification: A multistage approach. IEEE Transactions on Learning Technologies, 12(2), 148-157.

[26] Fincham, E., Gašević, D., Jovanović, J., & Pardo, A. (2018). From study tactics to learning strategies: An analytical method for extracting interpretable representations. IEEE Transactions on Learning Technologies, 12(1), 59-72.

[27] Kaur, D., Aujla, G. S., Kumar, N., Zomaya, A. Y., Perera, C., & Ranjan, R. (2018). Tensor-based big data management scheme for dimensionality reduction problem in smart grid systems: SDN perspective. IEEE Transactions on Knowledge and Data Engineering, 30(10), 1985-1998.

[28] Dong, D., & Herbert, J. (2017). Content-aware partial compression for textual big data analysis in hadoop. IEEE Transactions on Big Data, 4(4), 459-472.

[29] Edstrom, J., Chen, D., Gong, Y., Wang, J., & Gong, N. (2017). Data-pattern enabled self-recovery low-power storage system for big video data. IEEE Transactions on Big Data, 5(1), 95-105.

[30] Yang, Y., & Chen, T. (2019). Analysis and visualization implementation of medical big data resource sharing mechanism based on deep learning. IEEE Access, 7, 156077-156088.

[31] Kumar, S., & Singh, M. (2019). A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem. Big Data Mining and Analytics, 2(4), 240-247.

[32] Parsola, J., Gangodkar, D., & Mittal, A. (2019, July). Mobile Application for Storage and Retrieval of e-learning videos Using Hadoop. In 2019 International Conference on Communication and Electronics Systems (ICCES) (pp. 757-762). IEEE.

[33] Jagtap, A., Bodkhe, B., Gaikwad, B., & Kalyana, S. (2016, January). Homogenizing social networking with smart education by means of machine learning and Hadoop: A case study. In 2016 International Conference on Internet of Things and Applications (IOTA) (pp. 85-90). IEEE.

[34] Tian, X., Cui, B., Deng, J., & Yang, J. (2016, July). The Performance Optimization of Hadoop during Mining Online Education Packets for Malware Detection. In 2016 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS) (pp. 305-309). IEEE.

[35] Cholissodin, I., & Supianto, A. A. (2019, September). Enhancement Full Open Source Hadoop Distribution Universal Big Data Up Projects (UBig) From Education To Enterprise. In 2019 International Conference on Sustainable Information Engineering and Technology (SIET) (pp. 90-93). IEEE.

[36] Wu, C. H. (2019, July). A Concept Framework of Using Education Game With Artificial Neural Network Techniques to Identify Learning Styles. In 2019 International Conference on Machine Learning and Cybernetics (ICMLC) (pp. 1-6). IEEE.

[37] Huang, L., & Ma, K. S. (2018, October). Introducing Machine Learning to First-year Undergraduate Engineering Students Through an Authentic and Active Learning Labware. In 2018 IEEE Frontiers in Education Conference (FIE) (pp. 1-4). IEEE.

[38] Gouripeddi, P. S., Gouripeddi, R., & Gouripeddi, S. P. (2019, December). Toward Machine Learning and Big Data Approaches for Learning Analytics. In 2019 IEEE Tenth International Conference on Technology for Education (T4E) (pp. 256-257). IEEE.

[39] Jeon, H., Oh, H., & Lee, J. (2018, October). Machine Learning based Fast Reading Algorithm for Future ICT based Education. In 2018 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 771-775). IEEE.

[40] Sriyanong, W., Moungmingsuk, N., & Khamphakdee, N. (2018, July). A Text Preprocessing Framework for Text Mining on Big Data Infrastructure. In 2018 2nd International Conference on Imaging, Signal Processing and Communication (ICISPC) (pp. 169-173). IEEE.

[41] Wang, H., Wang, Q., & Wang, W. (2018, September). Text mining for educational literature on big data with Hadoop. In 2018 IEEE International Conference on Smart Cloud (SmartCloud) (pp. 166-170). IEEE.

[42] Çakir, M. U., & Güldamlasioğlu, S. (2016, June). Text mining analysis in Turkish language using big data tools. In 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC) (Vol. 1, pp. 614-618). IEEE.

[43] Larson, R. R., Marciano, R., Hou, C. Y., Watry, P., Harrison, J., Aguilar, L., & Fuselier, J. (2014, October). Integrating Data Mining and Data Management Technologies for Scholarly Inquiry. In 2014 IEEE International Conference on Big Data (Big Data) (pp. 67-71). IEEE.

[44] Cavallaro, G., Riedel, M., Richerzhagen, M., Benediktsson, J. A., & Plaza, A. (2015). On understanding big data impacts in remotely sensed image classification using support vector machine methods. IEEE journal of selected topics in applied earth observations and remote sensing, 8(10), 4634-4646.

[45] Ifenthaler, D., & Yau, J. Y. K. (2020). Utilising learning analytics to support study success in higher education: a systematic review. Educational Technology Research and Development, 68(4), 1961-1990.

[46] Gibson, D., & Ifenthaler, D. (2020). Adoption of learning analytics. In Adoption of data analytics in higher education learning and teaching (pp. 3-20). Springer, Cham.

[47] Beerwinkle, A. L. (2021). The use of learning analytics and the potential risk of harm for K-12 students participating in digital learning environments. Educational Technology Research and Development, 69(1), 327-330.

[48] Lee, L. K., & Cheung, S. K. (2020). Learning analytics: Current trends and innovative practices. Journal of Computers in Education, 7(1), 1-6.

A Systematic Mapping Study of Software Usability Studies

Abdulwahab Ali Almazroi
University of Jeddah, College of Computing and Information Technology at Khulais
Department of Information Technology, Jeddah, Saudi Arabia

***Abstract*—Among software quality attributes "software usability" is considered as one of the vital factors in software engineering literature. Software usability is the ability for users to generally understand, use, and learn a software with ease. Due to the importance of usability in software quality, a considerable amount of literature is published in the past decade. Few review and survey studies are also published to critically review the existing literature in the domain. However, there is limited research covering systematic mapping study of software usability. Mapping studies help in analyzing the general trends and research productivity in a research area. To fill this gap, this work critically examines the overall research productivity, demographics, trends, and challenges of software usability. The objective is to classify the current contributions and trends in the area of software usability. We retrieved 9,874 research articles from six research databases and 62 works are selected as primary studies using an evidence-based approach. The result of this mapping study shows that software usability is an active research area, with a promising number of works published in the last decade (2011 - 2020). We identified that the current literature spans over multiple article classes of which investigative papers, model proposals and evaluation papers are the most frequently published article types. We found experiments and theoretical validations to be the most common validation techniques. In terms of application domains; web, software development and mobile applications are the most frequent domains where usability studies are conducted. We identified that future usability studies should focus more on field studies as well as on the usability testing of scientific software packages. It will be of importance to consider ethical issues in usability testing as well.**

Keywords*—Software usability; usability study; systematic mapping study; systematic literature review; software engineering

I. Introduction

Based on IEEE Std.610.12, software usability is the user's simplicity in learning to provide inputs and operate a given component or system [1]. Although software and system usability are considered as non-functional requirements, its importance cannot be overstated [2]. Sagar and Saha [3] listed six software quality criteria which includes usability as well. As modern-day software is evolving to be more complex and omnipresent, software usability has become an indispensable non-functional requirement for ensuring software quality.

In the field of software engineering, various studies have identified usability issues and its implication on software quality [4-6]. As per ISO/IEC 9126, software usability is the ability of users to generally understand, use, and learn the software with ease. Based on [7], there are five important usability attributes, which are learnability, attractiveness, understandability, operability, and usability compliance. Thus, the definition for software usability differs among different standards and researchers [3].

In the real world, various application domains have duly considered usability engineering as an important area. It is widely considered to be a far-reaching research area at large [3] and is used in domains such as aerospace [8], computer-based medical devices [9], defense[10], mobile devices [11], web applications [12] etc. Various research and review studies were conducted on software usability [3, 11-21]. However, based on our findings from the existing literature, we observed that a systematic mapping study (SMS) in this domain is lacking, confirming the claim of Bitkina, et al. [9] regarding the absence of UX/usability studies. Hence, this SMS is undertaken to fill this research gap through the extensive analysis of important studies that were published in the last decade (2011-2020). Filling this research gap will help researchers and experts better understand software usability's efficiency, effectiveness, and development.

In an effort to fill the research gap and inclusion of significant works, this research followed systematic mapping methodology. This approach allows the research to capture key facts and details from literature using a well-defined process. To this end, an SMS protocol composed of search strategy, data extraction, selection criteria, and rejection criteria was formed. The main objective of this research is to investigate factors affecting software usability. Additionally, this researched also aimed at classifying the selected studies by knowing the existing contributions, research facets conducted, validation methods used, evaluation measures utilized, application domains, and lastly the overall demographics of the literature reviewed. The selected parameters will provide an overview of the general trends of the publications as well as the evolution of research in the domain of software usability.

In this study, the main contributions are as follows:

- A detailed examination and synthesis of key studies on software usability.
- The study reviews primary studies (PS) and identify their distinct contributions.
- The mapping study analyzes the overall research productivity, demographics, trends, and challenges in software usability.

- The research identifies area of research that are least addressed and provides directions for future research.

This research study is divided into six sections. Section II presents various aspects of software usability and discusses related surveys in the domain of software usability. Section III describes the research methodology including research questions, data acquisition and processing techniques. Results are presented in Section IV. Section V discusses the research findings as well as directions for future research, whereas threats to validity are discussed in Section VI. Section VII concludes the work.

II. Software Usability and Related Work

This section discusses software usability by reviewing articles published in the area.

A. Software Usability

Software usability is a key characteristic of software quality. As per ISO/IEC 9126, usability is defined as "a set of attributes that bear on the effort needed for use, and on the individual assessment of such use, by a stated or implied set of users" [3, 7]. Based on [7], software quality model is composed of six dimensions, which are functionality, portability, maintainability, efficiency, usability and reliability.

However, usability is inconsistently defined in the literature, where different standards and researchers have distinct definitions. In the following text, some definitions of software usability from various standards are outlined. In a study by Nielsen, the author defined usability based on five key dimensions. These dimensions are efficiency, errors, learnability, memorability, and satisfaction [22]. Efficiency means that the system should be efficient to be utilized by a user. When the system is efficiently utilized, productivity will increase. It is also expected that such systems have low error rates. Learnability is the ability to understand the system, and a good system is expected to have small learning curve. Memorability refers to the ability of the system to be easily remembered without requiring users to relearn things again. Lastly, satisfaction means that the user should feel gratified when using the system. Moreover, in the ISO 9241-11, usability is defined as the level where a product can be utilized by a given user in achieving a specific objective with efficiency, effectiveness, and satisfaction in a defined context [23, 24].

In software engineering, usability issues in software directly or indirectly contribute to software quality problems. These problems often cause low efficiency and effectiveness. Based on this, users encounter various difficulties when using a specific software [25]. Further, the usability issues can also lead to low acceptance rate of software applications [26].

B. Related Work

This section summarizes the existing literature in software usability and identifies the research gaps.

Bastien [27] focused on the growing applications of Information and Communication Technology (ICT) based medical devices incorporating human-computer interfaces. The author highlighted that usability of these medical devices' hinges on their ease of usability. The author identified inspection-, user- and model-based evaluations as standard industry approaches for validating usability. The work considered various scenarios and provided directions for further research in usability testing for mobile applications. Usability testing can be performed in a number of manners such as questionnaires, interviews, automated testing etc. Sure [21] considered questionnaire as a method for usability testing where information about usability, user satisfaction, expectations, software behavior and other related information is collected using a set of questions. The author identified thirteen questionnaires from literature. The analysis of the questionnaires revealed that 'satisfaction' is the most common factor that is measured by the majority of the questionnaires.

Esmeria and Seva [18] presented a survey of works in relation to the usability of the websites. The authors argued that the process of usability testing should result in a usability index that can describe the usability of a website. A drawback of the work is the lack of rigor in selection of data sources. The authors relied on Science Direct and ACM Digital Library only and no specific query string was mentioned. Sagar and Saha [3] conducted a review of the software usability and reached to a variety of conclusions. The authors identified that so far researchers have not been able to unanimously agree on usability models. Further, the work highlighted efficiency, effectiveness, satisfaction, and learnability as standard measures used by a majority of the usability models. In terms of usability evaluation methods, questionnaires, usability testing and heuristic evaluations are the leading mechanisms.

Maramba et al. [28] conducted a systematic review of usability testing of e-health applications. The authors observed that although there is an exponential increase in the development and subsequence deployment of e-health applications, very few of these applications have published their usability evaluation results. The work further highlighted that questionnaire has been the de-facto method for usability analysis and argued that more qualitative and automated methods must be used in usability evaluation. Key limitations of the work include the limited time frame for the selection of potential works (April 2014 – October 2017) and the scope of the work which is limited to e-health applications only. Weichbroth [29] used Scopus as data source to conduct a systematic literature review of usability aspects of mobile applications. The author identified 790 relevant documents spanning 2001 to 2018. It was observed that the usability definition as given in ISO 9241-11 is used by majority of the works. The work identified 75 attributes associated with the usability of mobile applications and identified efficiency, satisfaction and effectiveness as most important considerations as indicated in the literature. Memorability, cognitive load and errors are identified as least considered attributes. Like Maramba, et al. [28], and Weichbroth [29] observed that controlled observations and surveys are mostly used in usability studies and suggested the use of eye-tracking, thinking loud, and interviews to be used in the future usability studies. Besides these studies, other important works include Coursaris and Kim [30], Harrison, et al. [31], and Quiñones and Rusu [32].

TABLE I. SUMMARY OF THE SELECTED SURVEYS

S. No	Study and Year	Selected Databases	Keywords	No. of Selected Articles	Limitations
1	Harrison, et al. [31] (2013)	ACM Digital Library, Google Scholar, IEEE Xplore	Mobile application evaluations, mobile application usability evaluations, usability of mobile applications	131	Narrow scope focusing on mobile applications only Unavailability of 6% of the selected papers
2	Sure [21] (2014)	ACM Digital Library, Inspec	Usability, Questionnaire, Reliability, Validity	35	Generic query terms Use of Inspec and ACM Digital library for literature search Limited scope focusing on questionnaires for a usability study
3	Esmeria and Seva [18] (2017)	ACM Digital Library, ScienceDirect	Website usability evaluation, measures of web usability	42	Lack of rigor in selection of data sources Reliance on ACM Digital Library and Science Direct only
4	Quiñones and Rusu [32] (2017).	ACM Digital Library, Science Direct, IEEE Xplore, Springer Link, Scopus, Google Scholar	Usability heuristic(s), methodology, heuristic evaluation, formal process, usability design	73	Focused on usability heuristics only
5	Maramba, et al. [28] (2019)	ACM Digital Library, CINAHL, IEEE Xplore, Medline / PubMed.	eHealth, mHealth, usability	133	Limited focus on e-health applications only Very short time frame selection
6	Weichbroth [29] (2020)	Scopus	Usability, mobile applications	66	A narrow focus on mobile applications Use of single source for data collection

Table I presents a summary of the selected related works. Note that some works such as Bastien [27] and Coursaris and Kim [30] are not included in the table as these works did not use the standard methodology for conducting systematic literature review.

Despite the usefulness of these studies, none of the highlighted studies in this section conducted a systematic mapping study for software usability. Most papers study selection process is also arbitrary with no rigor or repeatability. Hence, we observed that there is no study in the software usability that categorize and analyze existing research with respect to their research facets, contribution facets, publication forums/trends, citation impacts, and so on. Thus, the aim of this research is to fill these gaps in the field.

III. METHODOLOGY

Systematic mapping studies (SMS) are conducted to provide a general overview of a research area by systematically classifying the existing works and identify the contributions of researchers in the area of study. The studies (SMS) largely explore current literature to examine the reporting of areas, publication frequency, research trends, and publication venues where the primary studies are published [33, 34]. There are many shared characteristics between a systematic literature review (SLR) and systematic mapping studies. Some of the characteristics are the use of evidence-based searching and study selection procedures. However, SMS has a distinct objective from SLR and took a different approach to data analysis. SMS is primarily aimed at mapping and structuring an area of study. Hence, in this study, the SMS follows the general guidelines suggested by Petersen, et al. [34] and Kitchenham and Brereton [35]. Consequently, this study follows the pathway of similar studies that adopt these guidelines [36-38]. Fig. 1 outlines the process for our study.

Fig. 1. The Systematic Mapping Process.

As depicted in Fig. 1, the SMS process comprises of five distinct stages. The initial stage is to define key research questions (RQs). In stage two, the search process is conducted by specifying the search terms for retrieving the primary studies (PS). In the third and fourth stages, the retrieved studies are screened to remove unnecessary and irrelevant studies. Lastly, the data extraction is conducted, and the systematic maps of the study are created. These stages are defined in the following text.

A. Research Questions

In this section, the research questions (RQs) of the study are presented with respect to our main research objective. The main objective of this SMS is to classify the selected studies by knowing the existing contributions, research facets, validation methods, evaluation measures, application domains, and the overall demographics of the selected studies. The key question of this SMS is "what is the state-of-the-art in software usability studies?". Based on the objective of the work, the key question is divided into six distinct RQs as presented in Table II.

B. Data Sources

A SMS heavily relies on the selected Primary Studies (PS). To achieve a good mapping, it is vital that the process of selecting the PS is conducted carefully. The literature on software usability was collected from 2011 to 2020. Six data sources were selected for our literature search. The data sources include ACM Digital Library, IEEE Xplore, Google Scholar, Science Direct, SpringerLink, and Taylor and Francis.

It is important to mention that the selected sources include most the publications in the area of usability. Scopus could not be included as the author was unable to obtain access to the restricted (subscription based) database of Scopus. However, the selected six sources for data retrieval are broad and comprehensive enough to provide coverage to most of the reputable publications' outlets. In addition, the selected sources are used in other studies such as Esmeria and Seva [18] and Maramba, et al. [28]. Table III highlights the data sources with respect to the studies identified in our initial result search. The initial search resulted in 9,874 studies, out of which 62 studies were shortlisted based selection criteria covered in Section 3.D.

TABLE II. THE DEFINED RESEARCH QUESTIONS

RQ#	Research Question	Motivation
RQ1	What are the demographic characteristics of the PS?	To identify publication trend, publication forums and citation impact of the primary studies.
RQ2	What contribution facets have the primary studies provided?	To identify the contribution facets (model, method, investigation, and so on).
RQ3	What are the research types (facet) focused on by the primary studies in the domain?	To ascertain the research facets (evaluation research, solution proposal, and so on) in the area of study.
RQ4	What validation methods are generally utilized for software usability evaluation?	To identify various software usability evaluation methods.
RQ5	What are the various application domains for usability?	To investigate the domains of application for usability.
RQ6	What are the evaluation measures used by the PS?	To identify the evaluation measures used by the selected studies.

TABLE III. STUDIES IDENTIFIED IN EACH DATA SOURCES

Data Source		Initial Results of Search	Final Selected Studies
ID	*Name*		
D1	IEEE Xplore	729	18
D2	Science Direct	1,208	15
D3	Taylor and Francis	76	12
D4	ACM	957	7
D5	SpringerLink	467	5
D6	Google Scholar	6,437	5
Total		9,874	62

C. Search Terms

We performed automatic searches to retrieve important studies from our selected data sources. This is achieved using our search string or terms developed based on the guidelines of Petersen, et al. [34]. Basically, a search string is a composition of characters used by a researcher to identify the most relevant set of documents from a data source. Therefore, selecting the right search string is imperative because the outcome is connected to the information given by the data source. Hence, the selection of right terms requires careful attention to ensure important studies are not missed in the mapping process. In doing so, a generic search string was formulated to work on all the data sources. The search string is outlined as follows.

((Software usability AND Usability models) OR (Usability metrics))

It is important to point out that the search term is relatively generic to ensure that maximum results are obtained. A carefully crafted inclusion and exclusion criterion is then applied to shortlist the publications for further study. After execution of the search terms on respective data sources, the resultant publications obtained from each of the data source are reflected in Table III. A total of 9,874 publications were retrieved. Note that the search query was executed to collect the raw data from 19 November 2020 to 25 December 2020.

D. Inclusion and Exclusion Criterion

When the search results utilizing the articulated search string are acquired, we anticipated that works that are not important to the objective of this SMS study can also be retrieved. If so happens, we cautiously designed a clear inclusion and exclusion criteria that will be used on the retrieved studies to eliminate those that are not in-line with the objective of the paper. The inclusion-exclusion criteria for this study are outlined as follows;

Inclusion Criteria:

- Include studies on software usability
- Include studies that were published in the past 10 years (2011 - 2020)
- Include only peer-reviewed studies

Exclusion Criteria:

- Exclude survey and review studies
- Short Papers
- Editorials
- Summaries of keynotes
- Exclude the studies that are not on software usability
- Exclude the studies that are not written in English

After a careful implementation of inclusion and exclusion criterion, and thorough manual analysis including removal of duplicates, we identified 62 primary studies for further analysis.

E. Extraction of Data

To extract data from each of the 62 primary studies for answering the formulated RQs, a systematic data extraction method has to be clearly defined. We created a form to extract important data from the 62 identified articles for this study. The author as well as the two volunteers filled out the form for each of the 62 selected papers. For each publication, title, publication venue, research type, contributions, validation method, publication year, evaluation measures and application domain were recorded.

F. Classification Scheme

We followed Petersen, et al. [34] to develop the classification scheme for this study. The 62 final selected studies were examined by their titles, abstracts, keywords, research contributions, theoretical models, and general demographics. These studies were comprehensively studied for a thorough understanding of various characteristics of the classification. The first step is to classify the PS into the contributions made by various researchers. These contributions are investigative study, evaluation study, model, framework, application, scheme, method, usability concepts, usability principles, approach, and system. We further extended the classification to identify the research facets. These facets are experience papers, evaluation research, solution proposals, and validation research. These classifications are standard and in line with the existing literature for conducting a mapping study [34, 38]. Subsequently, further classifications of the PS were done with respect to the validation methods used to validate software usability, evaluation measures, application domains, and the selected studies general demographic characteristics.

IV. Results

This section covers the results of the research. The RQs formulated are all answered by critically analyzing the PS studies. Table IV presents the PS for this study. For the sake of brevity, only research article IDs is provided. For mapping between IDs and research articles description, the reader is referred to Appendix A.

TABLE IV. Overview of Selected Studies

PS Paper ID	Year of Publication	Publication Channel	Citation Count	Contribution
B1	2020	IEEE	2	Approach
B2	2017	ACM	5	Model
B3	2017	ACM	2	Model
B4	2019	ACM	0	Evaluation
B5	2020	Springer	1	Evaluation
B6	2019	Springer	1	Metrics
B7	2017	IEEE	4	Evaluation
B8	2017	IEEE	1	Evaluation
B9	2017	IEEE	3	Investigation
B10	2018	ACM	2	Investigation
B11	2018	IEEE	0	Investigation
B12	2018	IEEE	1	Method
B13	2016	Taylor and Francis	13	Usability concepts
B14	2015	Taylor and Francis	62	Investigation
B15	2016	Taylor and Francis	10	Investigation
B16	2016	Taylor and Francis	3	Evaluation
B17	2015	Taylor and Francis	42	Investigation
B18	2011	Elsevier	92	Investigation
B19	2013	Taylor and Francis	78	Usability principles
B20	2016	Taylor and Francis	2	Investigation
B21	2012	Elsevier	284	Investigation
B22	2013	IEEE	78	Usability guidelines
B23	2011	Elsevier	74	Evaluation
B24	2012	Elsevier	75	Investigation
B25	2013	Elsevier	63	Investigation
B26	2011	ACM	17	Model
B27	2020	IEEE	0	Evaluation
B28	2020	IEEE	1	Investigation
B29	2011	Independent	60	Investigation
B30	2013	Springer	23	Evaluation
B31	2016	Springer	103	Evaluation
B32	2014	Independent	113	Evaluation
B33	2019	IEEE	1	Evaluation
B34	2019	IEEE	0	Investigation
B35	2018	IEEE	0	Model
B36	2018	IEEE	1	Investigation
B37	2019	IEEE	1	Model
B38	2013	Elsevier	66	Framework
B39	2019	IEEE	4	System
B40	2014	Elsevier	24	Scheme
B41	2020	IEEE	0	Investigation
B42	2015	Independent	1	Approach
B43	2019	IEEE	0	Investigation
B44	2014	Elsevier	34	Evaluation
B45	2015	Taylor and Francis	4	Investigation
B46	2015	Taylor and Francis	118	Evaluation
B47	2012	Elsevier	25	Application
B48	2015	Elsevier	20	Model
B49	2012	Taylor and Francis	83	Evaluation
B50	2013	Springer	12	Evaluation
B51	2012	ACM	20	Investigation
B52	2013	Taylor and Francis	14	Investigation
B53	2013	IEEE	1	Model
B54	2011	ACM	8	Evaluation
B55	2013	Taylor and Francis	33	Evaluation
B56	2015	Elsevier	15	Application
B57	2013	Independent	3	Model
B58	2015	Elsevier	29	Investigation
B59	2013	Elsevier	8	Investigation
B60	2015	Elsevier	21	Approach
B61	2013	Elsevier	47	Evaluation
B62	2011	Independent	1	Investigation

A. RQ1. What are the Demographics Characteristics of the PS?

In answering this RQ, the primary studies were analyzed critically with the purpose of answering the RQ. Three aspects of the PS were analyzed including publication trend, publication forums, and citation impact.

Publication trend: From 2011 to 2020, 62 studies were retrieved from the data sources. In Fig. 2, the year-wise publications in the domain of software usability are graphically presented. We observed that in 2013 and 2015, more studies were published with 12 and 9 studies, which are the most active years in the research domain. 2019 was also moderately active, with 7 studies. In general, even though the number of studies is linear, the research output continues to stabilize with stable yearly publication.

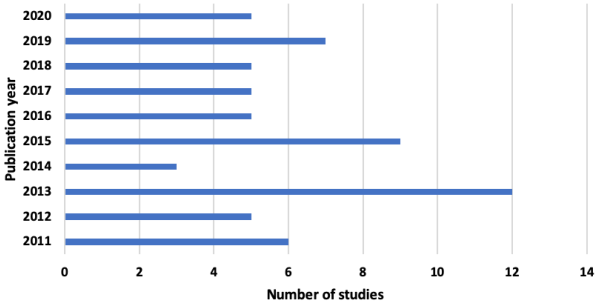


Fig. 2. Publication Trend Per Year.

Publication forums: This SMS study covered 20 different journals, 19 different conference proceedings, and 1 symposium and workshop each, respectively (see Appendix B). 39 papers are published in journals, 21 in conferences and one study each in workshop and symposium. From the analysis, we found that the International Journal of Human-Computer Interaction and the International Journal of Human-Computer Studies were the venues that contribute the most with 7 and 6 publications respectively followed by Journal of Systems and Software and Behavior and Information Technology with 4 and 2 publications each.

Citation impact: In Table V, the number of citations for the top 10 most cited papers is presented. We obtained the citation count of each paper from Google Scholar. Hence, the citation count may or will change at any point in time. In general, from our PS, we found four papers that have more than 100 citations each, which are Lee and Kozar [39], Kortum and Sorber [40], Mirkovic, et al. [41], and Maitama, et al. [36]. The total number of citations from the PS was 1809 as presented in Table IV. Therefore, the average number of citations per paper is 29.17.

B. RQ2. What Contribution Facets have the PS Provided?

To answer the RQ, we conducted a thorough analysis of the selected PS. Based on the analysis, we found 13 key contributions. These contributions are summarized in Fig. 3. The most significant contributions are Investigation (23 papers), Evaluation (18 papers), and Model (8 papers). 37% of the PS conducted an investigative study on software usability, followed by evaluation (29%), and model (13%), respectively. The rest of the contributions have less than 5% coverage. From our analysis, we observed that majority of the studies conducted an investigation into usability of existing web applications (B6, B15, B8, B38) or usability evaluation models (B37, B57, B53). Other studies focused more on proposing new models to help in facilitating or understanding key factors that facilitate or hinder usability of software or web application.

TABLE V. TOP CITED PAPERS

PS Paper ID	Paper Title	Citation	Year
B21	Understanding of website usability: Specifying and measuring constructs and their relationships	284	2012
B46	Measuring the usability of mobile applications for phones and tablets	118	2015
B32	Supporting cancer patients in illness management: usability evaluation of a mobile app	113	2014
B31	Usability evaluation of mobile applications using ISO 9241 and ISO 25062 standards	103	2016
B18	Reliability, validity, and sensitivity of a single-item measure of online store usability	92	2011
B49	A comparison of usability evaluation methods for evaluating e-commerce websites	83	2012
B19	Usability principles for augmented reality applications in a smartphone environment	78	2013
B22	Usability through software design	78	2013
B23	Aesthetics and usability of in-vehicle navigation displays	74	2011
B24	How do usability professionals construe usability?	75	2012

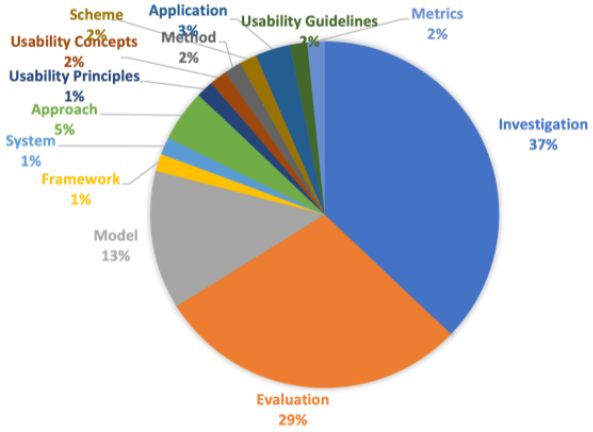


Fig. 3. Contributions by the Selected Studies.

Generally, we observed that a majority of the studies are not tailored to solution proposals, rather, they are focused on understanding software usability and ways to understand the factors that hinders users' acceptability and understandability of a software system. Despite the fact that software usability is not a recent research area, we observed that both investigative studies and evaluation studies are gaining attention from the researchers in this domain. This trend should be tailored into the proposition of new ways (in terms of framework, method, models, and so on) to help solve the usability issue in software engineering rather than just investigations and general evaluations. However, this is understandable because usability issues need to be understood using investigative and evaluation approaches.

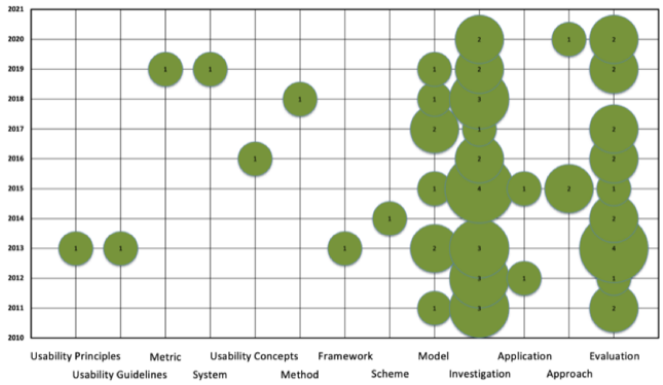


Fig. 4. Mapping of the Yearly Distribution of the Identified Contributions.

Fig. 4 shows the contributions on yearly basis. We observed that in 2018, no evaluation study was conducted. We further observed that from 2011 onward, research output is consistent and predictable. However, contributions such as framework, usability principles, usability guidelines, metric, system, usability concepts, method, scheme, and approach have received less attention in the field of study.

C. RQ3. What are the Research Types (Facet) Focused by the PS in the Domain?

From the PS, we identified four key research facets. These research facets are experience paper, evaluation research, solution proposal, and validation research. Experience papers are intended to give an explanation of how issues are handled in practice. Hence, it is generally the personal experience of the authors conducting the work. An example of the experience paper is Komiyama, et al. [42] where the authors attempted to obtain the developers point of view on the usability of intensive software system. Solution proposal seeks to provide new and novel solutions to an established problem in a research domain. The example of a solution proposal is Diaz, et al. [43]. In their work, the authors proposed a new usability metrics for e-commerce website. Evaluation research is conducted to understand how a method is implemented in practice. An example of evaluation research is the work of Al-Maani and Salameh [44]. Validation research papers present a novel proposal that is not fully implemented in practice. Examples of validation research are Störrle [45], and Christophersen and Konradt [46].

From our analysis, as presented in Table VI, we found that most of the studies conducted Evaluation research with 33% of the PS. Furthermore, experience paper constitutes 29% of the total publications, followed by solution proposal with 24%, and validation research with 13%. The analysis shows that there is an urgent need for more solution proposals and research to validate the proposed proposals. Fig. 5 depicts the map for the identified research facets in correspondence to the validation methods. We observe that more experience papers are needed in this research domain to understand users' perspective with respect to usability issues.

D. RQ4. What Validation Methods are Generally Utilized for Software Usability Evaluation?

Validation methods are key to evaluating one's work in any scientific work. In answering this RQ, we identify eight validation methods utilized by the respective PS. These methods are experiment (22 studies), theoretical validation (10), case study (5), questionnaire (3), interview (3), interview and questionnaire (2), simulation (2), and field study (1). We observed that experiment and theoretical validation are the most used approaches by the PS in this domain. In Table VII, the identified validation methods with respect to the studies that used them are highlighted.

TABLE VI. RESEARCH FACETS

Research Facet	Studies	No. of Studies	%
Evaluation Research	B3, B40, B52, B54, B50, B61, B57, B16, B44, B7, B45, B32, B55, B2, B23, B30, B31, B33, B38, B46, B49	21	33%
Experience Paper	B5, B36, B41, B15, B43, B58, B59, B4, B29, B9, B24, B11, B14, B17, B20, B34, B51, B62	18	29%
Solution Proposal	B6, B37, B48, B19, B42, B26, B39, B53, B1, B12, B13, B22, B47, B56, B60	15	24%
Validation Research	B10, B18, B21, B25, B27, B28, B8, B35	8	13%

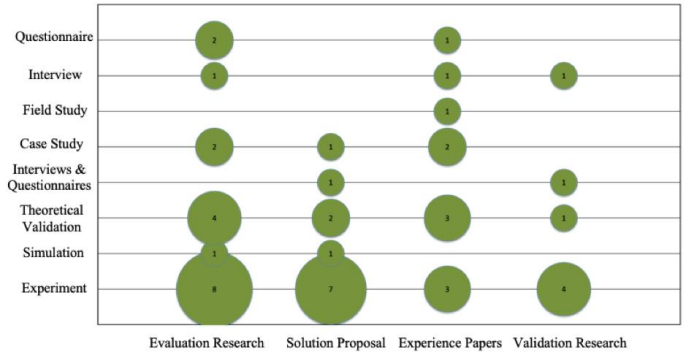


Fig. 5. Map for Research Facets against Validation Methods.

TABLE VII. VALIDATION METHODS

Research Facet	Studies	No. of Studies
Experiment	B46, B22, B60, B1, B30, B31, B21, B9, B10, B18, B29, B25, B56, B47, B2, B33, B4, B23, B49, B12, B13, B38	22
Not clearly defined	B57, B53, B26, B35, B61, B50, B51, B17, B11, B62, B34, B20, B14, B39	14
Theoretical Validation	B19, B42, B41, B36, B52, B3, B40, B27, B5, B54	10
Case Study	B59, B45, B58, B37, B7	5
Questionnaire	B15, B44, B16	3
Interview	B24, B8, B32	3
Interview and Questionnaire	B6, B28	2
Simulation	B48, B55	2
Field Study	B43	1

E. RQ5. What are the Various Application Domains for Usability?

In answering this RQ, we identified six application domains in the research area, which are web (17 studies), software development (16), mobile (8), industry and research (6), navigation (3), and robotics (1). We identified 11 studies that have no well-defined domain. Web is the most considered domain with 27% of the PS, followed by software development (25%). Table VIII presents the application domains in the research area.

TABLE VIII. APPLICATION DOMAINS

Application Domain	Studies	No. of Studies
Web	B6, B15, B8, B55, B37, B7, B30, B21, B18, B29, B56, B2, B49, B13, B38, B3, B61	17
Software Development	B43, B44, B22, B60, B10, B42, B41, B40, B27, B54, B53, B26, B51, B11, B34, B14	16
No Clear Domain Identified	B45, B58, B48, B24, B28, B12, B36, B57, B35, B50, B62	11
Mobile	B32, B46, B1, B31, B47, B33, B19, B52	8
Industry & Research	B16, B9, B4, B5, B17, B20	6
Navigation	B59, B25, B23	3
Robotics	B39	1

F. RQ6. What are the Evaluation Measures used by the PS?

With respect to evaluation metrics, we observed that most of the studies (16) used usability as a metric. This is followed by effectiveness (4 studies) and learnability (4 studies). This research observed that some studies did not clearly identified the evaluation metric that were utilized by the research. This observation is quite alarming, because almost of the primary studies have no well-defined evaluation measures in their studies. Hence, this need to be address and more work need to clarify or use an evaluation metric for their study.

V. DISCUSSION

The discussion in this section comprises of two facets. Firstly, the main findings of the study are summarized and presented clearly. It is followed by highlighting areas in usability that have not received considerable attention from the researchers.

From 2011 - 2020, the research in software usability is generally stable. Most of the PS are published in journals (39), followed by conferences (21), workshop and symposium (1 each). We found that International Journal of Human-Computer Interaction and the International Journal of Human-Computer Studies are the venues that contributed the most with 7 and 6 publications, respectively. With respect to contribution facets, multiple key contributions were identified. 37% of the PS conducted an investigative study on software usability, followed by evaluation with 29% of the PS, and model with 13%, respectively.

From our analysis, we observed that majority of the studies focused on usability testing of existing web applications (such as B6, B15, B8, B38) or usability evaluation models (such as B37, B57, B53). Other studies focused on proposing new models to help in understanding key factors that facilitate or hinder usability of software or web application. We observed that a majority of the studies are not tailored to solution proposals, rather, they are anchored in the direction of understanding software usability and ways to understand the factors that hinders users' acceptability and understandability of a software system.

From the PS, we identified four key research facets which are evaluation papers (33%), experience papers (29%), solution proposal (24%), and validation research (13%). We also identified eight validation methods utilized by the respective PS. These methods are experiment with 22 studies, followed by theoretical validation (10), case study (5), questionnaire (3), interview (3), interview and questionnaire (2), simulation (2), and field study (1). The research also identified three major evaluation metrics for evaluating proposals, which are used in this area. These metrics are usability, effectiveness and learnability. The proposals have been evaluated using a combination of metrics and sub-metrics as well. Some of the primary studies did not use any evaluation measure for validation. Hence, this needs to be addressed.

From our study, we observed that although a significant research is conducted to improve the usability of the software applications, there are still a number of research directions that needed to explore in more detail. For instance, when evaluating usability of software, majority of the works relied on closed experiments, questionnaires and surveys. It is recommended that future studies should focus more on field studies where an application usability is evaluated in real world conditions. Field studies are more beneficial as these are conducted by the end users in real operational environment resulting in identification of problems that might be overshadowed in laboratory-based tests. However, conducting the field-based study can be expensive both in terms of time and monetary value. An important aspect of software testing is the reproducibility of the errors/bugs. It is important to consider reproducibility of usability testing as a dimension in usability metrics.

Usability testing are used in a variety of domain such as aviation [47, 48], banking [49], bioinformatics [20] and medicine [28]. One area that is neglected by the researcher is the usability of software used by the scientific community. Most of the software applications developed for the use of scientific community are either command line or has poor user interfaces and faces numerous usability challenges. It will be important to consider usability testing of the scientific software and design a set of standard guidelines.

Ethics is an important aspect of software engineering process [50], however ethics in usability testing has received the least consideration by the researcher. It will be of interest to examine compliance of various ethical guidelines in usability testing. The ethical dimension of the usability testing is more important when considering the applications in domains such as health and finance.

Although considerable efforts are devoted in the literature to investigate the usability of mobile applications, we could not find any studies that focused on the cross-platform usability of

various mobile applications. Generally, mobile applications are developed for different platforms (notably android and iOS). It will be helpful to investigate how the usability of mobile applications varies across different platforms.

In the recent past machine learning applications have seen widespread use in various domains [51, 52]. However, there is limited work to evaluate the usability aspect of machine learning applications [53]. As the machine learning applications are adopted for more widespread use, it is recommended to conduct usability testing of these applications. Although, open source software has gained traction and acceptance, there is little work to assess the usability of such systems [54]. It will be important to assess the usability of open source software systems.

VI. Threat to Validity

After the rigorous analysis of the primary studies, this section discusses some identified issues that can be regarded as threat to validity. These limitations are discussed as follows.

Selection bias can be regarded as an external threat to the validity of our study. With respect to selection bias, even though the selected data sources were thoroughly searched by the author, there is a possibility that some important studies might be missed. To reduce this bias, this study employs inclusion and exclusion criteria for the selection of quality papers. The study is conducted by a single author which might result in selection bias. In order to mitigate the effect of the personal selection bias, two volunteer researchers oversaw the process and provided the feedback on the selection of final research papers for the study. Therefore, the threat level of personal selection bias is mitigated.

Conclusion validity is minimized by drawing all the relationships and conclusions from the literature based on the search query and then analyzing them using statistics such as citation count, classification into various sub-fields etc. Publication bias was mitigated by searching six important data sources. We also employed forward and backward snowballing techniques to make sure that all relevant studies related to software usability is identified, properly vetted and considered.

With respect to misclassification, some primary studies showed a limited information. Hence, some information was inferred while other where classified as “Nil” during the classification process. Hence, the inadequacy of information during classification may result in bias. In such a situation, the general methodology of a given study is carefully considered including the experimental setup to infer other classification entities. Therefore, this threat is mitigated to a certain level.

The selection of keywords can be regarded as threat of construct validity. We selected generalized keywords to identify a larger set of research papers in our domain from the various databases. Although it helped in reducing the probability of missing a relevant article, it has resulted in a large number of hits. In order to ensure that only relevant papers are selected for the study, a careful approach involving two volunteers as well as a carefully crafted approach is employed ensuring selection of relevant papers for the study.

Finally, the data was retrieved from the six selected data sources from 19 November 2020 to 25 December 2020. Therefore, if the same search query is executed at any later stage, different results might be obtained. Likewise, the citation count can be different as well as the research articles might have accrued more citations since 25th December 2020.

VII. Conclusion

This work presented a mapping study that analyzed research work from 2011 to 2020 in the domain of software usability. From an initial pool of 9,874 papers, 62 papers were carefully selected based on our inclusion/exclusion criterion. This study examined the existing contributions, research facets, evaluation measures, validation methods, application domains, demographics, publication trends, and publication forums in the research domain. With respect to contributions, investigative studies and evaluation studies are the two most common approaches. We identified four key research facets, which are experience paper, solution proposal, evaluation research, and validation research. We also identified eight validation methods utilized by the respective PS. Usability, effectiveness and learnability are found to be the common evaluation metrics. Rather alarmingly, some of the primary studies have no clear evaluation metrics defined. Hence, this need to be address and more work need to clarify or use an evaluation metric for their study.

In conclusion, the aim of this mapping study was to allow researchers and experts to have a clear understanding of the general research productivity, trends and demographics that shaped the research domain of software usability. This work will help in highlighting potential opportunities for both new and experienced researchers to conduct more works with the aim of improving the research domain.

References

[1] D. Gupta, A. K. Ahlawat, A. Sharma, and J. J. P. C. Rodrigues, "Feature selection and evaluation for software usability model using modified moth-flame optimization," *Computing*, vol. 102, pp. 1503-1520, 2020/06/01 2020. https://doi.org/10.1007/s00607-020-00809-6.

[2] K. Curcio, R. Santana, S. Reinehr, and A. Malucelli, "Usability in agile software development: A tertiary study," *Computer Standards & Interfaces*, vol. 64, pp. 61-77, 2019/05/01/ 2019. https://doi.org/10.1016/j.csi.2018.12.003.

[3] K. Sagar and A. Saha, "A systematic review of software usability studies," *International Journal of Information Technology*, 2017/12/11 2017. https://doi.org/10.1007/s41870-017-0048-1.

[4] T. Alahmadi and S. Drew, "Subjective Evaluation of Website Accessibility and Usability: A Survey for People with Sensory Disabilities," presented at the Proceedings of the 14th International Web for All Conference, Perth, Western Australia, Australia, 2017. https://doi.org/10.1145/3058555.3058579.

[5] B. Aryana and T. Clemmensen, "Mobile Usability: Experiences From Iran and Turkey," *International Journal of Human–Computer Interaction*, vol. 29, pp. 220-242, 2013/03/01 2013. 10.1080/10447318.2013.765760.

[6] A. D. Nuovo, S. Varrasi, D. Conti, J. Bamsforth, A. Lucas, A. Soranzo, *et al.*, "Usability Evaluation of a Robotic System for Cognitive Testing," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Daegu, Korea (South), 2019, pp. 588-589. https://doi.org/10.1109/HRI.2019.8673187.

[7] I. ISO, "Information technology-software product evaluation-quality characteristics and guide lines for their use," *Iso/iec is*, vol. 9126, 1991.

[8] K. Bershinsky and R. Narciso, "An Experimental Framework for Determining the Usability of Mixed Reality Interfaces for Aerospace Applications," in *AIAA Scitech 2019 Forum*, ed. https://doi.org/10.2514/6.2019-0064.

[9] O. V. Bitkina, H. K. Kim, and J. Park, "Usability and user experience of medical devices: An overview of the current state, analysis methodologies, and future challenges," *International Journal of Industrial Ergonomics*, vol. 76, p. 102932, 2020/03/01/ 2020. https://doi.org/10.1016/j.ergon.2020.102932.

[10] M. A. Razzak and M. N. Islam, "Exploring and Evaluating the Usability Factors for Military Application: A Road Map for HCI in Military Applications," *Human Factors and Mechanical Engineering for Defense and Safety*, vol. 4, p. 4, 2020/01/10 2020. https://doi.org/10.1007/s41314-019-0032-6.

[11] H. M. Az-zahra, N. Fauzi, and A. P. Kharisma, "Evaluating E-marketplace Mobile Application Based on People At the Center of Mobile Application Development (PACMAD) Usability Model," in *2019 International Conference on Sustainable Information Engineering and Technology (SIET)*, Lombok, Indonesia, 2019, pp. 72-77. https://doi.org/10.1109/SIET48054.2019.8986067.

[12] R. P. Bringula, "Factors Affecting Web Portal Information Services Usability: A Canonical Correlation Analysis," *International Journal of Human–Computer Interaction*, vol. 32, pp. 814-826, 2016/10/02 2016. https://doi.org/10.1080/10447318.2016.1199180.

[13] M. Bures, M. Macik, B. S. Ahmed, V. Rechtberger, and P. Slavik, "Testing the Usability and Accessibility of Smart TV Applications Using an Automated Model-Based Approach," *IEEE Transactions on Consumer Electronics*, vol. 66, pp. 134-143, 2020. https://doi.org/10.1109/TCE.2020.2986049.

[14] M. Hertzum and T. Clemmensen, "How do usability professionals construe usability?," *International Journal of Human-Computer Studies*, vol. 70, pp. 26-42, 2012/01/01/ 2012. https://doi.org/10.1016/j.ijhcs.2011.08.001.

[15] W. Isa, M. R. Suhami, N. I. Safie, and S. S. Semsudin, "Assessing the usability and accessibility of Malaysia e-government website," *American Journal of Economics and Business Administration*, vol. 3, pp. 40-46, 2011. https://doi.org/10.3844/ajebasp.2011.40.46.

[16] W. P. N. H. Pathirana and D. N. Wickramaarachchi, "Software usability improvements for Generation Z oriented software application," in *2019 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, Colombo, Sri Lanka, 2019, pp. 151-157. https://doi.org/10.23919/SCSE.2019.8842779.

[17] M. V. Waardhuizen, J. McLean-Oliver, N. Perry, and J. Munko, "Explorations on Single Usability Metrics," presented at the Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland Uk, 2019. https://doi.org/10.1145/3290607.3299062.

[18] G. J. Esmeria and R. R. Seva, "Web usability: a literature review," in *DLSU Research Congress*, Manila, Philippines, 2017.

[19] A. Linja, "Review of usability assessment tools and standards," Michigan Technological University, Applied Cognitive Science and Human Factors, Michigan Technological University2019.

[20] S. Mangul, L. S. Martin, E. Eskin, and R. Blekhman, "Improving the usability and archival stability of bioinformatics software," *Genome Biology*, vol. 20, p. 47, 2019/02/27 2019. https://doi.org/10.1186/s13059-019-1649-8.

[21] M. Sure, "Questionnaires for usability: A systematic literature review," Independent thesis Advanced level (degree of Master (Two Years)), Department of Computer and Information Science, Human-Centered systems. , Linköping University, Linköping University, 2014.

[22] J. Nielsen, *Usability engineering*, 1st edition ed. Harcourt Place, 32 Jamestown Road, London, NW1 7BY, UK: Academic Press, Inc, 1993.

[23] N. Bevan, J. Carter, and S. Harker, "ISO 9241-11 Revised: What Have We Learnt About Usability Since 1998?," in *17th International Conference, HCI International*, Los Angeles, CA, USA, 2015, pp. 143-151. https://doi.org/10.1007/978-3-319-20901-2_13.

[24] J. M. Ferreira, S. T. Acuña, O. Dieste, S. Vegas, A. Santos, F. Rodríguez, *et al.*, "Impact of usability mechanisms: An experiment on efficiency, effectiveness and user satisfaction," *Information and Software Technology*, vol. 117, p. 106195, 2020/01/01/ 2020. https://doi.org/10.1016/j.infsof.2019.106195.

[25] R. Abiri, S. Borhani, J. Kilmarx, C. Esterwood, Y. Jiang, and X. Zhao, "A Usability Study of Low-Cost Wireless Brain-Computer Interface for Cursor Control Using Online Linear Model," *IEEE Transactions on Human-Machine Systems*, vol. 50, pp. 287-297, 2020. https://doi.org/10.1109/THMS.2020.2983848.

[26] F. P. Tulinayo, P. Ssentume, and R. Najjuma, "Digital technologies in resource constrained higher institutions of learning: a study on students' acceptance and usability," *International Journal of Educational Technology in Higher Education*, vol. 15, p. 36, 2018/09/27 2018. https://doi.org/10.1186/s41239-018-0117-y.

[27] J. M. C. Bastien, "Usability testing: a review of some methodological and technical aspects of the method," *International Journal of Medical Informatics*, vol. 79, pp. e18-e23, 2010/04/01/ 2010. https://doi.org/10.1016/j.ijmedinf.2008.12.004.

[28] I. Maramba, A. Chatterjee, and C. Newman, "Methods of usability testing in the development of eHealth applications: A scoping review," *International Journal of Medical Informatics*, vol. 126, pp. 95-104, 2019/06/01/ 2019. https://doi.org/10.1016/j.ijmedinf.2019.03.018.

[29] P. Weichbroth, "Usability of Mobile Applications: A Systematic Literature Study," *IEEE Access*, vol. 8, pp. 55563-55577, 2020. https://doi.org/10.1109/ACCESS.2020.2981892.

[30] C. K. Coursaris and D. J. Kim, "A meta-analytical review of empirical mobile usability studies," *Journal of Usability Studies*, vol. 6, pp. 117-171, 2011.

[31] R. Harrison, D. Flood, and D. Duce, "Usability of mobile applications: literature review and rationale for a new usability model," *Journal of Interaction Science*, vol. 1, p. 1, 2013/05/07 2013. https://doi.org/10.1186/2194-0827-1-1.

[32] D. Quiñones and C. Rusu, "How to develop usability heuristics: A systematic literature review," *Computer Standards & Interfaces*, vol. 53, pp. 89-122, 2017/08/01/ 2017. https://doi.org/10.1016/j.csi.2017.03.009.

[33] I. Ahmad, G. Ahmed, S. A. A. Shah, and E. Ahmed, "A decade of big data literature: analysis of trends in light of bibliometrics," *The Journal of Supercomputing*, vol. 76, pp. 3555-3571, 2020/05/01 2020. https://doi.org/10.1007/s11227-018-2714-x.

[34] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Information and Software Technology*, vol. 64, pp. 1-18, 2015/08/01/ 2015. https://doi.org/10.1016/j.infsof.2015.03.007.

[35] B. Kitchenham and P. Brereton, "A systematic review of systematic review process research in software engineering," *Information and Software Technology*, vol. 55, pp. 2049-2075, 2013/12/01/ 2013. https://doi.org/10.1016/j.infsof.2013.07.010.

[36] J. Z. Maitama, N. Idris, and A. Zakari, "A Systematic Mapping Study of the Empirical Explicit Aspect Extractions in Sentiment Analysis," *IEEE Access*, vol. 8, pp. 113878-113899, 2020. https://doi.org/10.1109/ACCESS.2020.3003625.

[37] S. Ouhbi, A. Idri, J. L. Fernández-Alemán, and A. Toval, "Requirements engineering education: a systematic mapping study," *Requirements Engineering*, vol. 20, pp. 119-138, 2015/06/01 2015. https://doi.org/10.1007/s00766-013-0192-5.

[38] A. Zakari, S. P. Lee, K. A. Alam, and R. Ahmad, "Software fault localisation: a systematic mapping study," *IET Software*, vol. 13, pp. 60-74, 2019. https://doi.org/10.1049/iet-sen.2018.5137.

[39] Y. Lee and K. A. Kozar, "Understanding of website usability: Specifying and measuring constructs and their relationships," *Decision Support Systems*, vol. 52, pp. 450-463, 2012/01/01/ 2012. https://doi.org/10.1016/j.dss.2011.10.004.

[40] P. Kortum and M. Sorber, "Measuring the Usability of Mobile Applications for Phones and Tablets," *International Journal of Human–Computer Interaction*, vol. 31, pp. 518-529, 2015/08/03 2015. https://doi.org/10.1080/10447318.2015.1064658.

[41] J. Mirkovic, D. R. Kaufman, and C. M. Ruland, "Supporting Cancer Patients in Illness Management: Usability Evaluation of a Mobile App," *JMIR Mhealth Uhealth*, vol. 2, 2014. https://doi.org/10.2196/mhealth.3359.

[42] T. Komiyama, S. i. Fukuzumi, M. Azuma, H. Washizaki, and N. Tsuda, "Usability of Software–Intensive Systems from Developers' Point of View," in *22nd HCII: International Conference on Human-Computer Interaction*, Copenhagen, Denmark, 2020, pp. 450-463. https://doi.org/10.1007/978-3-030-49059-1_33.

[43] E. Diaz, S. Flores, and F. Paz, "Proposal of Usability Metrics to Evaluate E-commerce Websites," in *21st HCI International Conference on Design, User Experience, and Usability. Practice and Case Studies*, Orlando, FL, USA, 2019, pp. 85-95. https://doi.org/10.1007/978-3-030-23535-2_6.

[44] D. I. Al-Maani and H. B. Salameh, "A generic model for evaluating the usability of learning management systems," presented at the Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing, Cambridge, United Kingdom, 2017. https://doi.org/10.1145/3018896.3018921.

[45] H. Störrle, "Improving model usability and utility by layered diagrams," presented at the Proceedings of the 10th International Workshop on Modelling in Software Engineering, Gothenburg, Sweden, 2018. https://doi.org/10.1145/3193954.3193958.

[46] T. Christophersen and U. Konradt, "Reliability, validity, and sensitivity of a single-item measure of online store usability," *International Journal of Human-Computer Studies*, vol. 69, pp. 269-280, 2011/04/01/ 2011. https://doi.org/10.1016/j.ijhcs.2010.10.005.

[47] J. McSorley, J. Kleber, and B. Blickensderfer, "Usability Analysis of Aviation Weather Products for General Aviation," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, pp. 1903-1907, 2019/11/01 2019. https://doi.org/10.1177/1071181319631339.

[48] H. Xue, T. Li, X. Zhang, and R. Wang, "Integrated Usability Evaluation Method for Cockpit of Civil Aircraft," in *17th International Conference on Man–Machine–Environment System Engineering*, Jinggangshan, China, 2018, pp. 745-752. https://doi.org/10.1007/978-981-10-6232-2_89.

[49] I. Aboobucker and Y. Bao, "What obstruct customer acceptance of internet banking? Security and privacy, risk, trust and website usability and the role of moderators," *The Journal of High Technology Management Research*, vol. 29, pp. 109-123, 2018/01/01/ 2018. https://doi.org/10.1016/j.hitech.2018.04.010.

[50] I. Ozkaya, "Ethics Is a Software Design Concern," *IEEE Software*, vol. 36, pp. 4-8, 2019. https://doi.org/10.1109/MS.2019.2902592.

[51] I. Ahmad, M. A. Alqarni, A. A. Almazroi, and A. Tariq, "Experimental Evaluation of Clickbait Detection Using Machine Learning Models," *Intelligent Automation And Soft Computing*, vol. 26, pp. 1335-1344, 2020. https://doi.org/10.32604/iasc.2020.013861.

[52] I. Ahmad, M. Hamid, S. Yousaf, S. T. Shah, and M. O. Ahmad, "Optimizing Pretrained Convolutional Neural Networks for Tomato Leaf Disease Detection," *Complexity*, vol. 2020, p. 8812019, 2020/09/23 2020. https://doi.org/10.1155/2020/8812019.

[53] F. Bernardo, M. Zbyszyński, M. Grierson, and R. Fiebrink, "Designing and Evaluating the Usability of a Machine Learning API for Rapid Prototyping Music Technology," *Frontiers in Artificial Intelligence*, vol. 3, 2020-April-03 2020. https://doi.org/10.3389/frai.2020.00013.

[54] K. A. Dawood, K. Y. Sharif, A. A. Ghani, H. Zulzalil, A. A. Zaidan, and B. B. Zaidan, "Towards a unified criteria model for usability evaluation in the context of open source software based on a fuzzy Delphi method," *Information and Software Technology*, vol. 130, p. 106453, 2021/02/01/ 2021. https://doi.org/10.1016/j.infsof.2020.106453.

APPENDIX A. LIST OF PRIMARY STUDIES

ID	Research Article Reference
B1	Bures, M., Macik, M., Ahmed, B. S., Rechtberger, V., & Slavik, P. (2020). Testing the usability and accessibility of smart tv applications using an automated model-based approach. IEEE transactions on consumer electronics, 66(2), 134-143.
B2	Alahmadi, T., & Drew, S. (2017, April). Subjective evaluation of website accessibility and usability: A survey for people with sensory disabilities. In Proceedings of the 14th International Web for All Conference (pp. 1-4).
B3	D Al-Maani, D. I., & Bani-Salameh, H. (2017, March). A generic model for evaluating the usability of learning management systems. In Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing (pp. 27-1).
B4	Van Waardhuizen, M., McLean-Oliver, J., Perry, N., & Munko, J. (2019, May). Explorations on Single Usability Metrics. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-8).
B5	Komiyama, T., Fukuzumi, S. I., Azuma, M., Washizaki, H., & Tsuda, N. (2020, July). Usability of Software–Intensive Systems from Developers' Point of View. In International Conference on Human-Computer Interaction (pp. 450-463).
B6	Diaz, E., Flores, S., & Paz, F. (2019, July). Proposal of usability metrics to evaluate e-commerce websites. In International Conference on Human-Computer Interaction (pp. 85-95).
B7	Komarkova, J., Sedlak, P., Habrman, J., & Cermakova, I. (2017, July). Usability evaluation of web-based GIS by means of a model. In 2017 international conference on information and digital technologies (pp. 191-197).
B8	Alotaibi, K. J. (2017, May). Gathering of usability requirements by Saudi e-learning software developers. In 2017 8th International Conference on Information Technology (pp. 255-261).
B9	Okike, E. U., & Morogosi, M. (2017, July). Measuring the usability probability of learning management software using logistic regression model. In 2017 Computing Conference (pp. 1217-1223).
B10	Störrle, H. (2018, May). Improving model usability and utility by layered diagrams. In Proceedings of the 10th International Workshop on Modelling in Software Engineering (pp. 59-66).
B11	Kwon, H., & Choi, W. (2018, June). A Feld Study on Improving the API Usability of Software Platforms for Consumer Electronics Devices. In 2018 IEEE International Conference on Consumer Electronics-Asia (pp. 206-212).
B12	Geszten, D., Hámornik, B. P., & Hercegfi, K. (2018, August). Exploring awareness related usability problems of collaborative software with a team usability testing approach. In 2018 9th IEEE International Conference on Cognitive Infocommunications (pp. 45-50).
B13	McClellan, M. A., Karumur, R. P., Vogel, R. I., Petzel, S. V., Cragg, J., Chan, D., ... & Geller, M. A. (2016). Designing an educational website to improve quality of supportive oncology care for women with ovarian cancer: an expert usability review and analysis. International journal of human-computer interaction, 32(4), 297-307.
B14	Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability. International Journal of Human-Computer Interaction, 31(8), 496-505.

B15	Bringula, R. P. (2016). Factors affecting web portal information services usability: A canonical correlation analysis. International Journal of Human–Computer Interaction, 32(10), 814-826.
B16	Shamim, A., Balakrishnan, V., Tahir, M., & Ahsan Qureshi, M. (2016). Age and domain specific usability analysis of opinion visualisation techniques. Behaviour & Information Technology, 35(8), 680-689.
B17	Hanrath, S., & Kottman, M. (2015). Use and usability of a discovery tool in an academic library. Journal of web librarianship, 9(1), 1-21.
B18	Christophersen, T., & Konradt, U. (2011). Reliability, validity, and sensitivity of a single-item measure of online store usability. International Journal of Human-Computer Studies, 69(4), 269-280.
B19	Ko, S. M., Chang, W. S., & Ji, Y. G. (2013). Usability principles for augmented reality applications in a smartphone environment. International Journal of Human-Computer Interaction, 29(8), 501-515.
B20	Leow, M. C., Wang, L. Y. K., Lau, S. H., & Tan, C. K. (2016). Usability of rpg-based learning framework. International Journal of Human-Computer Interaction, 32(8), 643-653.
B21	Lee, Y., & Kozar, K. A. (2012). Understanding of website usability: Specifying and measuring constructs and their relationships. Decision support systems, 52(2), 450-463.
B22	Carvajal, L., Moreno, A. M., Sanchez-Segura, M. I., & Seffah, A. (2013). Usability through software design. IEEE Transactions on Software Engineering, 39(11), 1582-1596.
B23	Lavie, T., Oron-Gilad, T., & Meyer, J. (2011). Aesthetics and usability of in-vehicle navigation displays. International Journal of Human-Computer Studies, 69(1-2), 80-99.
B24	Hertzum, M., & Clemmensen, T. (2012). How do usability professionals construe usability?. International Journal of Human-Computer Studies, 70(1), 26-42.
B25	Roberts, M. J., Newton, E. J., Lagattolla, F. D., Hughes, S., & Hasler, M. C. (2013). Objective versus subjective measures of Paris Metro map usability: Investigating traditional octolinear versus all-curves schematics. International Journal of Human-Computer Studies, 71(3), 363-386.
B26	Lallemand, C. (2011, June). Toward a closer integration of usability in software development: a study of usability inputs in a model-driven engineering process. In Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems (pp. 299-302).
B27	Baum, D., Bechert, S., Eisenecker, U., Meichsner, I., & Müller, R. (2020, August). Identifying Usability Issues of Software Analytics Applications in Immersive Augmented Reality. In 2020 Working Conference on Software Visualization (pp. 100-104).
B28	Abiri, R., Borhani, S., Kilmarx, J., Esterwood, C., Jiang, Y., & Zhao, X. (2020). A usability study of low-cost wireless brain-computer interface for cursor control using online linear model. IEEE Transactions on Human-Machine Systems, 50(4), 287-297.
B29	Isa, W. A. R. W. M., Suhami, M. R., Safie, N. I., & Semsudin, S. S. (2011). Assessing the usability and accessibility of Malaysia e-government website. American Journal of Economics and Business Administration, 3(1), 40-46.
B30	Rivero, L., & Conte, T. (2013). Using an empirical study to evaluate the feasibility of a new usability inspection technique for paper based prototypes of web applications. Journal of Software Engineering Research and Development, 1(1), 1-25.
B31	Moumane, K., Idri, A., & Abran, A. (2016). Usability evaluation of mobile applications using ISO 9241 and ISO 25062 standards. SpringerPlus, 5(1), 1-15.
B32	Mirkovic, J., Kaufman, D. R., & Ruland, C. M. (2014). Supporting cancer patients in illness management: usability evaluation of a mobile app. JMIR mHealth and uHealth, 2(3), e3359.
B33	Az-zahra, H. M., Fauzi, N., & Kharisma, A. P. (2019, September). Evaluating E-marketplace Mobile Application Based on People at the Center of Mobile Application Development (PACMAD) Usability Model. In 2019 International Conference on Sustainable Information Engineering and Technology (pp. 72-77).
B34	Pathirana, W. P. N. H., & Wickramaarachchi, D. N. (2019, March). Software usability improvements for Generation Z oriented software application. In 2019 International research conference on smart computing and systems engineering (pp. 151-157).
B35	Hamzah, N., Mageswaran, G., Nagappan, S. D., & Chuprat, S. (2018, October). Assessing Usability of Ubiquitous Systems Using Quality Model. In 2018 Fourth International Conference on Advances in Computing, Communication & Automation (pp. 1-4).
B36	Tamimi, H., & Bensefia, A. (2018, November). Software Usability Challenges for Native Arab Users. In 2018 3rd International Conference on System Reliability and Safety (pp. 6-12).
B37	Abuqaddom, I., Alazzam, H., Hudaib, A., & Al-Zaghoul, F. (2019, June). A measurable website usability model: Case Study University of Jordan. In 2019 10th International Conference on Information and Communication Systems (pp. 83-87).
B38	Torrente, M. C. S., Prieto, A. B. M., Gutiérrez, D. A., & De Sagastegui, M. E. A. (2013). Sirius: A heuristic-based framework for measuring web usability adapted to the type of website. Journal of Systems and Software, 86(3), 649-663.
B39	Di Nuovo, A., Varrasi, S., Conti, D., Bamsforth, J., Lucas, A., Soranzo, A., & McNamara, J. (2019, March). Usability evaluation of a robotic system for cognitive testing. In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (pp. 588-589).
B40	Vilbergsdottir, S. G., Hvannberg, E. T., & Law, E. L. C. (2014). Assessing the reliability, validity and acceptance of a classification scheme of usability problems (CUP). Journal of Systems and Software, 87, 18-37.
B41	Wang, W., Cheng, J., & Guo, J. L. (2020). How Do Open Source Software Contributors Perceive and Address Usability? Valued Factors, Practices, and Challenges. IEEE Software.
B42	Polgár, P. B. (2015). Using the cognitive walkthrough method in software process improvement. E-Informatica Software Engineering Journal, 9(1).

B43	Geszten, D., Hámornik, B. P., & Hercegfi, K. (2019, October). Usability evaluation of a collaborative design software in the wild. In 2019 10th IEEE International Conference on Cognitive Infocommunications (pp. 101-106).
B44	Teruel, M. A., Navarro, E., López-Jaquero, V., Montero, F., & González, P. (2014). A CSCW requirements engineering CASE tool: development and usability evaluation. Information and Software Technology, 56(8), 922-949.
B45	Lindgaard, G. (2015). Challenges to assessing usability in the wild: a case study. International Journal of Human-Computer Interaction, 31(9), 618-631.
B46	Kortum, P., & Sorber, M. (2015). Measuring the usability of mobile applications for phones and tablets. International Journal of Human-Computer Interaction, 31(8), 518-529.
B47	Mansar, S. L., Jariwala, S., Shahzad, M., Anggraini, A., Behih, N., & AlZeyara, A. (2012). A usability testing experiment for a localized weight loss mobile application. Procedia Technology, 5, 839-848.
B48	Hurtado, N., Ruiz, M., Orta, E., & Torres, J. (2015). Using simulation to aid decision making in managing the usability evaluation process. Information and Software Technology, 57, 509-526.
B49	Hasan, L., Morris, A., & Probets, S. (2012). A comparison of usability evaluation methods for evaluating e-commerce websites. Behaviour & Information Technology, 31(7), 707-737.
B50	Hua, L., & Gong, Y. (2013, July). Usability evaluation of a voluntary patient safety reporting system: Understanding the difference between predicted and observed time values by retrospective think-aloud protocols. In International Conference on Human-Computer Interaction (pp. 94-100). Springer, Berlin, Heidelberg.
B51	Bruun, A., & Stage, J. (2012, November). Training software development practitioners in usability testing: an assessment acceptance and prioritization. In Proceedings of the 24th Australian Computer-Human Interaction Conference (pp. 52-60).
B52	Aryana, B., & Clemmensen, T. (2013). Mobile usability: experiences from Iran and Turkey. International Journal of Human-Computer Interaction, 29(4), 220-242.
B53	Moorthy, J. T. S., bin Ibrahim, S., & Mahrin, M. N. R. (2013, December). Formulation of usability risk assessment model. In 2013 IEEE Conference on Open Systems (pp. 168-173).
B54	Dubey, S. K., & Rana, A. (2011). Usability estimation of software system by using object-oriented metrics. ACM SIGSOFT Software Engineering Notes, 36(2), 1-6.
B55	Erickson, W., Trerise, S., Lee, C., VanLooy, S., Knowlton, S., & Bruyère, S. (2013). The accessibility and usability of college websites: Is your website presenting barriers to potential students?. Community College Journal of Research and Practice, 37(11), 864-876.
B56	Rodríguez, F. D., Acuña, S. T., & Juristo, N. (2015). Design and programming patterns for implementing usability functionalities in web applications. Journal of Systems and Software, 105, 107-124.
B57	Mapayi, T., Olaniyan, O., Isamotu, N., & Moses, O. (2013). Evaluating usability factors in different authentication methods using artificial neural network. African Journal of Computing & ICT, 6(1), 69-78.
B58	Wale-Kolade, A. Y. (2015). Integrating usability work into a large inter-organisational agile development project: Tactics developed by usability designers. Journal of systems and software, 100, 54-66.
B59	Brown, M., Sharples, S., & Harding, J. (2013). Introducing PEGI: A usability process for the practical evaluation of Geographic Information. International journal of human-computer studies, 71(6), 668-678.
B60	Panach, J. I., Juristo, N., Valverde, F., & Pastor, O. (2015). A framework to identify primitives that represent usability within Model-Driven Development methods. Information and Software Technology, 58, 338-354.
B61	Castilla, D., Garcia-Palacios, A., Breton-Lopez, J., Miralles, I., Baños, R. M., Etchemendy, E., ... & Botella, C. (2013). Process of design and usability evaluation of a telepsychology web and virtual reality system for the elderly: Butler. International Journal of Human-Computer Studies, 71(3), 350-362.
B62	Winter, J., & Hinley, M. (2011). Examining correlations in usability data to effectivize usability testing. E-Informatica Software Engineering Journal, 5(1).

APPENDIX B. LIST OF JOURNALS AND PROCEEDINGS

IEEE transactions on consumer electronics	Journal
Proceedings of the 14th International Web for All Conference	Conference
Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing	Conference
Conference on Human Factors in Computing Systems	Conference
International Conference on Human-Computer Interaction	Conference
International conference on information and digital technologies	Conference
International Conference on Information Technology	Conference
Computing Conference	Conference
Proceedings of the 10th International Workshop on Modelling in Software Engineering	Workshop
IEEE International Conference on Consumer Electronics-Asia	Conference
9th IEEE International Conference on Cognitive Infocommunications	Conference

International journal of human-computer interaction	Journal
Behaviour & Information Technology	Journal
Journal of web librarianship	Journal
International Journal of Human-Computer Studies	Journal
Decision support systems	Journal
IEEE Transactions on Software Engineering	Journal
Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems	Symposium
2020 Working Conference on Software Visualization	Conference
IEEE Transactions on Human-Machine Systems	Journal
American Journal of Economics and Business Administration	Journal
Journal of Software Engineering Research and Development	Journal
SpringerPlus	Journal
JMIR mHealth and uHealth	Journal
International Conference on Sustainable Information Engineering and Technology	Conference
International research conference on smart computing and systems engineering	Conference
Fourth International Conference on Advances in Computing, Communication & Automation	Conference
3rd International Conference on System Reliability and Safety	Conference
10th International Conference on Information and Communication Systems	Conference
Journal of Systems and Software	Journal
14th ACM/IEEE International Conference on Human-Robot Interaction	Conference
IEEE Software	Journal
E-Informatica Software Engineering Journal	Journal
10th IEEE International Conference on Cognitive Infocommunications	Conference
Information and Software Technology	Journal
Procedia Technology	Journal
24th Australian Computer-Human Interaction Conference	Conference
2013 IEEE Conference on Open Systems	Conference
ACM SIGSOFT Software Engineering Notes	Journal
Community College Journal of Research and Practice	Journal
African Journal of Computing & ICT	Journal

Multimedia Transmission Mechanism for Streaming Over Wireless Communication Channel

Shwetha M[1]
Assistant Professor
Department of Electronics and Communication Engineering
Dr. Ambedkar Institute of Technology
Bangalore, India

Yamuna Devi C R[2]
Associate Professor
Department of Electronics and Telecommunication Engineering
Dr. Ambedkar Institute of Technology
Bangalore, India

***Abstract*—With the evolution of wireless communication technologies (i.e., 4G/5G), the explosion of multimedia transmission of content sharing has become an integral part of users' daily lives. It expects further growth in Quality of Service (QoS) and Quality-of-Experience (QoE) performance. Therefore, multimedia service providers are developing new technologies to offer higher video streaming quality content along with video compression standards, which is highly demanded by the receivers. Thus, inventing precise and efficient quality-based media transmission protocol will significantly help to improve the multimedia QoS over wireless networks. This comprehensive research study discusses standard research work progress in multimedia transmission protocol for wireless communication networks. It also investigates the limitations of such literature found some challenging factors that play a significant role in managing the superior signal quality for digital or video content transmission over heavy traffic conditions. The final section provides a briefing on crucial open research issues to develop a multimedia transmission model that can seamlessly communicate multimedia content irrespective of adverse traffic conditions.**

***Keywords*—Multimedia transmission; video encoding; multimedia streaming; quality of service; quality of experience; video compression standards**

I. Introduction

The evolution of media communication systems (i.e., 4G and 5G) has led to increased digital services and applications, such as IPTV, social networking, video conferencing, multimedia games, educational digital presentation, etc. These multimedia applications are becoming an integral part of our daily lives and are provisioned to grow exponentially. Various multimedia service providers, e.g., subway system [1], chroma-keying [2], 2D and 3D animation industries, etc., are discovering differing technologies to offer a higher quality of experience, which end users are progressively insisting. In the current scenario, almost all users having/utilizing smart devices for multi-purpose like generating data, communicating, or sharing information from person to person or device to device anywhere and anytime. Most of the users spending time viewing and sharing multimedia content from the internet. At present, viewing video content over the internet is almost free for public users. It is also studied that there will be a phenomenon increase in user base for using multimedia streamed contents over the internet [3]. Apart from this, it is also figured out that it reaches up to 80% by 2020. In the future, most of the multimedia access traffic will be transmitted wirelessly. Nowadays, video transmission has become common for all internet users.

Despite the increasing growth of advanced technologies, the audio-video transmission process suffers from impairments by lossy transmission and source encoding over the network channels, thereby degrading the quality of multimedia content [4]. For example, the user may receive a sample video file that may group different quality ranges due to rendering errors or another transmission. However, other technologies and network standards have been developed that facilitate a high communication range among digital devices. Such standard networks are; IEEE 802.15 WPAN [5], IEEE 802.11 WLAN [6], IEEE 802.16 WMAN [7], and 4G telecommunication networks [8]. The high-speed network availability and video transmission with high speed and minimum cost provide a new era for video communication that has not been implemented over the past decades. Video communication technology dominates the high traffic over the wireless networks and is envisioned for multiple applications. The deployment of high-standard networks like 3G and 4G and advancements in intelligent device development had led to the massive demand for digital media transmission over wireless channels. The increasing requirement for multimedia content creates challenges for all digital media streaming systems, such as wireless network service providers [9], content providers, and mobile device makers.

The mobile network service providers and content providers strive to enhance their services while adopting advanced technologies—for example, improving processing power and high-quality displays. Ultimately, the common goal of all service providers is to improve the quality of experience (i.e., QoE) for end-users. The objective of QoE is t evaluate the video streaming quality by end-users. It can be assessed in-display smoothness, streaming bit-rate, video quality range like PSNR, etc. Therefore, in [10], the authors provided a comprehensive survey study on existing video transmission methods and offered a research direction towards defining high QoE and new transmission methods for 3D video streaming. Video streaming or video transmission over wireless channels remains challenging, for example, signal interference between nodes, unreliable quality due to multi-

path fading, and dynamicity in connectivity. Routing problems always influence the end-to-end QoS of video applications; an example is finding an optimal route that could help for video transmission with high quality. The conventional routing protocols rely on packet delay and packet loss metrics to achieve the high perceived video quality. The case study shows that multiple techniques exist to address the routing issues for real-time video transmission over wireless networks to enhance the quality of video at the end user. The practical use of these theoretical results of existing studies offers better guidelines toward formulating new studies for better QoS and QoE performance in routing strategies.

Therefore, most research was carried out on developing optimal routing for video transmission in wireless networks, mainly focused on network-oriented QoS (i.e., delay, throughput, and packet loss) and less concentrated on application-oriented QoS perceived video quality. Thus, the present survey study overviews different challenges in designing a multimedia transmission protocol to improve the signal quality in heavy traffic over wireless networks. From the prior research study, it can be observed that very little research has been done towards multimedia compression, which does not have an explicit module of wireless networks and its associated problems. The transmission protocols designed to date seriously lack multi-level optimization, showing that the existing algorithms can provide a one-way solution and cannot go beyond that. Since multimedia transmission protocols play a significant role in delivering the requirements of WSN applications, there is less work done that needs to extend the conventional real-time communication systems. Multimedia transmission protocols must be upgraded to be suitable for wireless environments. The significant challenge in implementing multimedia transmission protocol involves real-time data streaming over wireless networks with high QoS. Another challenging task is the compression technique, where a slight increment in the compression level causes data quality to decrease. Hence, the present study's contribution presents the relevant prior research study towards multimedia transmission protocol over wireless channels that have been recently introduced. Also, investigate such literature's limitations and find some challenging factors that play a significant role in managing the superior signal quality for digital or video content transmission over heavy traffic conditions. With the increasing usage of streaming-based services over various commercial applications for different causes, a smoother streaming experience is needed. This streaming is eventually carried out using different variants of the wireless network. Adoption of wireless network offers cost-effective utilization towards the user, but it also introduces various challenges. There are multiple archives of research-based solutions for dealing with the difficulties in data transmission over a wireless network. However, the challenges still exist, and there is yet evolving research work.

Hence, the biggest problem is a snapshot of the existing transmission methods for multimedia contents over a wireless network. Therefore, this manuscript contributes towards more detailed insight into the strength and weaknesses of existing multimedia transmission in a wireless network. The overall organization of the current manuscript is as follows; Section II briefs about multimedia transmission standards. Section III discusses different researchers who have introduced various theoretical and implementation research studies. Section IV presents the other multimedia transmission protocols that help improve QoS in the streaming process concerning compression standards. Section V reviews current research challenges found from existing studies. Finally, in the last Section VI, the summary of the study is presented in the form of a conclusion.

II. Multimedia Transmission Protocols

The multimedia transmission means forwarding data packets that usually consist of audio, video, or audio-visual streaming. Multimedia transmission is the fundamental process for sending media content to mobile users. The scalable transmission process will need efficient and robust routing protocols which provide high-quality output video content. Therefore, effective and scalable video streaming protocols intend to transmit multimedia content through the internet while enabling users to access it without completing the transmission process. Generally, all video transmission protocols preferred a transport layer where transmission occurs via live video streaming. The functionality of the transmission protocol is to offer real-time, sequential, less packet loss, low delay, minimum energy consumption for video/digital data streaming. This section discusses the most common and frequently adopted multimedia transmission protocol.

- Real-Time Messaging Protocol (RTMP): RTMP is best and significantly utilized for media streaming technology in all listed transmission protocols. Macromedia developed RTMP to stream digital data over the internet. It is a TCP-based protocol that provides low latency communication with a persistent connection. It contains multiple features such as; it is very flexible and enables audio, video, audio-visual streaming, even text content in several formats to various devices. Another significant feature is the multiplatform transmission protocol; users can access the media content using any platform (i.e., Android, Mac, Windows, etc.).
- Nevertheless, one of the drawbacks of this protocol is that users need to consider it before selecting it for video streaming activities. RTMP is an old protocol and well-proof multimedia streaming technology that has been adopted for years now. The Flash-Player helps in viewing media streams via RTMP, which is very famous and utilized over the globe [11].
- Real-time Transport / Control Protocol (RTP/RTCP): A transport layer protocol built on UDP enables real-time multimedia content transport. It may be exploited for single-way transport services like video on demand and internet telecalls. One of the features of RTP is closely associated with RTCP, which performs at the session layer of the ISO model. It offers feedback for the quality of content distribution. RTP is mainly designed

to utilized UDP/IP protocols at the transport layer. Both RTP & RTCP are compressed in UDP/IP packets.

- Real-time Streaming Protocol (RTSP): Primarily, RTSP is utilized to control real-time media streaming applications such as HD-video streaming. It is a network control protocol that establishes communication sessions among the endpoints. This protocol uses TCP protocol to balance the end-to-end session, and RTP is utilized to deliver the media content over the UDP. Additionally, RTSP protocol may interact with HTTP server such that hand over devices is specified among the media and web server. This makes, as, the delivered file content to be requested via HTTP or RTSP. "VOCAL" optimized software is utilized for media transmission, which supports RTSP protocol [12].
- HTTP Streaming: In the current multimedia streaming technique, HTTP streaming is the new trend that supports adaptive bit rates. Specifically, HTTP plus TCP/IP protocol is designed for reliable transmission to maintain the transmission flow. APPLE's company developed HTTP streaming for IOS, and it does not apply to other products, i.e., it supports only Apple's products [13].
- Adobe HTTP Dynamic Streaming (AHDS): AHDS is a proprietary solution for real-time on-demand streaming of high-quality content. It mainly operates on HTTP. Like RTMP protocol, AHDS is associated with flash. However, unlike RTP, AHDS utilizes an adaptive bit-rate mechanism to transfer the MP4 media files over HTTP servers. Additionally, it supports encrypted files and can be used for HD quality video content delivery up to 1080 pixels with a 6MBps bit rate. It also supports different compression algorithms, for example, H.264, VP6, AAC, and MP3 video [14].
- Microsoft Smooth Streaming (MSS): Microsoft invented it in 2008 to develop silver light architecture [15]. It was primarily utilized to deliver the on-demand video clips of the 2008 Olympics. The MSS technology can optimize the media playback by switching real-time video quality. Now, this technology goes beyond, and it can be utilized to reach different types of devices/clients like as browsing with Xbox, Silver light, Apple devices, TV set-top boxes, etc. typical example is the Xfinity TV application for iPhone, which has been operated on top of MSS and transmits media content to IOS devices. MSS is generally utilized for television and premium media content delivery. MSS can also host on Apache web server and support advanced operations like rewind, fast-forward, etc.
- Shout-Cast: It is one of the popular technologies which delivers broadcast streaming. Shoutcast utilizes its protocols, and it was developed by Nullsoft and named ICY; at present, Ultravox is using for ShoutCast-2. This protocol can operate over either UDP or TCP. The major drawback of this protocol is only applicable for broadcasting, not for on-demand video delivery.
- Moving Picture Expert Group-DASH: (MPEG-DASH): Generally, MPEG was developed for multimedia streaming with multiple standards, i.e., MPEG-2, MPEG-4, and MPEG-7. The dynamic adaptive streaming standard over HTTP is the MPEG-DASH, which can solve media delivery problems to several devices with a unified standard [16].

Above discussed, all multimedia protocols have different methods and formats as well as unique features. Therefore, a dynamic supportable protocol must be needed to deliver or receive media content from servers to users. However, all these standard protocols can provide multimedia content over wireless channels. Apart from this, there are some recent studies carried out in this direction. The study presented by Li et al. [17] has implemented a parallel coding scheme to improve multimedia transmission considering turbo coding. Reinforcement learning is another scheme that offers better video transmission performance considering a case study of the internet of things, as seen in the analysis of Xiao et al. [18]. Ta et al. [19] have developed a cooperation scheme of image transmission considering sensor networks as a wireless medium. Huang et al. [20] have presented a Q-Learning scheme over the cognitive network for improving multimedia transmission. Liu et al. [21] have used the beamforming approach for enhancing multimedia transmission in 5G, harnessing the potential of the relay network. The following section discusses existing strategies for multimedia transmission over different variants of the wireless network.

III. Existing Approaches

Different theoretical and implementation research studies have been introduced by other researchers that achieved multimedia transmission over other network technologies. Therefore, this section discusses relevant research studies on multimedia transmission protocols in wireless networks. At present, various wireless protocols are claimed to offer the quality of service for multimedia transmission. Table I highlights this comparative analysis of these standards.

TABLE I. Comparison of Performance of some Existing Standards

	H1	H2	H3	H4	H5
[22]	Delay	No	Yes	No	No
[23]	Delay	Limited	No	Limited	No
[24]	Delay	No	Yes	Yes	Yes
[25]	Delay	Yes	Yes	Yes	Yes
[26]	Delay	No	Yes	Yes	Yes

H_1: QoS Metric, H_2: Energy-Aware, H_3: Location-aware, H_4: Scalability, and H_5: Service differentiation

At present, most online users are widely utilizing multimedia content for various reasons to ensure public and personal services. The multimedia content is transferred or distributed over different networks, i.e., ranging from classical wireless networks to IoT (Internet of Things). However, the deployment of the network has multiple reasons: ensuring high QoS and QoE with minimum latency reduction specific to traffic concerns. To tackle this challenging task, Bennis et al. [27] introduced a cross-layer protocol that handles the video transmission procession over WSN's. The proposed approach also cooperated with the application layer to frame an aware strategy for queuing policy that solves the various functions (i.e., enqueuing, dequeuing) to optimum latency reduction and enhances the video streaming quality. Asha and Mahadevan [28] adopted the exact cross-layer mechanism, which addressed QoS challenges in mobile networks for multimedia applications. To enhance the QoS for the mobile web, the authors proposed a combined approach that improves the network lifetime. This represents three objectives; network modeling, threshold-based packet transmission, and queuing model on physical-layer, which support the QoS.

Current cloud computing environment, multimedia transmission via IoT technology presents lots of challenges to nodes' diversity. Said et al. [29] explored adaptive real-time transport and control protocols over IoT environments. The experimental study considered the heterogeneous network for transmission, threshold value, and various multimedia sources. The primary intention was to split the scalable multimedia sessions into multiple sessions with network status awareness. The proposed technique can decrease network overload under critical traffic conditions. Also advantageous for an end-to-end delay, minimum packet loss, and energy consumption. Another research study by Huang et al. [30] focused on the concept of improving QoE over multimedia IoTs for network users. First, the author introduced a Quality-of-Experience (QoE) optimization mechanism for multimedia IoTs that leverages the data fusion technique. Initially, the proposed method involves two core phases; the data fusion model builds a QoE mapping among the un-controllable streaming data with controllable network system data. Then, another automatic QoE optimization model was designed to automatically adjust the network systems and achieve higher optimization results.

Multiple approaches have been proposed which ensures the energy trade-off for network performance [31][32]. However, the routing challenge has been considered a significant problem and needs to resolve to support future communication technologies. Therefore, in the context of work carried out by Khernane et al. [31], who have addressed the different routing problems based on a routing matrix. As a solution, a single selective routing protocol has been introduced. The solution strategy allows the end-to-end routing for each video sensor without any path discovery. Therefore, it is named as a change of dynamic network topology.

Another challengeable issue in the transmission process is the security because voluminous data content is quite impossible by conventional methods to encrypt the video content fully. Almasalha et al. [32] presented a scalable model for securing multimedia content on low energized mobile devices. The proposed technique is mainly applicable to the compressed video stream and will not require any decoding. The system encrypts 3% of packet load and offers equivalent security by doing bit-stream encryption. This phenomenon has experimented on laptops, desktops, notebooks, and mobile phones. Canovas et al. [33] have proposed a multimedia distribution system that delivers the video streams over the IP network. The proposed mechanism adopted a heuristic decision method and a probabilistic distribution system that provides the media streams among the service providers. Clients can upload and download the media files. The proposed approach takes into account energy conservation as well as enhances the QoE of end-users. However, the wireless multimedia transmission process contains multiple constraints and faces several problems: bit rate, storage problem, power consumption, bandwidth, and processing rate. Hassan et al. [34] explored an advanced multimedia compression technique that lacks such challenges, i.e., H.264 has been developed jointly with MPEG. The proposed compression standard offers multiple tunable parameters which tailor the video encoding operation as per the provisions. Additionally, the proposed framework to resolve the multi-objective problems achieved relevant results in bit-rate improvement, power consumption, and quality enhancement in multimedia content.

Due to wireless sensor technology's advanced improvement, sensor nodes can perform multimedia data processing, but the significant challenge is the real-time routing system over wireless multimedia networks [35]. Therefore, Ahmed et al. [35] introduced a real-time routing protocol for video streaming over next-generation wireless multimedia sensor networks. This study elaborates an algorithm to accomplish adaptive traffic shaping for video streaming. It exploits a multi-route forwarding approach with dynamic cost computation for a section of the next node. The author mainly focused on video streaming and real-time routing. Majeed et al. [36] have provided a comprehensive survey study on several problems in the art of information-centric networking systems and discussed respective architectures and literature concerning multimedia streaming.

Additionally, a roadmap is provided on the research community studying in a similar domain. Huang et al. [30] introduced analytical modeling for multimedia data flow scheduling systems over SDNs (software-defined networks). This study author presented a hybrid data flow scheduling system by integrating priority-based queuing packets and offering QoS for multimedia applications in SDN. Several researchers provided different multimedia transmission methods. Some of them addressed multimedia transmission over VANETs. For example, Xu et al. [37] have explored an information-centric networking model that delivers multimedia content over mobile vehicular networks. The proposed mechanism implements two significant factors, data mobility, and provider supply-demand balance. They also formulated an optimized mixed-integer programming module that is cost-effective concerning QoS multimedia. In another research study, author Moussaoui et al. [38] have adopted VANETs technology to implement cost-free and efficient multimedia content sharing among the two vehicles and their

passengers. In this study, the authors proposed an improved cross-layer protocol that deals with routing challenges over VANETs. Yang et al. [39] have presented a movie recommendation model based on user scores. From the viewpoint of the movie formulation system, the level of access control & media security are analyzed, along with cloud storage security architecture was implemented. The primary objective is to ensure the safety of multimedia content during the data transmission process. Dien et al. [40] have presented cross-layer architecture to implement a security-based routing protocol for multimedia transmission on the wireless sensor network. The primary focus was on energy consumption during packet transmission and path scheduling. The authors concluded that the proposed framework is suitable for enhancing real-time video quality and prolonging the network performance from the implementation results.

Rapid advancement in wireless technology infrastructure and smart devices, video streaming like cloud gaming, live sports watching, YouTube video uploading and downloading, etc., has dominated the harmful applications over the web. With the increasing rate in emerging multimedia applications, providing a better quality of video services (e.g., YouTube and many more) provides multimedia streaming up to sixty frames in a second. Therefore, Wu et al. [41] mainly focused on real-time video transmission problems on mobile devices like the example video conferencing or calls, video games, etc. hence, nowadays, it is becoming a highly challenging issue for the service providers to provide and ensures about high-quality content delivery as well as high-quality video streaming. For that purpose, the authors proposed a frame scheduling and error resilience model for mobile devices over heterogeneous wireless networks. Hameed et al. [42] have introduced an energy-efficient video quality prediction model for wireless communications. The entire work mainly consists of two components: real-time video quality with low complexity. Another is the content and energy-aware model to balance the video quality during packet transmission over the network. The authors showed that the proposed prediction model achieves ~ 90% accuracy, and as compared to conventional techniques, the proposed communication model reduces the network overhead by 41%. With the growth of new generation networks and communication technologies, video services are becoming pervasive for large-scale heterogeneous wireless networks. More and more uploading, downloading, and accessing video information with the help of various devices (PCs, tablets, smart TVs, smartphones, etc.) is becoming very common for all users. Offering heterogeneity with QoE, which supports a wide range of multiple multimedia devices, is crucial and challenging to broadcast the video over new generation wireless networks. Chen et al. [43] have reviewed different existing video broadcasting technologies and founds present requirements ranging from homogeneous to heterogeneous transmission network technologies. Also presented is a typical modeling approach for video broadcasting with large-scale heterogeneous network support that enables QoE, joint coding, cross-layer transmission, optimal and dynamic adaptation to enhances the receiving quality of heterogeneous devices.

One of the challenging factors for digital media transmission over multimedia WSN is the spectrum scarcity with high radio interference in the current digital world. In that context, Bradai et al. [44] proposed a solution mechanism for multimedia transmission over multimedia WSNs which exploit radio interferences for spectrum scarcity and clustering method for energy efficiency. Also highlighted significant issues and challenges of multimedia WSNs, i.e., high bandwidth requirements, energy efficiency, QoS, data processing cross-layer routing issues, and compressing techniques. On the other hand, Gur [45] has mainly focused on QoE and QoS requirements over multimedia applications or services deployed over mobile networks. Also defined the network performance parameters utilized to balance the service performance like; throughput, latency, packet loss, reliability, and availability. Han et al. [46] have investigated the new concept of a fast directional hand-off mechanism that helps to enhance or improves the quality of multimedia over WSN. Also introduced is a lightweight retransmission protocol that reduces the packet loss on WiFi without generating any acknowledgment. The proposed mechanisms can be applied on android based smart devices, and their performance has been evaluated in the indoor wireless LAN environment. The experiment results demonstrated that the proposed mechanism balances the seamless quality for video streaming under hand-off operation. Some studies have analyzed the challenges facing multimedia transmission in the IoT environment. Alvie et al. [47] have introduced a novel of the internet of multimedia things where smart multimedia devices can cooperate and interact with each other and connect with the internet to provide multimedia-based services to global users. Jiang and Meng [48] have designed an IoT-based multimedia platform that improves real-time multimedia transmission protocol quality. The following Table II highlights a summary of the existing multimedia transmission techniques that different authors have proposed.

TABLE II. SUMMARY OF THE PRIOR RESEARCH STUDY TOWARDS MULTIMEDIA TRANSMISSION TECHNIQUES

Author	Problem	Technique	Application	Performance
Bennis,[27]	To ensure better QoE with Low latency for multimedia transmission	Cross-layer scheme	Enhance the video quality and reliability	Bit-rate, delay, and packet loss rate.
Asha et al.[28]	Improve the QoS for mobile applications	Channel modeling	Able to select the transmission channel through the threshold. Less resource utilization	Throughput, bandwidth, delay
Said et al. [29]	Multimedia Transmission over IoT environment	RTP/RTCP protocols	End to end delay, minimum energy consumption	Delay jitter, packet loss, throughput.
Huang et al. [30]	Increase QoE in multimedia IoT	Machine learning	Optimize the QoE by adjusting network parameters, automatic bandwidth allocation	Comparative analysis with different topologies
Khernane et al. [31]	To improve network performance	Video encoding method	It consumes less amount energy and increases the network lifetime	Energy cost, network lifetime.
Almasalha et al. [32]	Secure media streaming technique for low energized mobile nodes	Selective encryption method, RTP protocol	It is suitable for laptop, desktop, tablet, and Nokia N-series platform	Speed and streaming rate
Canovas et al.[33]	To ensure high quality of experience for IP multimedia with minimum resource allocation	Heuristic and probability distribution model	Can upload multimedia files with high speed	Delay jitter, packet loss, and energy consumption
Hassan et al.[34]	To achieve a high bit rate for H.264 encoding	joint parameter cost-function	Maintain good video quality with low energy consumption	Bit-rate, PSNR, processing delay
Ahmad et al. [35]	Real-time data routing for media streaming	Dynamic routing mechanism for wireless multimedia applications	The packet can switch from single path to multi-path, balance the traffic load	Limited bandwidth, minimum processing power
Majeeb et al. [36]	Multicast delivery, security, QoS, and mobility	Information-centric networking mechanism	Multimedia streaming	Comparative analysis with existing methods.
Xu et al. [37]	Provide high-quality media streaming with a minimum cost of mobile vehicular networks	Information-centric networking mechanism	Reduce the economic cost, caching enhancement, and less resource utilization	Delay, jitter, QoE, playback continuity.
Moussaoui et al. [38]	Multimedia Data Dissemination over VANET's	Cross-Layer technique	Maintain the road traffic, mitigate traffic congestion and reduce air pollution	Decrease the packet control overhead.
Yang et al. [39]	To ensure multimedia security and access control	Hybrid cloud storage security model	Mitigates the attacks	Downloading time Vs. user request
Din et al. [40]	To improve the QoS for multimedia applications	Cross-layer approach, packet, and route scheduling algorithm	Can improve the network lifetime through packet scheduling	Loss ratio Vs. Lost frames, Energy usage per nodes
Wu et al. [41]	To achieve high-quality video frames via mobile devices	Frame scheduling approach	Can reduce the intraframe probability, outperformance w.r.t video transmission	End to end delay Vs. overdue frames
Hameed et al. [42]	Maintain the video quality with good QoE	Decision tree-based QoE support model	Predict the high quality of real-time video content with minimum complexity	Prediction accuracy about ¬ 90%, Network overhead by an average of 41%
Chen et al. [43]	Large video streaming with QoE over large-scale heterogeneous networks	Video broadcasting method, experimental analysis	Can improve the QoE of heterogeneous devices	PSNR, PSPNR, and SSIM
Bradri et al. [44]	Multimedia Transmission in an urban area using wireless multimedia sensor network	Cognitive radio spectrum technique, the clustering approach	Perform high video quality, minor transmission delay, less frame loss ratio	PSNR Vs. several channels and several media sources.
Gur et al. [45]	Multimedia transmission over WSN	Theoretical analysis	real-time communications	Comparative study between existing methods
Han et al. [46]	To provide high-quality services with minimum packet loss.	Hand-off mechanism, the retransmission protocol	Minimizes the packet loss ratio using WiFi, implemented over Android platform applications	The rate of packet loss during the hand-off process
Alvie et al. [47]	Multimedia streaming over IoT environment	Theoretical approach	Heterogeneous intelligent devices can interact and cooperate	NA
Jiang et al. [48]	Real-time multimedia streaming over IoT	Efficient communication protocol (i.e., UDP & TCP) is called control over UDP.	Rate control and retransmission	The result was calculated in terms of PSNR.

IV. Multimedia Compression Standards

Multimedia compression is a technique that transmits the media content over a wired or wireless channel by encoding digital video content. The compression for media content transmission offers multiple benefits like fewer storage requirements and minimum bandwidth requirements. The compression technique typically involves deletion of information not considered critical to viewing the video content and a good video codec technique that provides multiple benefits mentioned above: without significant degradation in the visual content experience, post-compression, and without requiring significant hardware overhead achieve the compression. Even within a particular video compression technique, different levels of compression standards can be applied. Hence, the more aggressive compression, high storage space and transmission bandwidth efficiency, and the higher computing power required. However, ISO/IEC and ITU-T are the influential international organizations that classified the multimedia compression standards into two major categories; i.e., ISO/IEC includes MPEG standards like MPEG-1, 2, 3, 4, MPEG-4(AVC), and Motion-JPEG [49]. At the same time, ITU-T has H.26x series standards, viz. H.261, H.263, H.264, and H.265 (HEVC) [50].

Most video retailers utilize few standard compression techniques (i.e., M-JPEG, MPEG-4, and H.264). However, such standards techniques are mainly relevant to video compression since video can be used for several purposes, such as video surveillance. Therefore, this section has briefly discussed work in state of the art in standard video compression techniques. Additionally, Fig. 1 illustrates the media transmission process at a different layer of the IOS model. Motion-JPEG: M-JPEG is a digital video sequence that contains a series of JPEG images. An image in the video file has equal quality determined from the compression level selected for the video encoder. The high compression level lowers the video file size as well as quality. Some image files require more bandwidth and memory during the compression process since the file size is more significant. Thus, to prevent more storage and bandwidth requirements, video retailers allow the users to set up the file size range for the image frame. The higher quality video content requires additional bandwidth and more storage for file transmission. The mezzanine image compression technique reduces the file transmission capacity and provides higher resolution with high-quality video content [51].

Additionally, the authors introduced a term called JPEG-XS which addressed the requirement for interoperable video over IP. In another research study, Willeme et al. [52] adopted a similar approach to JPEG-XS for image buffer compression. The proposed research aims to reduce the frame buffers' bandwidth and make HEVC more reliable for energy-aware applications.

Moving Pictures Experts Group (i.e., MPEG) is the traditional video compression technique that can compress the media contents like images and audio and combine both files. Several MPEG compression standards are currently available, i.e., MPEG-1, 2, 3, and MPEG-4. All MPEG versions have their features concerning the data rates variation. For example MPEG-1 intended for intermediate data-rates (i.e. 1.5 Mbit/sec), whereas MPEG-4 intended for very less data rates (i.e. <64 KB/sec). The study of Hameed et al. [53] has introduced a decision-tree-based media quality prediction model using the MPEG-4 compression standard. This technique extracts frames from the compressed bit-streams and predicts the video quality based on resultant features. The proposed model provides high video quality content with low complexity.

Meanwhile, Seethram et al. [54] investigated a scheduling algorithm to deliver a multimedia stream from the server to mobile users. Also introduced is an epoch-by-epoch model which allocates the transmission slots for video streaming. The experimental results were validated by applying the MPEG-4 compression technique and trace the wireless channel.

3G mobile video services were widely based on MPEG-4 and H.263 compression standards. Still, from the recent advancement in wireless transmission technology, almost all video service providers exclusively utilize the H.264 media transmission technique, providing a better video streaming quality. Thus, nowadays, all mobile operators are widely exploiting efficient video streaming applications for multimedia broadcasting. Several researchers have introduced multiple approaches that intended the different features with different H.264/AVC compression standards. For example, Hassan et al. [55] have introduced advanced video compression techniques (i.e., improved H.264) to overcome wireless networks' bit rate overhead challenge. Authors developed video compression standards jointly, i.e., H.264/MPEG-4, which reduces the cost function and maintains good video quality without performing compression. Many researchers developed the H.264/AVC compression technique [56]-[59] to reduce and delete the redundant media content such that compressed video files can be successfully transmitted over the wireless network. However, the significant challenge of any technique is to reduce the content size and provide high visual quality without any packet loss. Therefore, Chang et al. [56] presented a multi-pooling control access scheme that ensures low latency during the transmission of video frames and reduces transmission overhead. At the same time, Wu et al. [57] presented a video frame scheduling mechanism that reduces the total distortion. The result performance can be evaluated in terms of analyzing video PSNR values.

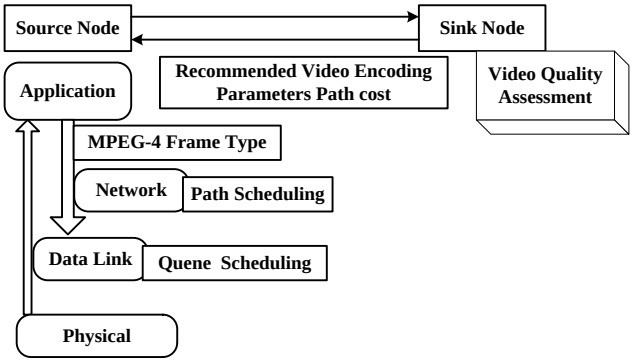


Fig. 1. Multimedia Transmission Process at different Layers.

V. RESEARCH GAP

Despite continuous progress and usage of multimedia content quality assessment, some challenges are addressed to some extent. Therefore, this section has discussed some significant research challenges in multimedia transmission over wireless networks.

- Media specific QoE demands: Generally, a multimedia file contains different types of media with other characteristics. For example, real-time video streaming, i.e., video, audio, or audio-visual, is delay-sensitive and error-resilient. On the other hand, non-real-time media files, i.e., web data, have minor delays but need an end-to-end transmission. Hence, due to the scalability in multimedia encoding technologies, different media content is essential for their users QoE. At present, there is no generalized scheme to offer benchmarked QoE for multimedia transmission over a wireless network.
- The trade-off between improvements in the physical layer and network technologies: Due to the greater demand for higher quality multimedia content, there is a provision to improve the network efficiency with available bandwidth. In this context, the role of media compression remains of core importance, ensuring that media transmission at a particular layer with appropriate quality while staying compatible with available bandwidth over the transmission channel. Furthermore, in the context of the growing requirement of multimedia streaming at higher QoS and QoE, there is a provision to introduce a transformational mechanism to solve the media compression problems and perform well beyond the conventional coding standards. Hence, more advancement in network technologies and lower supportability by its physical medium are translated. This demands more computational modeling.
- Less Efficient network coding for successful transmission: The massive requirement for multimedia streaming from the receiver devices to wireless channels faces multiple difficulties in which the bandwidth demand is limited. The efficient network coding allows the content transmitted over the network to be improved compared to conventional routing systems. However, there are several theoretical approaches have been introduced to assess the network capacity. For example, Zhang et al. [60] presented a physical layer network coding scheme for wireless transmission intending to improve the network performance with a high degree of freedom for media transmission. However, implementing a physical network coding system faces multiple practical problems that need to be solved, such as bandwidth issues, spectral efficiency, etc.
- High Media Content Loss Rate and Bit Error Rate: In wired transmission, there are more chances of content loss caused by the congestion of intermediator nodes. Meanwhile, wireless networks have a high bit error rate owing to multi-path fading and interference. The increasing packet loss rate or content loss and the bit error rate can directly affect media quality. Hence, there is a need to discover a reliable routing protocol to improve media quality with minimum loss.
- Energy Consumption Rate: As compared with fixed devices, there is good battery life in mobile devices. Usually, maintaining a media quality with minimum power consumption, i.e., transmitting and processing media content on mobile devices, are conflicting tasks. These trade-offs are valid, especially for wireless networks, i.e., wireless multimedia sensor networks. From the viewpoint of media coding, generating high-quality multimedia content usually consumes high power processing. Meanwhile, from the perspective of network performance, interference and multi-path fading require high transmission energy. So that, there is a need to design a reliable network that performs multimedia transmission with a minimum power consumption rate.
- Inadequacies of transport layer protocols performance: The existing transport layer protocol analyzes the primary reason behind the packet loss: congestion and unusual delay in the network. These two factors, i.e., packet loss and delay rate, affect the transmission process and media quality. However, in the wireless communication system, packet loss may occur due to network errors. Therefore, a QoS and transmission viewpoint should focus on designing a reliable transport protocol because the transmission process mainly occurs at the transport layer.
- Heterogeneity between the receivers and networks: The end-user is in the multimedia transmission process is quite different in QoS requirements, latency reduction, power consumption, processing capabilities, bandwidth demand, etc. Additionally, multimedia may deliver to varying networks with non-similar characteristics (i.e., delay, jitter, reliability, many more) and MAC (medium access control) mechanism.
- Lack of performance in the video compression standard: Existing multimedia transmission standards do not emphasize the signal quality, and the techniques lack a decision mechanism to perform compression. Moreover, existing methods do not ensure the perceptual quality of the multimedia framework to support higher pixel resolution. Although used more frequently in current times, HEVC is a new protocol; currently, no standard and potential studies exist in literature archival to further improve it. Very few works were found to adopt HEVC on a wireless mobile networking platform to check the efficiency of the HEVC algorithm and its potential to mitigate the loading impact of a dynamic traffic system (especially in a wireless environment).

Discussion: Apart from the points mentioned above, Table III highlights the contribution of the proposed survey work with some of the existing survey work. It is seen that existing review papers do not possess discussion of the

research gap. At the same time, their emphasis is particular, while the proposed study intends to offer a clear discussion about research challenges and contributes towards a simplified debate on the strength/weaknesses of related work.

TABLE III. COMPARISON WITH EXISTING REVIEW WORK

	Research Gap	Emphasis
Proposed Manuscript	Yes	Strength & weakness of existing schemes
Pal et al. [61]	No	Application
Wang et al. [62]	No	Mobile internet
Barakabitze et al. [63]	No	Quality of experience

The review findings show that multiple aspects can be classified into two parts, viz. i) standard protocol for wireless transmission and ii) research-based protocol for wireless transmission. There is a significant trade-off between this two. The first standard methods are meant for the theoretical formulation of transmission, while the second research part is particular and narrowed in its applicability process. Hence, not much higher scope is witnessed in the existing system. The existing literature is found with highly scattered technique implementation where multiple methods have been used to improve the performance. However, apart from this, a potential research gap exists, which requires immediate attention for future research work.

VI. CONCLUSION

The extreme growth in the various wireless communication technologies, the convergences of standard protocols, availability of small-sized hardware devices, software tools, and collaborative frameworks have paved a solid basis to visualizing real-time multimedia content, including; audio, video, and audio-visual applications in future endeavors. Furthermore, the advancement in the emerging wireless network technologies (i.e., 4G/5G) leads to having many multimedia-based new applications in the direction of augmented reality. However, order to perform transmission of multimedia content over the wireless channel is always a challenging factor. Various studies have focused on multimedia contents transmission over wireless sensor networks, mobile ad-hoc networks, using IEEE standard, etc. However, the mechanism adopted mainly applies conventional encoding techniques evolved from conventional discrete cosine transform, which does not hold much validity in upcoming encoders, e.g., HEVC (High-Efficiency Video Coding) standards. Therefore, from the comprehensive research study, there is a provision to investigate the better and reliable multimedia transmission proclaiming for wireless channels. The contribution of the present research study is to provide a dept investigational research study on understanding different existing techniques, challenges over multimedia transmission, and its significant impact on perceptual quality and traffic rate in wireless communication. The novelty of the study is in terms of its findings as research gap, i.e., i) The trade-off between improvements in the physical layer and network technologies, ii) Less Efficient network coding for successful transmission, iii) High Media Content Loss Rate and Bit Error Rate, iv) Energy Consumption Rate, v) Inadequacies of transport layer protocols performance, vi) Heterogeneity between the receivers and networks, vii) Lack of performance in the video compression standard.

Therefore, for the future, there is an aim to develop a multimedia transmission model that can perform seamless transmission of multimedia content irrespective of any adverse traffic conditions. Additionally, there is a need to design a novel and efficient compression mechanism using H.265 for next-generation wireless networks.

REFERENCES

[1] D. Striccoli, G. Piro and G. Boggia, "Multicast and Broadcast Services Over Mobile Networks: A Survey on Standardized Approaches and Scientific Outcomes," in IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1020-1063, Second quarter 2019, doi: 10.1109/COMST.2018.2880591.

[2] P. Lestari, S. Niyas and D. Krisnandi, "Depth Data based Chroma Keying using Grab-cut Segmentation," 2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA), 2018, pp. 118-123, doi: 10.1109/IC3INA.2018.8629501.

[3] A. Nauman, Y. A. Qadri, M. Amjad, Y. B. Zikria, M. K. Afzal and S. W. Kim, "Multimedia Internet of Things: A Comprehensive Survey," in IEEE Access, vol. 8, pp. 8202-8250, 2020, doi: 10.1109/ACCESS.2020.2964280.

[4] N. Minallah, I. Ahmed, M. Ijaz, A. S. Khan, L. Hasan, and A. Rehma, "On the Performance of Self-Concatenated Coding for Wireless Mobile Video Transmission Using DSTS-SP-Assisted Smart Antenna System", Hindawi-Wireless Communication and Mobile Computing, 2021.

[5] Kütner, Thomas. "Turning THz Communications into Reality: Status on Technology, Standardization, and Regulation." In 2018 43rd International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz), pp. 1-3. IEEE, 2018.

[6] M. C. Caballé, A. C. Augé, E. Lopez-Aguilera, E. Garcia-Villegas, I. Demirkol, and J. P. Aspas, "An Alternative to IEEE 802.11ba: Wake-Up Radio With Legacy IEEE 802.11 Transmitters," in IEEE Access, vol. 7, pp. 48068-48086, 2019, doi: 10.1109/ACCESS.2019.2909847.

[7] M. Jia, W. Liang, Z. Xu, M. Huang, and Y. Ma, "QoS-Aware Cloudlet Load Balancing in Wireless Metropolitan Area Networks," in IEEE Transactions on Cloud Computing, vol. 8, no. 2, pp. 623-634, 1 April-June 2020, doi: 10.1109/TCC.2017.2786738.

[8] Santos, Raul Aquino, ed. Broadband Wireless Access Networks for 4G: Theory, Application, and Experimentation: Theory, Application, and Experimentation. IGI Global, 2013.

[9] Huan Chen, Lei Huang, Sunil Kumar, C.C. Jay Kuo, Radio Resource Management for Multimedia QoS Support in Wireless Networks, Springer Science & Business Media, pp. 256, 2012.

[10] Su, Guan-Ming, Xiao Su, Yan Bai, Mea Wang, Athanasios V. Vasilakos, and Haohong Wang. "QoE in video streaming over wireless networks: perspectives and research challenges." Wireless networks 22, no. 5 (2016): 1571-1593.

[11] Hasan, Mohammed Zaki, Hussain Al-Rizzo, and Fadi Al-Turjman. "A survey on multi-path routing protocols for QoS assurances in real-time wireless multimedia sensor networks." IEEE Communications Surveys & Tutorials 19, no. 3 (2017): 1424-1456.

[12] Stokking, Hans Maarten, Mattijs Oskar Van Deventer, Fabian Arthur Walraven, and Omar Aziz Niamut. "Method and system for transmitting a multimedia stream." U.S. Patent 9,654,330, issued May 16, 2017.

[13] Kua, Jonathan, Grenville Armitage, and Philip Branch. "A survey of rate adaptation techniques for dynamic adaptive streaming over HTTP." IEEE Communications Surveys & Tutorials 19, no. 3 (2017): 1842-1866.

[14] Robert, Antoine, Omar Alvarez, and Gwenaël Doërr. "Adjusting bit-stream video watermarking systems to cope with HTTP adaptive streaming transmission." In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 7416-7419. IEEE, 2014.

[15] Chan, K. M., and Jack YB Lee. "Improving adaptive HTTP streaming performance with predictive transmission and cross-layer client buffer estimation." Multimedia Tools and Applications 75, no. 10 (2016): 5917-5937.

[16] Lim, Seong Yong, Joo Myoung Seok, Jeongil Seo, and Tag Gon Kim. "Tiled panoramic video transmission system based on MPEG-DASH." In Information and Communication Technology Convergence (ICTC), 2015 International Conference on, pp. 719-721. IEEE, 2015.

[17] Z. Li, M. Miao and Z. Wang, "Parallel Coding Scheme With Turbo Product Code for Mobile Multimedia Transmission in MIMO-FBMC System," in IEEE Access, vol. 8, pp. 3772-3780, 2020, doi: 10.1109/ACCESS.2019.2958482.

[18] Y. Xiao, G. Niu, L. Xiao, Y. Ding, S. Liu and Y. Fan, "Reinforcement learning-based energy-efficient internet-of-things video transmission," in Intelligent and Converged Networks, vol. 1, no. 3, pp. 258-270, Dec. 2020, doi: 10.23919/ICN.2020.0021.

[19] V. K. Ta and H. Oh, "A Pipelined Cooperative Transmission Protocol for Fast and Reliable Image Delivery in Wireless Sensor Networks," in IEEE Access, vol. 8, pp. 142758-142771, 2020, doi: 10.1109/ACCESS.2020.3013738.

[20] X. -L. Huang, Y. -X. Li, Y. Gao and X. -W. Tang, "Q-Learning-Based Spectrum Access for Multimedia Transmission Over Cognitive Radio Networks," in IEEE Transactions on Cognitive Communications and Networking, vol. 7, no. 1, pp. 110-119, March 2021, doi: 10.1109/TCCN.2020.3027297.

[21] F. Liu, Y. Liu, Y. Liu, and J. Yu, "Secure Beamforming in Full-Duplex Two-Way Relay Networks With SWIPT for Multimedia Transmission," in IEEE Access, vol. 8, pp. 26851-26862, 2020, doi: 10.1109/ACCESS.2020.2970612.

[22] T. He, J. Stankovic, C. Lu, T. Abdelzaher, SPEED: a stateless protocol for real-time communication in sensor networks, in Proc. IEEE International Conf. Distributed Computing Systems, 2003, pp. 46–55.

[23] K. Akkaya, M. Younis, Energy and QoS aware routing in wireless sensor networks, Cluster Comput. 8 (2–3) (2005) 179–188.

[24] E. Felemban, C. Lee, E. Ekici, MMSPEED: multi-path multi-SPEED protocol for QoS guarantee of reliability and timeliness in wireless sensor networks, IEEE Trans. Mobile Commun. 5 (2006) 738–754.

[25] S. Sanati, M.H. Yaghmaee, A. Beheshti, Energy-aware multi-path and multi-speed routing protocol in wireless sensor networks, in Proc. of 14th International CSI, CSICC 2009, Tahran, December 2009, pp. 640–645.

[26] S. Darabi, N. Yazdani, O. Fatemi, Multimedia-aware MMSPEED: a routing solution for video transmission in WMSN, in Proc. of 2nd International Symposium on Advanced Networks and Telecommunication Systems, Mumbai, India, December 2008, pp.1–3.

[27] Bennis, Ismail, Hacène Fouchal, Kandaraj Piamrat, and Marwane Ayaida. "Efficient queuing scheme through cross-layer approach for multimedia transmission over WSNs." Computer Networks 134 (2018): 272-282.

[28] Mahadevan, G. "A combined scheme of video packet transmission to improve cross-layer to support QoS for MANET." Alexandria engineering journal 57, no. 3 (2018): 1501-1508.

[29] Said Omar, Yasser Albagory, Mostafa Nofal, and Fahad Al Raddady. "IoT-RTP and IoT-RTCP: adaptive protocols for multimedia transmission over the internet of things environments." IEEE Access 5, no. 16 (2017): 757-16.

[30] Huang, Xiaohong, Kun Xie, Supeng Leng, Tingting Yuan, and Maode Ma. "Improving Quality of Experience in multimedia Internet of Things leveraging machine learning on big data." Future Generation Computer Systems 86 (2018): 1413-1423.

[31] Khernane, Nesrine, Jean-François Couchot, and Ahmed Mostefaoui. "Maximum network lifetime with optimal power/rate and routing trade-off for Wireless Multimedia Sensor Networks." Computer Communications 124 (2018): 1-16.

[32] Almasalha, Fadi, Farid Naït-Abdesselam, Goce Trajcevski, and Ashfaq Khokhar. "Secure transmission of multimedia contents over low-power mobile devices." Journal of information security and applications 40 (2018): 183-192.

[33] Cánovas, Alejandro, Miran Taha, Jaime Lloret, and Jesús Tomás. "Smart resource allocation for improving QoE in IP Multimedia Subsystems." Journal of Network and Computer Applications 104 (2018): 107-116.

[34] Hassan, Hammad, Muhammad Nasir Khan, Syed Omer Gilani, Mohsin Jamil, Hasan Maqbool, Abdul Waheed Malik, and Ishtiaq Ahmad. "H. 264 Encoder Parameter Optimization for Encoded Wireless Multimedia Transmissions." IEEE Access 6 (2018): 22046-22053.

[35] Ahmed, Adel A. "A real-time routing protocol with adaptive traffic shaping for multimedia streaming over next-generation of Wireless Multimedia Sensor Networks." Pervasive and Mobile Computing 40 (2017): 495-511.

[36] Majeed, Muhammad Faran, Syed Hassan Ahmed, Siraj Muhammad, Houbing Song, and Danda B. Rawat. "Multimedia streaming in information-centric networking: A survey and future perspectives." Computer Networks 125 (2017): 103-121.

[37] Xu, Changqiao, Wei Quan, Athanasios V. Vasilakos, Hongke Zhang, and Gabriel-Miro Muntean. "Information-centric cost-efficient optimization for multimedia content delivery in mobile vehicular networks." Computer Communications 99 (2017): 93-106.

[38] Moussaoui, Boubakeur, Soufiene Djahel, Mohamed Smati, and John Murphy. "A cross layer approach for efficient multimedia data dissemination in VANETs." Vehicular Communications 9 (2017): 127-134.

[39] Yang, Jiachen, Huanling Wang, Zhihan Lv, Wei Wei, Houbing Song, Melike Erol-Kantarci, Burak Kantarci, and Shudong He. "Multimedia recommendation and transmission system based on cloud platform." Future Generation Computer Systems 70 (2017): 94-103.

[40] Mohammed Ezz El Dien, Aliaa AA Youssif and Atef Zaki Ghalwash, "Energy-Aware Cross-Layer Framework for Multimedia Transmission over Wireless Sensor Networks", International Journal of Sensor Networks and Data Communications, Vo. 5(1), 2016

[41] Wu, Jiyan, Bo Cheng, Ming Wang, and Junliang Chen. "Delivering high-frame-rate video to mobile devices in heterogeneous wireless networks." IEEE Transactions on Communications 64, no. 11 (2016): 4800-4816.

[42] Hameed, Abdul, Rui Dai, and Benjamin Balas. "A decision-tree-based perceptual video quality prediction model and its application in FEC for wireless multimedia communications." IEEE Transactions on Multimedia 18, no. 4 (2016): 764-774.

[43] Chen, Bo-Wei, Wen Ji, Feng Jiang, and Seungmin Rho. "QoE-Enabled Big Video Streaming for Large-Scale Heterogeneous Clients and Networks in Smart Cities." IEEE Access 4 (2016): 97-107.

[44] Bradai, Abbas, Kamal Singh, Abderrezak Rachedi, and Toufik Ahmed. "EMCOS: Energy-efficient mechanism for multimedia streaming over cognitive radio sensor networks." Pervasive and Mobile Computing 22 (2015): 16-32.

[45] Gür, Gürkan. "Multimedia transmission over wireless networks fundamentals and key challenges." In Modeling and Simulation of Computer Networks and Systems, pp. 717-750. 2015.

[46] Han, Sangyup, Myungchul Kim, Ben Lee, and Sungwon Kang. "Fast Directional Hand-off and lightweight retransmission protocol for enhancing multimedia quality in indoor WLANs." Computer Networks 79 (2015): 133-147.

[47] Alvi, Sheeraz A., Bilal Afzal, Ghalib A. Shah, Luigi Atzori, and Waqar Mahmood. "Internet of multimedia things: Vision and challenges." Ad Hoc Networks 33 (2015): 87-111.

[48] Jiang, Wei, and Limin Meng. "IOT real-time multimedia transmission over CoUDP." International Journal of Digital Content Technology and its Applications 7, no. 6 (2013): 19.

[49] Mohammed, Anthony Olufemi Tesimi Adeyemi-Ejeye, Abdulrahman Alreshoodi, Geza Koczian Michael C. Parker, and Stuart D. Walker. "Ultra-High-Definition Video Transmission for Mission-Critical Communication Systems Applications." Multimedia Services and Applications in Mission Critical Communication Systems (2017): 115.

[50] Rao, K. R., Do Nyeon Kim, and Jae Jeong Hwang. "Video coding standards." The Netherlands: Springer (2014): 51-97.

[51] Richter, Thomas, Joachim Keinert, Siegfried Foessel, Antonin Descampe, Gaël Rouvroy, and Jean-Baptiste Lorent. "JPEG-XS—A

High-Quality Mezzanine Image Codec for Video Over IP." SMPTE Motion Imaging Journal 127, no. 9 (2018): 39-49.

[52] Willème, Alexandre, Benoit Macq, Antonin Descampe, and Gaël Rouvroy. "Power-Aware HEVC Compression Through Asymmetric JPEG XS Frame Buffer Compression." In 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 3598-3602. IEEE, 2018.

[53] Hameed, Abdul, Rui Dai, and Benjamin Balas. "A decision-tree-based perceptual video quality prediction model and its application in FEC for wireless multimedia communications." IEEE Transactions on Multimedia 18, no. 4 (2016): 764-774.

[54] Seetharam, Anand, Partha Dutta, Vijay Arya, Jim Kurose, Malolan Chetlur, and Shivkumar Kalyanaraman. "On managing quality of experience of multiple video streams in wireless networks." IEEE Transactions on Mobile Computing 3 (2015): 619-631.

[55] Hassan, Hammad, Muhammad Nasir Khan, Syed Omer Gilani, Mohsin Jamil, Hasan Maqbool, Abdul Waheed Malik, and Ishtiaq Ahmad. "H. 264 Encoder Parameter Optimization for Encoded Wireless Multimedia Transmissions." IEEE Access 6 (2018): 22046-22053.

[56] Chang, Che-Yu, Hsu-Chun Yen, Chun-Cheng Lin, and Der-Jiunn Deng. "QoS/QoE support for H. 264/AVC video stream in IEEE 802.11 ac WLANs." IEEE Systems Journal 11, no. 4 (2017): 2546-2555.

[57] Wu, Jiyan, Bo Cheng, Ming Wang, and Junliang Chen. "Delivering high-frame-rate video to mobile devices in heterogeneous wireless networks." IEEE Transactions on Communications 64, no. 11 (2016): 4800-4816.

[58] Khalek, Amin Abdel, Constantine Caramanis, and Robert W. Heath Jr. "Loss Visibility Optimized Real-Time Video Transmission Over MIMO Systems." IEEE Trans. Multimedia 17, no. 10 (2015): 1802-1817.

[59] de Miranda Regis, Carlos Danilo, Italo de Pontes Oliveira, Jose Vinicius de Miranda Cardoso, and Marcelo Sampaio de Alencar. "Design of objective video quality metrics using spatial and temporal informations." IEEE Latin America Transactions 13, no. 3 (2015): 790-795.

[60] S. Zhang, S. C. Liew, and P. P. Lam, "Hot topic: Physical-layer network coding," in Proc. 12th Annual Int. Con! Mobile Computing and Networking (MobiCom '06). New York, NY. USA: ACM, 2006. pp. 358-365

[61] Pal, Kunwar, Mahesh Chandra Govil, Mushtaq Ahmed, and Tanvi Chawla. "A Survey on Adaptive Multimedia Streaming." In *Recent Trends in Communication Networks*. IntechOpen, 2019.

[62] Wang, Mu, Changqiao Xu, Shijie Jia, and Gabriel-Miro Muntean. "Video streaming distribution over mobile Internet: a survey." *Frontiers Comput. Sci.* 12, no. 6 (2018): 1039-1059.

[63] A. A. Barakabitze et al., "QoE Management of Multimedia Streaming Services in Future Networks: A Tutorial and Survey," in IEEE Communications Surveys & Tutorials, vol. 22, no. 1, pp. 526-565, Firstquarter 2020, doi: 10.1109/COMST.2019.2958784.

A Systematic Literature Review on Regression Test Case Prioritization

Ani Rahmani[1], Sabrina Ahmad[2]*
Intan Ermahani A. Jalil[3]
Fakulti Teknologi Maklumat dan Komunikasi
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

Adhitia Putra Herawan[4]
Tokopedia Indonesia
Jakarta, Indonesia

***Abstract*—Test case prioritization (TCP) is deemed valid to improve testing efficiency, especially in regression testing, as retest all is costly. The TCP schedule the test case execution order to detect bugs faster. For such benefit, test case prioritization has been intensively studied. This paper reviews the development of TCP for regression testing with 48 papers from 2017 to 2020. In this paper, we present four critical surveys. First is the development of approaches and techniques in regression TCP studies, second is the identification of software under test (SUT) variations used in TCP studies, third is the trend of metrics used to measure the TCP studies effectiveness, and fourth is the state-of-the-art of requirements-based TCP. Furthermore, we discuss development opportunities and potential future directions on regression TCP. Our review provides evidence that TCP has increasing interests. We also discovered that requirement-based utilization would help to prepare test cases earlier to improve TCP effectiveness.**

Keywords*—Software testing; test case prioritization; regression testing; requirements-based test case prioritization; software engineering

I. Introduction

Software testing is a significant stage to confirm the quality of the software before it is released. Particularly in the software maintenance process, the study [1] demonstrated that the cost of testing implementation could reach 80% of the total maintenance costs. Therefore, further efforts are needed to reduce execution time in the testing process.

In the iterative-incremental process and the era of agile software development, new functions are increased by a short cycle [2]. Thereby, software development is also a process that is carried out continuously because of adding user needs. When there are changes in the software, new errors might appear. This situation will disrupt the previous stable system [3], [4]. For this reason, regression testing (RT) is needed, because it will verify the software to find the impact of changes to ensure its continued quality.

One of the popular techniques in RT is test case prioritization (TCP). This technique will order test cases in the test suite so that the testing execution will process the test cases with the most potential to find errors. The advantage of TCP implementation is that even if the testing process must be stopped for certain reason, the most significant errors have already been found. According to [5], there are two essential aspects of building TCP: determining the TCP approach and the technique to optimizing the TCP implementation.

In the past years, TCP studies gained significant attention and achievements to improve regression testing effectiveness. The study [6] emphasized that the researchers focus on five aspects: coverage criteria, algorithms, practical concerns involved, measurement techniques, and scenario to implement the technique. On the other side, studies [7], [8] explained that most of research efforts used source code as input resources to obtain the maximum number of faults within a certain period. Utilization of the code information is best applied to unit-level or block-level tests. Therefore, these efforts have limitations when applied to large systems since statements and block levels in source code will be challenging to manage [9], [10]. Utilizing code information will be expensive to implement because the tester must read and understand the source code, and this will take a long time.

Besides code-based, other TCP approaches have also been developed. According to a study [11], since a system is built from many requirements, the use of information from the requirements can increase error discovery. For this reason, some researchers argue it is essential to develop requirements-based TCP, while the studies in this area are still limited.

Therefore, the paper's main objective is to investigate TCP research's state of the art, emphasizing requirements-based TCP. The expected contributions of this study are:

1) To provide an overview of TCP developments in the years range from 2017 to 2020. We intend to highlight requirements-based TCP as one of the TCP approaches worth considering, and as far as we are concern, this is the first review on requirements-based TCP.

2) To present the variations of the TCP approaches and techniques explored so far, the diversity of software under test (SUT) used as an object for empirical evaluation, and the variation of metrics utilization to measure the TCP effectiveness. The results will be helpful to form a basis for future requirements-based TCP research.

Although there have been many studies in the form of TCP surveys, literature review, or mapping, each research has a different emphasis and perspective. In this regard, we have reviewed 48 credible papers from reputable journals and proceedings. Section II explains this in more detail.

*Corresponding Author

This paper is presented in stages, starting by looking at RT in general, followed by a study of TCP, and finally exploring the requirement-based TCP. Following the introduction, Section II presents the motivation and related work in RT and TCP. Section III describes the SLR method, including threats to validity and Section IV presents results and discussion. Subsequently, Section V describes the research findings, and Section VI offers future work and conclusions.

II. Motivation and Related Work

In this section, we explain the motivation and related work of the study conducted.

A. The Motivation

The ideal implementation of RT is to "retest all" or execute all test cases. However, in practice, not all test cases will be retested, especially those implementing RT manually. Several RT practice personal intuition based on experience, and even randomly [12]. Complete testing is complicated, and even worst, in several cases testing needs to be stopped. This condition causes other RT implementation problems, such as an error in the execution of the test case sequencing. On top of that, the RT process may be prolonged, or it may also run out of time. Studies [12], [13] stated that these approaches are inefficient and require high costs.

In many cases, RT is performed in high-pressure situations since testing execution requires a very long time. For example, the testing process conducted in an industry takes up to seven weeks to program with 20,000 lines [14]. In another case, Google has reported that there are more than 20 code changes every minute and that there is a change of 50% of files per month, resulting in a very long execution [15], [16]. The other example is a software development product with up to 30,000 functional test cases that need over 1000 hours. Besides, engineers need hundreds of hours to oversee the implementation of regression testing, supervise tests, monitor test results, and maintain test cases, oracles, and everything needed to support automated testing. Therefore, the study [17] concluded that RT is costly due to thousands of effort hours.

It is then understandably if several researchers emphasize that the common problem in RT is time constraint or insufficient [8], [18]–[20]. Through various surveys, research in the RT field will continue to grow, with the increasingly diverse types of approach or a broader application domain, for more effective methods.

RT techniques are divided into three types [2], [21]: regression test minimizing (RTM), regression test selection (RTS), and regression test prioritization (RTP), or also known as test case prioritization (TCP). A study [22] summarizes the comparison of the three techniques which are presented in Table I.

TABLE I. The Comparison of Regression Testing Approach [22]

Component	Regression Test Approaches		
	Minimizing (RTM)	Selection (RTS)	Prioritization (TCP)
Strategy	Eliminate test case	Modification aware test case	Test case permutation by ordereing and prioritizing
Strength	Effective in reducing test case	Effective in selecting modification-aware test cases	Usefull when new test case will always be considered in the test case permutation
Limitation	Test case are not modification-aware	New test cases might be missed out in the temporary selection that is modification-aware	Time consuming, larger test-suuite

RTM reduces test cases by removing many test cases for a particular reason, such as redundant ones. Meanwhile, RTS selects test cases that can potentially find errors. The selection process refers to specific criteria. Both RTM and RTS will permanently remove some test cases from the test suite. Unlike RTS and RTM techniques, TCP does not remove test cases but orders the test cases according to the criteria. The test case with the most potential to find an error in the program will have a higher priority and be executed earlier.

B. Related Work

Some surveys or reviews have been conducted on RT and the TCP techniques. This section describes the study, SLR, and mapping obtained from many digital libraries in 2010-2021 range.

Regression testing survey is available in several studies [2], [16], [21]. The study [2] surveyed RT in the scope of the technical side, metrics, strategy, software under test (SUT), and an overview of the optimization technique in the form of automation, or using a traditional approach. Meanwhile, the study [16] described the techniques and advantages of all three types of regression testing. Study [21] on the other hand, reviewed articles with the most extended ranges from 1977 to 2009. This study discussed the approaches and techniques covering test case minimizing effort, test case selection, and TCP in great detail.

The specific survey on TCP was performed in [5], [22], [23] [24], [25], [26], 33], [27], [28], and [29]. Survey [22] and [23] are two very detailed surveys and have been cited by many TCP researchers to date. The study [22] reviewed 80 articles from 1999 to 2016, while [23] reviewed 65 papers from 1997 to 2011. Generally, the aspects explored in the two studies include approaches and techniques on TCP, metrics, and software under test (SUT).

In analysing TCP, study [5] explained that there are two approaches to categorize TCP implementation: input resources (the information sources for the TCP process) and optimization strategies (methods or algorithms for executing the TCP technique). This study classified the TCP approaches and the TCP optimization strategies according to these two categories. This method is a more straightforward step to facilitate TCP classification. In measuring the TCP effectiveness, this study proposes another view of the metric used by many researchers, which is the average percentage error detection (APFD). However, APFD has limitations because it treats all test cases as having the same weight.

A survey [24] has mapped and reviewed 108 articles from 1999 to 2016. The author mapped article content into several aspects: the place of publication, the number of articles on the approach, and the number of metrics. Furthermore, the review includes the use of tools, the TCP effectiveness for each study investigated, the analysis of APFD factors, and a review of APFD in some SUT applications.

The model-based TCP has been studied [25] which reviewed 32 articles from 2005 to 2016. The authors classified the TCP models based on approaches, their characteristics, and how they can overcome obstacles in TCP implementation using model-based as an input resource.

The study conducted by Mukherjee and Patnaik [26] surveyed 90 TCP articles from 2001 to 2018. The purpose of the survey is to investigate several aspects: TCP Metric, the program or SUT, and identify the TCP method commonly used. This study concludes three essential perspectives: 1) the APFD metric is the most extensive to measure the effectiveness of TCP, 2) the program in the SIR repository is the most widely used as SUT, and 3) the coverage-aware, requirements-based, and model-based are the three approaches that are getting more attention, currently.

In 2019, Lio et al. [30] surveyed 191 articles on TCP published in the 1997 to 2016 range. They analyzed TCP trends based on six categories: constraints, algorithms, criteria, measurements, scenarios, and empirical studies. In addition to this, they highlighted several improvements during the development of test cases in 2004–2005, 2008–2009, and 2014–2015. They analyzed the trends of the period from various points of view as a basis. More specifically, the analysis was related to the emergence of technologies that allow online repositories to host software projects.

Meanwhile, a study [27] have reviewed TCP trends from 2017-2019. An essential aspect of this study is to answer whether the taxonomy proposed in the previous study [22] is still valid. This study further suggests other approaches: location-based, machine learning-based, neural network-based, and empirical, which are empirical studies of TCP in certain domains, with specific guidelines or software.

Recently, two more literature reviews on TCP are published in 2021. Samad et al. [28] reviewed TCP in general, and Hasnain et al. [29] specifically reviewed TCP's functional requirements. Samad et al. reviewed 52 TCP articles in the 2007-2020 range. Like most studies on regression testing and, in particular, TCP, the RQ proposed in this study is a state-of-the-art of TCP technique, parameters, dataset or object software used, and metrics to verify TCP techniques. The parameters used in the study include cost, code coverage, and fault detection ability.

The study conducted by Hasnain et al. [29] focuses on TCP studies that utilize the functional requirements approach, with 35 article from 2009 to 2019. The study answered 7 RQs: state-of-the-art regarding functional requirements-based TCP, the key factors discussed in the TCP requirements-based study, the essential aspects considered for proposing the TCP approach, the crucial issues addressed in the TCP functional-requirements study, test case size and type of defect, metrics used, software under test (SUT), and whether these studies can be applied in the real world or not.

There are five surveys on both RT and TCP for specific purposes. The study [15] reviewed the trend of the TCP approach in web applications and analysed the qualitative assessment of web applications. The analysis was carried out on three web application sizes: small, medium, and large, and was analysed from two categories: simple and complex web applications. Meanwhile, a study has been conducted [31] to map the regression testing applications on web services. The mapping aims to identify gaps between existing studies and the future studies in each article reviewed. The study mapped several things: stakeholders, SUT and related standards, validation methods, and web services, as well as mapping to validation services.

Moreover, to review the use of TCP techniques in web services, a study [32] has identified statistical methods, metrics to validate the proposed technique, and issues relating to current TCP concerning web services. Furthermore, a study [33] reviewed the scope of TCP's application for continuous interaction (TCPCI) environment. Some important aspects were analysed, including problems in continuous integration (CI), sources of information (input resources) for TCP in TCPCI, evaluating measures using metrics in TCP, and analysis of research opportunities to guide future research.

A study by [34] analysed 98 articles to support the research. The authors analysed and mapped several aspects, including the techniques and the efforts to improve the test's scope. The authors also construct a taxonomy that allows researchers to consider the relevance and applicability of regression testing to specific industries. Table II presents the secondary studies, whether in the form of SLRs, surveys, or mapping, from 2010 to 2020, grouped by RT, TCP technique, and RT or TCP for specific purposes.

TABLE II. SECONDARY STUDIES IN REGRESSION TESTING (RT) AND TEST CASE PRIORITIZATION(TCP)

	#Study	Publication Year	Type of Studied	Year Coverage	#of Primary Studies	Other Information
RT	[21]	2010	Survey	1977-2009	159	
	[16]	2016	Survey	-	-	
	[2]	2016	Survey	2000-2014	25	
TCP	[23]	2012	SLR	1997-2011	65	
	[24]	2017	Mapping	1999-2016	108	
	[22]	2017	SLR	1999-2016	80	
	[5]	2018	Survey	-	-	
	[26]	2018	Survey	2001-2018	90	TCP approaches
	[25]	2018	SLR	2005-2016	32	Model-based TCP
	[30]	2019	Survey	1997-2016	191	
	[27]	2020	SLR	2017-2019	320	
	[28]	2021	SLR	2007-2020	52	
	[29]	2021	SLR	2009 to 2019	35	Functional requirement-based
RT / TCP for Specific Purpose	[31]	2014	Mapping	2000-2013	30	RT for Web Service
	[15]	2015	SLR	1995-2014	64	RT for Web Appl.
	[34]	2019	SLR	x-2016	98	RT in Industry-relevant
	[33]	2020	Mapping	1979-2020	35	TCP in Continuous Integration
	[32]	2020	SLR	2001-2017	65	TCP for Web Service

III. REVIEW METHOD

We adopted a Systematic Literature Review (SLR) strategy [35] as a method. SLR is a research method for conducting a literature review with systematic and regular steps. According to the method, Table III presents three stages of review: the initial or planning stage, the selection and review process, and the reporting of the resulting process.

A. Research Question

The research questions (RQs) are intended to find the techniques, approaches, and empirical experiences from many researchers to formulate an efficient way to process regression testing using TCP techniques and requirement-based TCP. The results of the SLR must be able to answer several questions in Table IV.

B. Selecting and Review Process

This section explains several stages of activities in implementing the SLR.

1) Literature resources: The articles used in this study are taken only from journals and proceeding. We selected the most common and influential database sources and the ones most widely used by researchers, as listed below:

a) IEEE Xplore
b) Science Direct
c) Springer
d) Semantic Scholar
e) Google Scholar

TABLE III. SYSTEMATIC LITERATURE REVIEW STAGE

SLR Phase	Steps
Planning	Formulating the research questions
Selecting and Review	Determining the data sources
	Determining search strings/keyword
	Applying inclusion and exclusion criteria
	Selecting, classifying, and analyzing the references.
Reporting	Presenting the SLR result

TABLE IV. LIST OF RESEARCH QUESTIONS

#RQ	Research Questions	Motivations
RQ1	What is state of the art for TCP in regression testing based on TCP approaches and techniques?	To discover the development of approaches in TCP study
RQ2	What is the software under test (SUTs) in TCP studies?	To identify the variation of SUTs in TCP studies. This will be useful for researchers to prepare the SUT carefully.
RQ3	What is the trend of metrics to measure TCP effectiveness?	To provide insight into how the effectiveness of approaches or techniques is measured.
RQ4	What is the state of the art of requirement-based TCP in literature?	To explore techniques or approaches studied in the requirements-based TCP.

2) Search string criteria: We formulated a string for the search process considering its relevance to the research question. Sometimes we used several words by combining them into a query for words with similar meanings, such as "technique," "approach," or "strategy." Furthermore, to emphasize a string, quotes are also used in a phrase, such as "regression testing" or "test case," so that search results can be more specific. The keywords for the query search string used are: "test case" AND (prioritization OR prioritize) AND (approach OR technique OR strategies) AND "regression testing."

3) Inclusion/Exclusion criteria: The next stage is selecting articles based on inclusion criteria (ICs) and exclusion criteria (ECs). Table V explains four inclusion and exclusion criteria.

4) Selection and quality assessment: We decided to choose papers published started in 2017 to answer the Research Questions. The reason is because studies conducted from 1999 to 2016 [22] and from 1997 to 2011[23] have been in detail reviewed, and researchers to date have widely cited the results.

Using the query stated in sub-section 3.2.2, we discovered 501 papers in the primary studies from various databases. These papers are published in both journals and proceedings. Next, we selected the papers using the inclusion and exclusion criteria as presented in Table V, resulted in 122 papers being selected. Furthermore, we conducted a quality assessment based on the following five parameters:

a) The objectives are clearly described.

b) The article clearly states the used approach or technique.

c) There is sufficient information about the software under test (SUT) as a research object.

d) The research design is appropriate to answer the research question.

e) Conclusions are stated clearly and measurably using one or more metrics.

The five above parameters must be "true," otherwise the paper will be excluded to obtain the expected quality. At this selection stage, 48 papers were finally listed. Fig. 1 describes the process of sources search and selection.

5) Data extraction process: The data extraction stage aims to collect data from selected papers, which is done by extracting information to answer the research question (RQs) defined. Table VI is a list of extraction parameters along with the research question to be answered.

C. Threats to Validity

There is a risk of threats to validity in the review survey even though careful measure has been taken care of throughout the survey. In this survey, there are two threats of validity as listed below:

1) There may be missing credible sources, which is beyond our knowledge. To minimize the threat, we search from the most common and influential database sources.

2) There is a possibility there exist relevant studies but are not captured by the keywords due to the differences in terminology and mentions. For this matter, we have searched, and test various search string combinations as stated in Section III.B.2.

TABLE V. LIST OF INCLUSION: EXCLUSION CRITERIA

#ICs	Inclusion Criteria	#ECs	Exclusion Criteria
IC1	The document selected is an article from a journal or proceeding	EC1	Lecture note, book chapter
IC2	The articles taken are those related to the focus study in this research, whether explicitly proposing new approaches/techniques, or studies that examine the effectiveness of a technique, through comparisons, or empirical studies	EC2	Articles that discuss in the form of an overview of these concepts
IC3	The articles published 2017-2020	EC3	The articles published outside of 2017-2020
IC4	The articles written in English	EC4	The articles in languages other than English

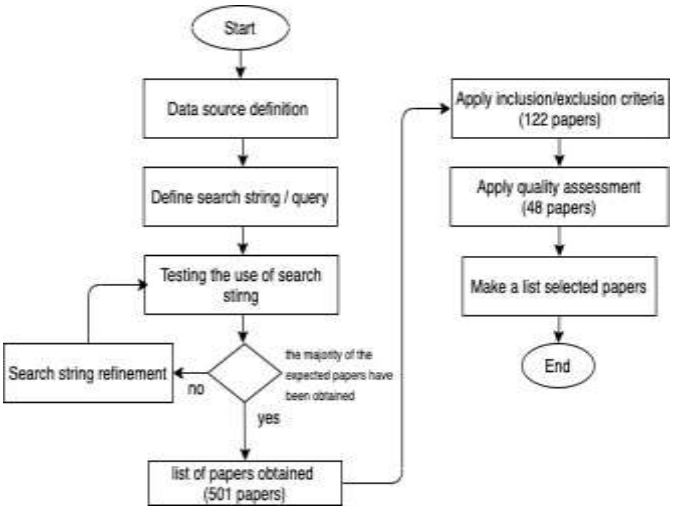


Fig. 1. Search and Selection Process.

TABLE VI. THE DATA EXTRACTION PARAMETERS

Research Question	Extraction Parameters
RQ1	TCP approaches and techniques
RQ2	SUT for empirical studies
RQ3	Metric used to measure the TCP effectiveness
RQ4	The strategies to implement the requirements-based TCP

IV. Results and Discussion

This section elaborates on the review results.

A. Primary Studies Overview

From the first search 501 articles were obtained from the databases. There are 235 journal articles, and 266 proceedings articles. Fig. 2 shows the distribution of articles obtained from 2017 to 2020. While Fig. 3 presents the comparison of the journal and proceeding in the first-round search.

When taking the inclusion and exclusion criteria into account, 122 papers were shortlisted, as shown in Table VII. Next, we filtered the shortlisted papers using the five quality assessment criteria, and only 48 were finalized (Table VIII). The detailed information of selected articles can be found in the Appendix.

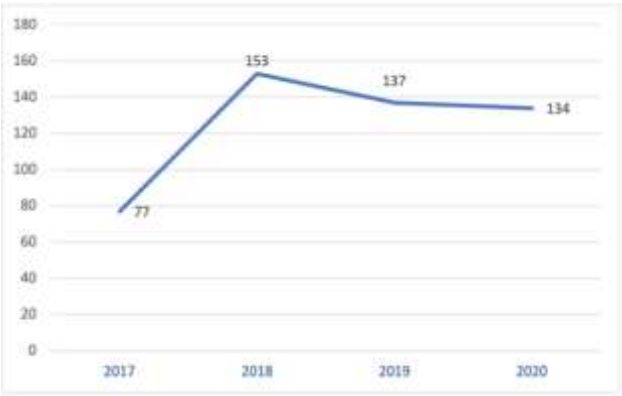


Fig. 2. First Search Results.

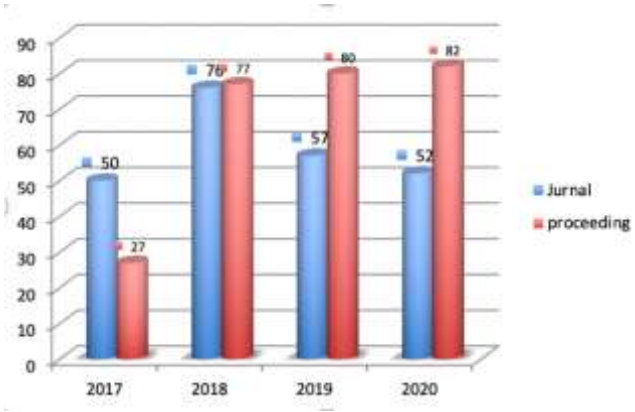


Fig. 3. Journal Articles and Proceedings Distribution.

TABLE VII. Total Articles during the Inclusion / Exclusion Selection

Year of Publication	Total articles selected	First-round	
		Included	Excluded
2020	134	39	95
2019	137	37	100
2018	153	28	125
2017	77	18	59
TOTAL	501	122	379

TABLE VIII. Total Articles during the Quality Assessment

Year	Result of the first round	Second round	
		Included	Excluded
2020	39	12	27
2019	37	16	21
2018	28	11	17
2017	18	9	9
TOTAL	122	48	73

The 48 primary studies consist of 31 journal articles and 17 proceedings, as illustrated in Fig. 4.

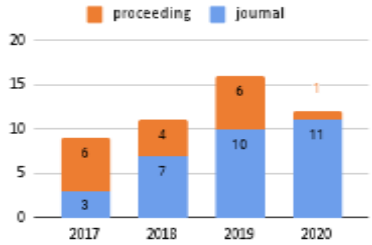


Fig. 4. Selected Papers through Two Rounds Selection.

Next, Fig. 5 shows the selected articles classification based on the origin (journal or proceeding) and quartiles in Scimago indexing. It is shown that 33.3% of articles are from Q1 journals, 18.8% are from Q2 journals, 8.3% are from Q3 journals, and 4.2 % are from Q4 journals.

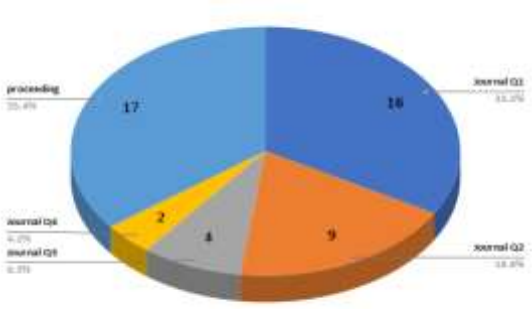


Fig. 5. Selected Papers Sources.

B. Current Research Efforts to Improve TCP for Regression Testing

This sub-section responds to RQ1, RQ2, and RQ3.

1) What is The State of the Art for TCP based on TCP Approaches and Techniques? (RQ1): The answer to RQ1 also covers the review done by Khatibsyarbini et al. [22] since it is essential to consider the improvement of TCP research before 2017. The significant discovery is the TCP taxonomy which portrays the regression testing types and some techniques in TCP. Fig. 6 shows the TCP approaches taxonomy proposed by [22] and portrays approaches added by [27]. Four items are added into the initial taxonomy: ' Location-based,' 'Machine-learning based,' 'Neural Network-based,' and 'Empirical.'

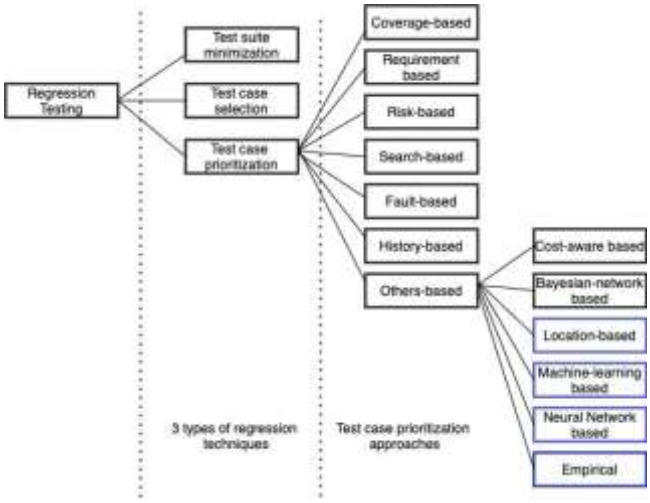


Fig. 6. Taxonomi of TCP Approach (Adapted from [22]).

Table IX presents the approaches in the TCP research during 2017-2020. While Fig. 7 visualizes the approaches distribution in the TCP research, it can be seen that several approaches are gaining popularity as they appeared in several researches.

TABLE IX. APPROACHES IN THE TCP RESEARCH

Approaches	Research
Risk based	[36][10]
Search based	[37] [38] [39] [40] [41] [42][43]
Fault based	[44] [45] [46] [47] [48]
Model based	[49][50] [46][51]
Modification based	[52][53][54]
Coverage based	[55] [56] [57] [58] [59] [60]
Similarity based	[61] [62][63] [64] [65][66] [67]
Requirement based	[68][69][70] [13][71]
User Interface based	[72]
History based	[73] [74]
Mutation based	[75]
Hybrid (combining more than 1 method)	[76][20][77] [53] [75]

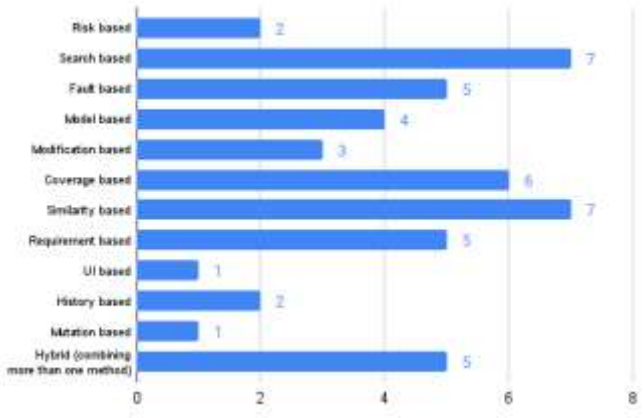


Fig. 7. The Trend of TCP Approaches.

In comparison to the approaches proposed by [22] and [27], we discover several other approaches through our survey: modification-based, user interface-based, model-based, mutation-based, and similarity-based. This discovery shows that researchers are still exploring and improving ways to better TCP by introducing more approaches.

Input resource, technique, and algorithm determination are essential to implement TCP [5]. Referring to the literature, there is no dominant technique or algorithm for implementing TCP. After we identified the TCP approaches, we then identified the techniques that researchers used in their study. Each researcher executes the chosen technique based on specific analysis and considerations. Some of the algorithms used include Greedy and Additional Greedy for search-based TCP [37], [40], Firefly Algorithm [38], [78], Neural Network Classifier [44], Ant Colony Optimization [55], [70], FAST Algorithm [79], Support Vector Machine/SVM [80], Genetic [42], [59], [76], [81], Fuzzy Expert [77], Dynamic Programming [45], Recommender System [58], Clustering Technique [73], [82], [83], Particle Swarm Optimization [61], Natural Language Processing (NLP) [74], and Bat-inspired Algorithm [48].

Based on our survey, we found that some researchers used more than one approach or technique in their study. For example, a study combined estimated risk value, coverage information, and fault detection [53]. Another study compared the mutation-based and diversity-aware [75]. Finally, there is also a study that looked into requirement and risk-based [84].

2) What is The Software under Test (SUT) in TCP Studies? (RQ2): Software or system under test (SUT) is a complete system as the object or target of testing. A well-structured and centralized SUT infrastructure can gradually build knowledge [85]. In this study, SUTs for evaluation are diverse. We classify the utilization of SUTs based on five categories: 1) researchers build their SUTs using open source from public resources, such as Github or other sources. In this case, the researchers design the fault and test cases for a specific purpose; 2) researchers utilize the SUT from the dataset or repository such as SIR, Defects4J, or others. In this case, researchers only need to explore and directly use the SUT from the repository; 3) researchers use the software from the industry as the cases with scale variations; 4) researchers build a software, create some faults and some test cases; 5) Others SUTs. Table X shows the distribution of SUT utilization according to the five categories.

The SIR and Defect4J repositories are still widely used as sources for SUTs. Besides, many researchers build and open-source SUT as a research object. Fig. 8 illustrates the distribution of SUT usage according to the five classifications described in Table X.

TABLE X. UTILIZATION OF SOFTWARE UNDER TEST (SUT)

SUT	Research
Building SUT by utilizing open source from public sources such as Github or the others	[86][20][1][45][41][58][60][48][42]
Utilizing SUT from the available database such as software - artifact infrastructure repository (SIR) and Defect4J	[37] [38] [39] [44] [49] [52] [55] [56] [40] [62] [77] [83] [66] [87][10] [59] [79] [43] [73] [61] [53] [64]
Real case from the industry	[49][49][20][1][46][58][51][60][47][42] [74]
Software, faults, and test-cases developed by the researcher	[82]
Others software	[36][71][76][69][81][70][13][71][42]

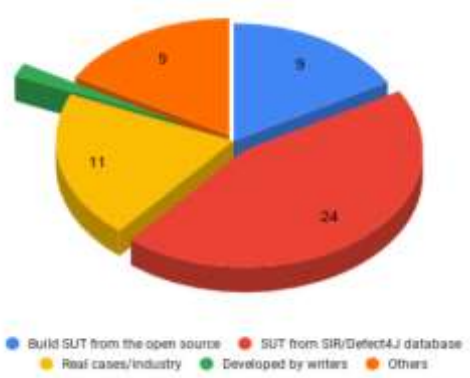


Fig. 8. Distribution of SUT Utilization.

3) What is the trend of metrics to measure the TCP effectiveness? (RQ3): In TCP studies, researchers generally aim to present the effectiveness of the developed techniques. In this regard, some metrics are known, as shown in Table XI. To answer what is the trend of metrics utilization to measure the TCP effectiveness, we identify metric utilization in all studies.

Several studies utilize more than one metric in their research. As in previous studies [22] and [23], the average percentage fault detection (APFD) was used dominantly in many TCP studies, while other metrics are spread out in less specific numbers. Table XII shows metric utilization in the studies, and Fig. 9 visualizes the distribution of metrics used.

C. What is the State of The Art of Requirement based-TCP (RQ4)?

Section 4.2 presents current research efforts to improve TCP for RT, while this section narrows the focus on current research efforts to improve requirements-based TCP.

Prior to conducting a review of the current research effort, we consider it is necessary to review the development of requirements-based TCP before 2017. Almost all TCP surveys that discuss requirements-based TCP start with prioritizing requirements for tests (PORT) [9] as the basis. The primary references for requirements-based TCP prior to 2017 are from studies [22] and [30]. The following is our exploration of the requirements-based TCP development including studies prior to 2017.

PORT is a value-driven approach to implementing TCP at the system level. Study[9] prioritizes the test-cases refer to four parameters: requirement volatility (RV), customer-assigned priority on requirements (CP), fault proneness of requirements (FP), and developer-perceived implementation complexity (IC). To determine the test case prioritization, each factor is carried out and given a score. For example, CP is rated with a range of 0-10, where 10 is the highest priority value. The evaluation result shows that the PORT technique can increase the detection rate of severe errors compared to executing a random test case. More specifically, CP is the most influential factor in increasing the PORT effectiveness and next IC. They used two metrics to measure the PORT effectiveness: the average severity of faults detected (ASFD) and total severity fault detection (TSFD).

TABLE XI. DESCRIPTION OF METRIC TO MEASURE THE TCP EFFECTIVENESS

Metric	Description
APFD	Average Percentage Fault Detection
APFDc	Average Percentage Fault Detection and Cost
Modified APFD	Modified Average Percentage Fault Detection
NAPFD	Normalize Average Percentage Fault Detection
EPS	Epsilon
ECC	Effectiveness of Change Coverage
PTRSW	Percentage of Total Risk Severity Weight
APCC	The average percentage of λ-wise combinations covered/ Average Percentage of Combinatorial Coverage
RP	The Average Relative Position
HMFD	The harmonic means of the rate of fault detection
APTC	Average percentage of test-point coverage
eAPWC	Enhanced average percentage of win-Cost coverage
NTE	The Number of Test to be Evaluated
HV	Hypervolume (HV) measures the volume in the objective space covered by the produced solutions with the range from 0 to 1 and a higher value of HV denotes a better performance of the algorithm
APSC	Average Percentage of Statement Coverage
Requirement coverage	The number of requirements covered by test
Code coverage	How many codes were executed while test performed? Can be in the form of number of line (line coverage), branch (branch coverage), or even path (path coverage).
Test case & patch diff.	Difference between the number of test cases and patches generated between approaches
Similarity	Test case similarity measures the distance between two test cases and returns a value within the range [0,1].
Prior-aware similarity	Prioritization-aware Test Case Similarity. Measures the average similarity of each of the test cases with its preceding test cases (i.e., test cases that were prioritized before)
Severity	Severity detection per test case execution (early detection of severe faults)

TABLE XII. METRIC UTILIZATION IN THE TCP RESEARCH

Metric	Research
APFD	[36][38][39][44][49][56][40][57][86] [62][72][20][77][83][88][66][87][41] [50] [79][43][73][75][60][82][47][48] [61][53][74][64]
APFDc	[38][80][1][45][87][51][10][59]
Modified APFD	[44]
NAPFD	[37][58]
EPS	[39]
ECC	[39]
Execution Time	[38][52][55][56][40][80][76][20][42] [41] [65][48]
PTRSW	[36]
APTC	[70]
eAPWC	[70]
APCC	[57][64]
RP	[54]
HMFD	[86]
NTE	[60]
Requirement coverage	[13][42][68][69]
Code Coverage	[55][56]
Fault detected	[40][46][71]
Test case and patch diff. count	[52]
APSC	[59]
Similarity	[42]
Prioritization-aware-Similarity	[42]
Severity	[45]

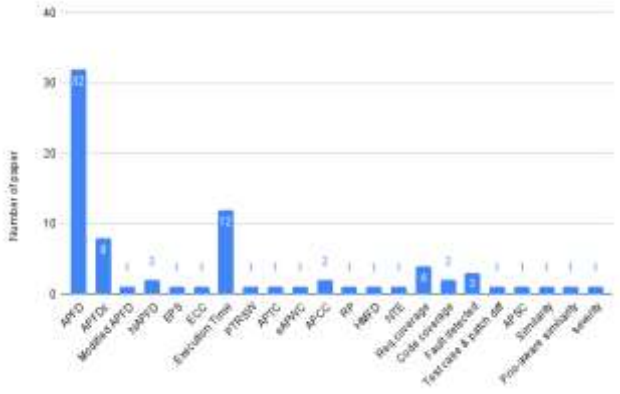


Fig. 9. Distribution of Metrics Utilization.

The following requirement-based TCP was introduced in 2009 [11], involving six factors: changes in requirements, customer assigned priority of requirements, fault impact, developer-perceived code implementation complexity, application flow, and usability. The authors divided the six factors into three factors for testing at the initial version stage and three factors for the regression testing stage. Furthermore, this study proposed a technique or steps to prioritize test cases using requirements-based factors. This stage information can be a reference for TCP requirement-based researchers.

A TCP through correlation of requirement and risk has been studied by Yoon et al. [89]. They reported TCP's risk-based testing (RBT) technique using defining risk items and estimated the risk exposure value derived from the requirement. The calculation of risk exposure value is determined based on requirement risk weight and the value of risk exposure. Specifically, they defined product risk items, which are expected to be helpful for the risk identification process. They also presented empirical studies comparing the effectiveness of their approach with other prioritization approaches. This empirical study shows that the utilization of risk exposure is promising in terms of effectiveness and can detect severe errors. This condition will have an impact on efforts to save time and costs.

In addition, Arafeen and Do [90] reported about TCP method using requirements-based clustering. They used a machine-learning algorithm to cluster the textual similarity among requirements. The clustering technique classified the distribution of words that co-occurs in their requirements. There are three tasks in this process: term-document matrix construction, term extraction, and k-means clustering. Their empirical study showed that the method could improve the effectiveness of TCP. Their empirical study showed that the method could improve the effectiveness of TCP.

Throughout years, several studies have been carried out to deal with requirements-risk in requirements-based TCP [8], [77], [91], [92]. These studies were seen as a series of efforts to further improve TCP based on requirement risk. Since PORT [9] was introduced, the researchers further explore the fuzzy expert system to prioritize test cases systematically [92] and later was investigated empirically with industry cases [77].

Many types of factors were utilized in the research conducted on test-case prioritization [8], such as utilized requirements modification status (RMS), requirements size (RS), requirements complexity (RC), and potential security threats (PST). Meanwhile, a study [77] reported four indicator risks to propose their approach, which are RC, fuzzification, a potential security risk (PSR), and requirements modification level (RML).

The other types of requirement risk factors was explained in [91], which proposed general steps to prioritize test: 1) estimating the risk and requirements correlation; 2) calculate the risk weight for all requirements; 3) calculate the exposure value; 4) evaluate additional factors for requirements prioritization; and 5) prioritize the requirements and test cases for all requirements.

The researchers utilized some risk factors to implement TCP, while the other researchers implemented the requirement-risk TCP for specific software. For example, a study has been carried out [36] to calculate the risk value from some parameters of requirement complexity, such as methods failure likelihood (MFL), method complexity (MC), change requirements (CR), methods failure impact (MFI), and method size (MS). The result of these calculations then used to determine the prioritized test suite. Meanwhile, study [93]

utilized five aspects of risk to formulate their framework: risk item type, characteristic, measurement method, calculation procedure, and risk level.

A study [94] utilized the correlation of requirements to build the TCP technique. Their study calculated requirements priority (RP) based on customer-perceive priority (CP), development perceives priority (DP). RP of the i-th calculation in the formula $RPi = CP_i + DP_i$. They assumed that the CP and DP have equal weight. The authors claimed, this TCP technique was efficient on a small-scale study, and their method was better than the sorting process.

The discussion of requirements-based TCP since 2017 was begun with an analysis of research [71] which implement the TCP using requirement dependency with four parameters: test cases, test requirements, errors, and costs. On the other hand, they defined other elements related to functional requirements and requirements dependencies. Therefore, the authors use the algorithm to prioritize test cases considers the objectives of optimization, error detection, and cost.

The study [68] presents the requirement-dependency TCP by modeling the requirements and information of the test-related and their relationships with some aspects such as stakeholder affiliation, stakeholder's assigned priority, cost, time, risk, and business value. In prioritizing the test cases, they utilize the PageRank algorithm.

Study [69] utilized information coverage as an input resource. The authors proposed the use of complex test cases to test the requirements coverage. With complete coverage, the error detection rate also increased. At the same time, study [94] explained TCP's usage based on requirement correlations. When the testing process detects errors in a functional requirement, other correlated requirements may contain similar errors or other errors depending on the correlation between the two requirements. This study gives a better understanding of requirements correlation and its impact to be further explored in future TCP research. The parameters for the prioritization process use customer priority (CP) and developer priority (DP). Both of which are assessed by humans to produce a requirement priority (RP) as the initialization stage for the test case prioritization process.

In 2019, a study [77] utilized requirements risks in requirements-based TCP which introduced the fuzzy logic to reduce the humans' role in estimating risk factors for prioritizing test cases. This study is a continuation of the previous requirements-risk survey, which started in 2014[91].

We investigate more on TCP using requirement dependencies researched by two studies [68] and [71]. These studies are essential to explore because requirements-based TCP research focuses on the use of information in requirements, such as the interactions between requirements that influence the feasibility of the functionalities. This interaction is known as requirements dependency. The study [71] compared the cost-effectiveness of testing between the Greedy Method and the Genetic Algorithm (GA). The study prioritizes the GA to form a test suite that ensures all the defined requirements and has the lowest cost and highest fault detection.

TABLE XIII. THE STUDY ON REQUIREMENT-BASED TCP

Studies	Requirement-information	Software Under Test (SUT)	Metric Used
[9]	requirements volatility, customer priority (CP), Implementation Complexity (IC), and requirement's fault proneness.	Four projects developed by students in the advanced graduate	ASFD, TSFD,
[11]	Customer assigned priority of requirements, developer-perceived, code, IC, change requirement (CR), application flow, fault impact, and usability	Five projects (Phase1); Project with 5000 LOC (Phase2); Industrial Case, Cosmosoft Technologies Limited (Phase3)	TSFD, ASFD, TTEI, ATEI
[89]	Requirement risk	Program from Siemens	APFD
[90]	Requirement clustering	Java programs containing multiple versions (two program)	APFD
[91]	Requirement risk	Open-source and capstone project.	APFD
[8]	Requirement risk	Enterprise-level IBM analytics application.	APFD
[92]	Requirement risk	Open-source and the industrial (one program)	APFD
[94]	Requirement risk	Industrial case study	APFD
[71]	Requirement dependency	A synthetic case study	APFD
[68]	Requirement dependency	Small example case	Requirement coverage
[69]	Requirement coverage	Own case study	APFD
[77]	Requirement risk	Industrial application	APFD

Meanwhile, Abbas et al. [68] made a requirement dependency meta-model on non-functional requirements and performed TCP using the Page Rank Algorithm. This requirement dependency value was used as an addition to the priority ranking weight for these requirements. In summary, Table XIII presents the requirement-based TCP research conducted to date.

V. RESEARCH FINDING

Regression testing is a crucial stage in the software development process especially in the era of Agile development. TCP appears as the most popular technique in regression testing due to testing efficiency. Even if testing must be stopped for some reason, the high-priority test cases have found the essential faults.

We have conducted a rigorous study through searchers from reputable resources and carefully filtered the findings through a quality assessment to review current efforts on TCP for RT. In RQ1, we found that the existing TCP approaches presented by [22] and later added by [27] are still of interest to

researchers. We discover five new approaches through the review: modification-based, user interface-based, model-based, mutation-based, and similarity-based. The discovery shows that the taxonomy is still progressing and further explored by researchers. In terms of the technique for TCP implementation, we discover that there is no dominant algorithm used. Researchers have their reasons for choosing a technique to implement TCP based on specific analysis and motivations. However, it is noticeable that the Greedy and Additional-Greedy Algorithm, which are classic search-based TCP algorithms, are still in demand. Likewise, Genetic Algorithms are also popular. On the other hand, we discover that Machine Learning Algorithms seem to be growing in use and gaining popularity.

As for RQ2, we have investigated that SIR is an SUT repository database with an excellent and complete SUT collection. However, although many researchers still use it, some studies use SUT for empirical studies. Through the review, we classify SUTs into six categories: SUTs built from open-source software, SUTs from repositories such as SIR and Defects, SUTs from industrial cases, SUTs constructed by researchers.

Answering RQ3, in terms of using metrics to measure the effectiveness of TCP, we did not find any significant progress. Some researchers add aspects of measurement on a fixed basis to the APFD. In this case, the APFD is dominantly used.

Finally, on RQ4, referring to reviews carried out by previous surveys, we found the overall development on the use of requirements-based TCP. Although requirements-based TCP is not as popular as many other approaches, it can be further developed. For example, referring to our survey, we found that requirement risk TCP introduced by [77], requirement dependency TCP by [71] and [68]. It is proven that the research conducted has allowed growing further and improving the effectiveness. As stated by [22], one of the advantages of requirement-based TCP is the privilege of utilizing information from requirements. Therefore, the preparation of test cases can be done earlier to save time in the testing process.

VI. Conclusion

This paper presents the SLR result on test case prioritization for regression testing. The study's main objective was to get the state of the art on TCP for regression testing in 2017-2020. Furthermore, this study also investigates the TCP explicitly based on requirements since the TCP was first introduced until 2020.

We found more TCP approaches not mentioned in other surveys, which have opportunities for further research. The new TCP approaches are modification-based, user interface-based, model-based, mutation-based, and similarity-based. In the use of SUT, it appears that there are more diverse variations of SUT. Even so, the utilization of SUT repositories such as SIR and Defect4J is still in great demand. We also discover that APFD is still a very dominant metric, and almost no specific new metrics are found.

For future work, it is beneficial to explore the utilization of requirements to improve TCP effectiveness. We view that requirement-based utilization will help prepare test cases earlier, so the testing process can be more efficient. Requirement risk is an essential aspect of a requirement that is considered in the test case prioritization development. Meanwhile, the dependency between requirements is a crucial issue to consider in software development, so it can be one of the factors for prioritizing test cases. We cannot ignore the relationship between requirements in the software development process, including in the testing stage. Requirement-based TCP still has many opportunities for improvement. We believe that there are many attributes in requirements that can improve TCP effectiveness.

VII. Acknowledgment

Thank you to Universiti Teknikal Malaysia Melaka for providing supports and resources for this research.

References

[1] Pradhan, S. Wang, S. Ali, T. Yue, and M. Liaaen, “Employing rule mining and multi-objective search for dynamic test case prioritization,” J. Syst. Softw., vol. 153, pp. 86–104, 2019, doi: 10.1016/j.jss.2019.03.064.

[2] [2]R. H. Rosero, O. S. Gómez, and G. Rodríguez, “15 Years of Software Regression Testing Techniques - A Survey,” Int. J. Softw. Eng. Knowl. Eng., vol. 26, no. 5, pp. 675–689, 2016, doi: 10.1142/S0218194016300013.

[3] Sebastian Ulewicz and Birgit Vogel-Heuser, “Industrially Applicable System Regression Test Prioritization in Production Automation,” 2 IEEE Trans. Autom. Sci. Eng., pp. 1545–5955, 2018.

[4] S. Souto and M. d’Amorim, “Time-space efficient regression testing for configurable systems,” J. Syst. Softw., vol. 137, pp. 733–746, 2018, doi: 10.1016/j.jss.2017.08.010.

[5] H. Hemmati, Advances in Techniques for Test Prioritization, 1st ed., vol. 112. Elsevier Inc., 2019.

[6] D. Hao, L. Zhang, and H. Mei, “Test-case prioritization : achievements and challenges,” pp. 1–9, 2016.

[7] M. J. Arafeen and H. Do, “Test case prioritization using requirements-based clustering,” Proc. - IEEE 6th Int. Conf. Softw. Testing, Verif. Validation, ICST 2013, pp. 312–321, 2013, doi: 10.1109/ICST.2013.12.

[8] H. Srikanth, C. Hettiarachchi, and H. Do, “Requirements based test prioritization using risk factors: An industrial study,” Inf. Softw. Technol., vol. 69, pp. 71–83, 2016, doi: 10.1016/j.infsof.2015.09.002.

[9] J. Srikanth, H, Williams, L, Osborne, “System Test Case Prioritization of New and Regression Test Cases,” in IEEEE International Symposium on Empirical Study, 2005, pp. 64–73.

[10] Y. Wang, Z. Zhu, B. Yang, F. Guo, and H. Yu, “Using reliability risk analysis to prioritize test cases,” J. Syst. Softw., vol. 139, pp. 14–31, 2018, doi: 10.1016/j.jss.2018.01.033.

[11] R. Krishnamoorthi and S. A. Sahaaya Arul Mary, “Requirement based system test case prioritization of new and regression test cases,” Int. J. Softw. Eng. Knowl. Eng., vol. 19, no. 3, pp. 453–475, 2009, doi: 10.1142/S0218194009004222.

[12] V. Garousi, R. Özkan, and A. Betin-Can, “Multi-objective regression test selection in practice: An empirical study in the defense software industry,” Inf. Softw. Technol., vol. 103, pp. 40–54, 2018, doi: 10.1016/j.infsof.2018.06.007.

[13] S. Ulewicz, B. Vogel-heuser, and S. Member, “Industrially Applicable System Regression Test Prioritization in Production Automation,” IEEE Trans. Autom. Sci. Eng., pp. 1–13, 2018.

[14] G. Rothermel, R. H. Untcn, C. Chu, and M. J. Harrold, “Prioritizing test cases for regression testing,” IEEE Trans. Softw. Eng., vol. 27, no. 10, pp. 929–948, 2001, doi: 10.1109/32.962562.

[15] A. Zarrad, “A Systematic Review on Regression Testing for Web-Based Applications,” J. Softw., vol. 10, no. 8, pp. 971–990, 2015, doi: 10.17706/jsw.10.8.971-990.

[16] H. Do, "Recent Advances in Regression Testing Techniques," Adv. Comput., vol. 103, pp. 53–77, 2016, doi: 10.1016/bs.adcom.2016.04.004.

[17] N. Sharma, Sujata, and G. N. Purohit, "Test case prioritization techniques 'an empirical study,'" 2014 Int. Conf. High Perform. Comput. Appl. ICHPCA 2014, 2015, doi: 10.1109/ICHPCA.2014.7045344.

[18] P. E. Strandberg, D. Sundmark, W. Afzal, T. J. Ostrand, and E. J. Weyuker, "Experience Report: Automated System Level Regression Test Prioritization Using Multiple Factors," Proc. - Int. Symp. Softw. Reliab. Eng. ISSRE, no. October, pp. 12–23, 2016, doi: 10.1109/ISSRE.2016.23.

[19] H. S. De Campos, C. A. De Paiva, R. Braga, M. A. P. Araujo, J. M. N. David, and F. Campos, "Regression tests provenance data in the continuous so ware engineering context," ACM Int. Conf. Proceeding Ser., vol. Part F1306, no. September 2018, 2017, doi: 10.1145/3128473.3128483.

[20] J. Anderson, M. Azizi, S. Salem, and H. Do, "On the use of usage patterns from telemetry data for test case prioritization," Inf. Softw. Technol., vol. 113, pp. 110–130, 2019, doi: 10.1016/j.infsof.2019.05.008.

[21] S. Yoo and M. Harman, "Regression testing minimization, selection and prioritization: A survey," Softw. Test. Verif. Reliab., vol. 22, no. 2, pp. 67–120, 2012, doi: 10.1002/stv.430.

[22] M. Khatibsyarbini, M. A. Isa, D. N. A. Jawawi, and R. Tumeng, "Test case prioritization approaches in regression testing: A systematic literature review," Inf. Softw. Technol., vol. 93, no. June 2018, pp. 74–93, 2018, doi: 10.1016/j.infsof.2017.08.014.

[23] Y. Singh, A. Kaur, B. Suri, and S. Singhal, "Systematic literature review on regression test prioritization techniques," Inform., vol. 36, no. 4, pp. 379–408, 2012, doi: 10.31449/inf.v36i4.420.

[24] H. S. De Campos Junior, M. A. P. Arajo, J. M. N. David, R. Braga, F. Campos, and V. Ströele, "Test case prioritization: A systematic review and mapping of the literature," ACM Int. Conf. Proceeding Ser., no. August 2018, pp. 34–43, 2017, doi: 10.1145/3131151.3131170.

[25] M. L. Mohd Shafie and W. M. N. Wan Kadir, "Model-based test case prioritization: A systematic literature review," J. Theor. Appl. Inf. Technol., vol. 96, no. 14, pp. 4548–4573, 2018.

[26] R. Mukherjee and K. S. Patnaik, "A survey on different approaches for software test case prioritization," J. King Saud Univ. - Comput. Inf. Sci., 2018, doi: 10.1016/j.jksuci.2018.09.005.

[27] M. D. C. De Castro-Cabrera, A. García-Dominguez, and I. Medina-Bulo, "Trends in prioritization of test cases: 2017-2019," Proc. ACM Symp. Appl. Comput., pp. 2005–2011, 2020, doi: 10.1145/3341105.3374036.

[28] A. Samad, H. Mahdin, R. Kazmi, and R. Ibrahim, "Regression Test Case Prioritization: A Systematic Literature Review," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 2, pp. 655–663, 2021, doi: 10.14569/IJACSA.2021.0120282.

[29] M. Hasnain, M. F. Pasha, I. Ghani, and S. R. Jeong, Functional Requirement-Based Test Case Prioritization in Regression Testing: A Systematic Literature Review, vol. 2, no. 6. Springer Singapore, 2021.

[30] Y. Lou, J. Chen, L. Zhang, and D. Hao, A Survey on Regression Test-Case Prioritization, 1st ed., vol. 113. Elsevier Inc., 2019.

[31] D. Qiu, B. Li, S. Ji, and H. Leung, "Regression testing of web service: A systematic mapping study," ACM Comput. Surv., vol. 47, no. 2, 2014, doi: 10.1145/2631685.

[32] M. Hasnain, I. Ghani, M. F. Pasha, C. H. Lim, and S. R. Jeong, "A Comprehensive Review on Regression Test Case Prioritization Techniques for Web Services," KSII Trans. Internet Inf. Syst., vol. 14, no. 5, pp. 1861–1885, 2020, doi: 10.3837/tiis.2020.05.001.

[33] J. A. Prado Lima and S. R. Vergilio, "Test Case Prioritization in Continuous Integration environments: A systematic mapping study," Inf. Softw. Technol., vol. 121, p. 106268, 2020, doi: 10.1016/j.infsof.2020.106268.

[34] N. bin Ali et al., On the search for industry-relevant regression testing research, vol. 24, no. 4. 2019.

[35] B. Kitchenham and P. Brereton, "A systematic review of systematic review process research in software engineering," Inf. Softw. Technol., vol. 55, no. 12, pp. 2049–2075, 2013, doi: 10.1016/j.infsof.2013.07.010.

[36] H. Jahan, Z. Feng, and S. M. H. Mahmud, "Risk-Based Test Case Prioritization by Correlating System Methods and Their Associated Risks," Arab. J. Sci. Eng., vol. 45, no. 8, pp. 6125–6138, 2020, doi: 10.1007/s13369-020-04472-z.

[37] R. Wang, Z. Li, S. Jiang, and C. Tao, "Regression Test Case Prioritization Based on Fixed Size Candidate Set ART Algorithm," Int. J. Softw. Eng. Knowl. Eng., vol. 30, no. 3, pp. 291–320, 2020, doi: 10.1142/S0218194020500138.

[38] W. Su, Z. Li, Z. Wang, and D. Yang, "A Meta-heuristic Test Case Prioritization Method Based on Hybrid Model," Proc. - 2020 Int. Conf. Comput. Eng. Appl. ICCEA 2020, pp. 430–435, 2020, doi: 10.1109/ICCEA50009.2020.00099.

[39] S. Mondal and R. Nasre, "Hansie: Hybrid and consensus regression test prioritization," J. Syst. Softw., vol. 172, no. October, 2021, doi: 10.1016/j.jss.2020.110850.

[40] J. Chi et al., "Relation-based test case prioritization for regression testing," J. Syst. Softw., vol. 163, 2020, doi: 10.1016/j.jss.2020.110539.

[41] M. Khanna, N. Chauhan, and D. K. Sharma, "Search for prioritized test cases during web application testing," Int. J. Appl. Metaheuristic Comput., vol. 10, no. 2, pp. 1–26, 2019, doi: 10.4018/IJAMC.2019040101.

[42] A. Arrieta, S. Wang, U. Markiegi, G. Sagardui, and L. Etxeberria, "Employing Multi-Objective Search to Enhance Reactive Test Case Generation and Prioritization for Testing Industrial Cyber-Physical Systems," IEEE Trans. Ind. Informatics, vol. 14, no. 3, pp. 1055–1066, 2018, doi: 10.1109/TII.2017.2788019.

[43] M. Azizi and H. Do, "Graphite: A Greedy Graph-Based Technique for Regression Test Case Prioritization," in 2018 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), 2018, pp. 245–251, doi: 10.1109/ISSREW.2018.00014.

[44] M. Mahdieh, S. H. Mirian-Hosseinabadi, K. Etemadi, A. Nosrati, and S. Jalali, "Incorporating fault-proneness estimations into coverage-based test case prioritization methods," Inf. Softw. Technol., vol. 121, no. January, p. 106269, 2020, doi: 10.1016/j.infsof.2020.106269.

[45] M. Khanna, A. Chaudhary, A. Toofani, and A. Pawar, "Performance Comparison of Multi-objective Algorithms for Test Case Prioritization During Web Application Testing," Arab. J. Sci. Eng., vol. 44, no. 11, pp. 9599–9625, 2019, doi: 10.1007/s13369-019-03817-7.

[46] I. Hajri, A. Goknil, F. Pastore, and L. C. Briand, "Automating system test case classification and prioritization for use case-driven testing in product lines," Empir. Softw. Eng., vol. 25, no. 5, pp. 3711–3769, 2020, doi: 10.1007/s10664-020-09853-4.

[47] S. Nayak, C. Kumar, and S. Tripathi, "Enhancing Efficiency of the Test Case Prioritization Technique by Improving the Rate of Fault Detection," Arab. J. Sci. Eng., vol. 42, no. 8, pp. 3307–3323, 2017, doi: 10.1007/s13369-017-2466-6.

[48] M. M. Öztürk, "A bat-inspired algorithm for prioritizing test cases," Vietnam J. Comput. Sci., 2017, doi: 10.1007/s40595-017-0100-x.

[49] K. W. Shin and D. J. Lim, "Model-based test case prioritization using an alternating variable method for regression testing of a UML-based model," Appl. Sci., vol. 10, no. 21, pp. 1–23, 2020, doi: 10.3390/app10217537.

[50] J. F. S. Ouriques, E. G. Cartaxo, and P. D. L. Machado, "Test case prioritization techniques for model-based testing: a replicated study," Softw. Qual. J., vol. 26, no. 4, pp. 1451–1482, 2018, doi: 10.1007/s11219-017-9398-y.

[51] T. Zhang, X. Wang, D. Wei, and J. Fang, "Test Case Prioritization Technique Based on Error Probability and Severity of UML Models," vol. 28, no. 6, pp. 831–844, 2018, doi: 10.1142/S0218194018500249.

[52] Y. Venugopal, P. Quang-Ngoc, and L. Eunseok, "Modification point aware test prioritization and sampling to improve patch validation in automatic program repair," Appl. Sci., vol. 10, no. 5, pp. 1–14, 2020, doi: 10.3390/app10051593.

[53] W. Fu, H. Yu, G. Fan, X. Ji, and X. Pei, "A Regression Test Case Prioritization Algorithm Based on Program Changes and Method Invocation Relationship," Proc. - Asia-Pacific Softw. Eng. Conf.

APSEC, vol. 2017-Decem, pp. 169–178, 2018, doi: 10.1109/APSEC.2017.23.

[54] H. Wang, M. Yang, L. Jiang, J. Xing, Q. Yang, and F. Yan, "Test Case Prioritization for Service-Oriented Workflow Applications: A Perspective of Modification Impact Analysis," IEEE Access, vol. 8, pp. 101260–101273, 2020, doi: 10.1109/ACCESS.2020.2998545.

[55] M. K. Pachariya, "Building Ant System for Multi-Faceted Test Case Prioritization: An Empirical Study," Int. J. Softw. Innov., vol. 8, no. 2, pp. 23–37, 2020, doi: 10.4018/IJSI.2020040102.

[56] R. Huang, Q. Zhang, D. Towey, W. Sun, and J. Chen, "Regression test case prioritization by code combinations coverage," J. Syst. Softw., vol. 169, p. 110712, 2020, doi: 10.1016/j.jss.2020.110712.

[57] R. Huang et al., "Abstract Test Case Prioritization Using Repeated Small-Strength Level-Combination Coverage," IEEE Trans. Reliab., vol. 69, no. 1, pp. 349–372, 2020, doi: 10.1109/TR.2019.2908068.

[58] M. Azizi, "A Collaborative Filtering Recommender System for Test Case Prioritization in Web Applications," pp. 1560–1567, 2018.

[59] D. Di Nucci, A. Panichella, A. Zaidman, and A. De Lucia, "A Test Case Prioritization Genetic Algorithm Guided by the Hypervolume Indicator," IEEE Trans. Softw. Eng., vol. 46, no. 6, pp. 674–696, 2020, doi: 10.1109/TSE.2018.2868082.

[60] R. Matinnejad, S. Nejati, L. C. Briand, and T. Bruckmann, "Test Generation and Test Prioritization for Simulink Models with Dynamic Behavior," IEEE Trans. Softw. Eng., vol. 45, no. 9, pp. 919–944, 2019, doi: 10.1109/TSE.2018.2811489.

[61] M. Khatibsyarbini, M. A. Isa, and D. N. A. Jawawi, "A hybrid weight-based and string distances using particle swarm optimization for prioritizing test cases," J. Theor. Appl. Inf. Technol., vol. 95, no. 12, pp. 2723–2732, 2017.

[62] M. Khatibsyarbini, M. A. Isa, D. N. A. Jawawi, H. N. A. Hamed, and M. D. Mohamed Suffian, "Test Case Prioritization Using Firefly Algorithm for Software Testing," IEEE Access, vol. 7, pp. 132360–132373, 2019, doi: 10.1109/access.2019.2940620.

[63] X. Lei, M. Huaikou, Z. Weiewei, and C. Shaojun, "An Empirical Study on Clustering Approach Combining Fault Prediction for Test Case Prioritization," in International Conference on Information Systems, 2017, pp. 815–820, doi: 10.1109/ICIS.2017.7960105.

[64] R. Huang, Y. Zhou, W. Zong, D. Towey, and J. Chen, "An Empirical Comparison of Similarity Measures for Abstract Test Case Prioritization," pp. 3–12, 2017, doi: 10.1109/COMPSAC.2017.271.

[65] B. Miranda, E. Cruciani, R. Verdecchia, and A. Bertolino, "FAST approaches to scalable similarity-based test case prioritization," Proc. - Int. Conf. Softw. Eng., vol. 2018-Janua, pp. 222–232, 2018, doi: 10.1145/3180155.3180210.

[66] S. A. Halim, D. N. A. Jawawi, and M. Sahak, "Similarity distance measure and prioritization algorithm for test case prioritization in software product line testing," J. Inf. Commun. Technol., vol. 18, no. 1, pp. 57–75, 2019, doi: 10.32890/jict2019.18.1.4.

[67] A. D. Shrivathsan et al., "Novel fuzzy clustering methods for test case prioritization in Software Projects," Symmetry (Basel)., vol. 11, no. 11, pp. 1–22, 2019, doi: 10.3390/sym11111400.

[68] M. Abbas, I. Inayat, M. Saadatmand, and N. Jan, "Requirements Dependencies-Based Test Case Prioritization for Extra-Functional Properties," in 2019 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), 2019, pp. 159–163, doi: 10.1109/ICSTW.2019.00045.

[69] R. Butool, A. Nadeem, M. Sindhu, and O. u. Zaman, "Improving Requirements Coverage in Test Case Prioritization for Regression Testing," in 2019 22nd International Multitopic Conference (INMIC), 2019, pp. 1–6, doi: 10.1109/INMIC48123.2019.9022761.

[70] W. Zhang, Y. Qi, X. Zhang, B. Wei, M. Zhang, and Z. Dou, "On Test Case Prioritization Using Ant Colony Optimization Algorithm," in 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2019, pp. 2767–2773, doi: 10.1109/HPCC/SmartCity/DSS.2019.00388.

[71] A. Vescan, C. Şerban, C. Chisăliţă-Creţu, and L. Dioşan, "Requirement dependencies-based formal approach for test case prioritization in regression testing," Proc. - 2017 IEEE 13th Int. Conf. Intell. Comput. Commun. Process. ICCP 2017, no. September 2017, pp. 181–188, 2017, doi: 10.1109/ICCP.2017.8117002.

[72] Z. Yu, F. Fahid, T. Menzies, G. Rothermel, K. Patrick, and S. Cherian, "TERMINATOR: Better automated UI test case prioritization," ESEC/FSE 2019 - Proc. 2019 27th ACM Jt. Meet. Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng., pp. 883–894, 2019, doi: 10.1145/3338906.3340448.

[73] M. Abdur, M. Abu, and M. Saeed, "Prioritizing Dissimilar Test Cases in Regression Testing using Historical Failure Data," Int. J. Comput. Appl., vol. 180, no. 14, pp. 1–8, 2018, doi: 10.5120/ijca2018916258.

[74] Y. Yang, X. Huang, X. Hao, Z. Liu, and Z. Chen, "An Industrial Study of Natural Language Processing Based Test Case Prioritization," in 2017 IEEE International Conference on Software Testing, Verification and Validation (ICST), 2017, pp. 548–549, doi: 10.1109/ICST.2017.66.

[75] D. Shin, S. Yoo, M. Papadakis, and D. H. Bae, "Empirical evaluation of mutation-based test case prioritization techniques," Softw. Test. Verif. Reliab., vol. 29, no. 1–2, pp. 1–28, 2019, doi: 10.1002/stvr.1695.

[76] D. B. Mishra, R. Mishra, A. A. Acharya, and K. N. Das, Test case optimization and prioritization based on multi-objective genetic algorithm, vol. 741. Springer Singapore, 2019.

[77] C. Hettiarachchi and H. Do, "A Systematic Requirements and Risks-Based Test Case Prioritization Using a Fuzzy Expert System," Proc. - 19th IEEE Int. Conf. Softw. Qual. Reliab. Secur. QRS 2019, pp. 374–385, 2019, doi: 10.1109/QRS.2019.00054.

[78] M. Khatibsyarbini, M. A. Isa, D. N. A. Jawawi, H. N. A. Hamed, and M. D. M. Suffian, "Test Case Prioritization Using Firefly Algorithm for Software Testing," IEEE Access, vol. 7, pp. 132360–132373, 2019, doi: 10.1109/ACCESS.2019.2940620.

[79] B. Miranda, E. Cruciani, R. Verdecchia, and A. Bertolino, "FAST Approaches to Scalable Similarity-Based Test Case Prioritization," in 2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE), 2018, pp. 222–232, doi: 10.1145/3180155.3180210.

[80] Z. Yu and T. Menzies, "TERMINATOR : Better Automated UI Test Case Prioritization."

[81] D. Kumar and Y. Sandip, "Regression test case selection and prioritization for object oriented software," Microsyst. Technol., vol. 7, 2019, doi: 10.1007/s00542-019-04679-7.

[82] S. C. Lei Xiao, Huaikou, Weiwei Zhuang, "An Empirical Study on Clustering Approach Combining Fault Prediction for Test Case Prioritization," in 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), 2017, pp. 815–820, doi: 10.1109/ICIS.2017.7960105.

[83] A. D. Shrivathsan, K. S. Ravichandran, R. Krishankumar, V. Sangeetha, and S. Kar, "Novel Fuzzy Clustering Methods for Test Case Prioritization in Software Projects," pp. 1–22.

[84] C. Hettiarachchi and H. Do, "A Systematic Requirements and Risks-Based Test Case Prioritization Using a Fuzzy Expert System," in 2019 IEEE 19th International Conference on Software Quality, Reliability and Security (QRS), 2019, pp. 374–385, doi: 10.1109/QRS.2019.00054.

[85] H. Do, S. Elbaum, and G. Rothermel, "Supporting controlled experimentation with testing techniques: An infrastructure and its potential impact," Empir. Softw. Eng., vol. 10, no. 4, pp. 405–435, 2005, doi: 10.1007/s10664-005-3861-2.

[86] H. Wang, M. Yang, L. Jiang, J. Xing, Q. Yang, and F. Yan, "Test Case Prioritization for Service-Oriented Workflow Applications: A Perspective of Modification Impact Analysis," IEEE Access, vol. 8, pp. 101260–101273, 2020, doi: 10.1109/ACCESS.2020.2998545.

[87] D. S. Silva, R. Rabelo, P. S. Neto, R. Britto, and P. A. Oliveira, "A test case prioritization approach based on software component metrics," Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern., vol. 2019-Octob, no. August, pp. 2939–2945, 2019, doi: 10.1109/SMC.2019.8914670.

[88] D. K. Yadav and S. Dutta, "Regression test case selection and prioritization for object oriented software," Microsyst. Technol., vol. 26, no. 5, pp. 1463–1477, 2020, doi: 10.1007/s00542-019-04679-7.

[89] M. Yoon, E. Lee, M. Song, and B. Choi, "A Test Case Prioritization through Correlation of Requirement and Risk," J. Softw. Eng. Appl., vol. 05, no. 10, pp. 823–835, 2012, doi: 10.4236/jsea.2012.510095.

[90] Arafeen and H. Do, "Test Case Prioritization using Requirement-Based Clustering," in 2013 IEEE Sixth International Conference on Software Testing, Verification and Validation, 2013, pp. 488–492, doi: 10.1109/ICST.2013.12.

[91] C. Hettiarachchi, H. Do, and B. Choi, "Effective regression testing using requirements and risks," Proc. - 8th Int. Conf. Softw. Secur. Reliab. SERE 2014, pp. 157–166, 2014, doi: 10.1109/SERE.2014.29.

[92] C. Hettiarachchi, H. Do, and B. Choi, "Risk-based test case prioritization using a fuzzy expert system," Inf. Softw. Technol., vol. 69, pp. 1–15, 2016, doi: 10.1016/j.infsof.2015.08.008.

[93] M. Felderer, C. Haisjackl, V. Pekar, and R. Breu, "A risk assessment framework for software testing," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8803 Lever, pp. 292–308, 2014, doi: 10.1007/978-3-662-45231-8_21.

[94] T. Ma, H. Zeng, and X. Wang, "Test case prioritization based on requirement correlations," 2016 IEEE/ACIS 17th Int. Conf. Softw. Eng. Artif. Intell. Netw. Parallel/Distributed Comput. SNPD 2016, no. 61170044, pp. 419–424, 2016, doi: 10.1109/SNPD.2016.7515934.

[95] W. Zhang, Y. Qi, X. Zhang, B. Wei, M. Zhang, and Z. Dou, "On test case prioritization using ant colony optimization algorithm," Proc. - 21st IEEE Int. Conf. High Perform. Comput. Commun. 17th IEEE Int. Conf. Smart City 5th IEEE Int. Conf. Data Sci. Syst. HPCC/SmartCity/DSS 2019, pp. 2767–2773, 2019, doi: 10.1109/HPCC/SmartCity/DSS.2019.00388.

APPENDIX

THE SELECTED ARTICLES

Study	Publication type	Publication year	Publisher
[36]	Journal	2020	Arabian Journal for Science and Engineering
[37]	Journal	2020	International Journal of Software Engineering and Knowledge Engineering
[38]	Conference paper	2020	International Conference on Computer Engineering and Application (ICCEA)
[39]	Journal	2020	Journal of Systems and Software
[44]	Journal	2020	Information and Software Technology
[49]	Journal	2020	Applied Sciences Multidisciplinary Digital Publishing Institute (MDPI)
[52]	Journal	2020	Applied Sciences Multidisciplinary Digital Publishing Institute (MDPI)
[55]	Journal	2020	International Journal of Software Innovation
[56]	Journal	2020	The Journal of Systems & Software
[40]	Journal	2020	The Journal of Systems and Software
[57]	Journal	2020	Journal of LaTeX Class Files
[54]	Journal	2020	IEEE Access
[78]	Journal	2020	IEEE Access
[68]	Conference paper	2019	IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)
[72]	Conference paper	2019	Association of Computing Machinery (ACM)
[76]	Conference Paper	2019	International Conference on Harmony Search Algorithm (ICHSA)
[20]	Journal	2019	Information and Software Technology
[69]	Conference paper	2019	IEEE 2019 22nd International Multitopic Conference (INMIC)
[77]	Conference paper	2019	IEEE 19th International Conference on Software Quality, Reliability and Security (QRS)
[1]	Journal	2019	The Journal of Systems and Software
[83]	Journal	2019	Symmetry Multidisciplinary Digital Publishing Institute (MDPI)
[81]	Journal	2019	Microsystems Technologies
[66]	Journal	2019	Journal of Information and Communication Technology
[95]	Conference paper	2019	IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th. International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems
[45]	Journal	2019	Arabian Journal for Science and Engineering
[87]	Conference paper	2019	IEEE International Conference on Systems, Man and Cybernetics (SMC)
[46]	Journal	2019	Empirical Software Engineering
[41]	Journal	2019	International Journal of Applied Metaheuristic Computing
[58]	Symposium paper	2018	Symposium on Applied Computing
[51]	Journal	2018	International Journal of Software Engineering
[50]	Journal	2018	Software Quality Journal
[10]	Journal	2018	The Journal of Systems and Software

[59]	Symposium paper	2018	Symposium on Search-Based Software Engineering 2015
[79]	Conference paper	2018	IEEE 40th International Conference on Software Engineering
[43]	Conference paper	2018	IEEE International Symposium on Software Reliability Engineering Workshops
[13]	Journal	2018	IEEE Transactions on Automation Science And Engineering
[73]	Journal	2018	International Journal of Computer Applications
[75]	Journal	2018	Software Testing, Verification and Reliability
[60]	Journal	2018	IEEE Transactions on Software Engineering
[82]	Conference paper	2017	International Conference on Information Systems
[71]	Conference paper	2017	IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)
[47]	Journal	2017	Arabian Journal for Science and Engineering
[48]	Journal	2017	Vietnam Journal of Computer Science
[42]	Journal	2017	IEEE Transactions on Industrial Informatics
[61]	Journal	2017	Journal of Theoretical and Applied Information Technology
[53]	Conference paper	2017	Asia-Pacific Software Engineering Conference
[74]	Conference paper	2017	IEEE International Conference on Software Testing, Verification and Validation
[64]	Conference paper	2017	IEEE 41st Annual Computer Software and Applications Conference

SNR based Energy Efficient Communication Protocol for Emergency Applications in WBAN

K. Viswavardhan Reddy[1]
Electronics Engineering, Jain University
Bengaluru, India

Navin Kumar[2]
Dept. of ECE, Amrita School of Engineering, AVV
Bengaluru, India

***Abstract*—Continuous remote monitoring of a patient's health condition in dynamic environment imposes many challenges. Challenges further get multiplied based on the size of body area sensor network. One such challenge is energy efficiency of sensors. Maintaining longer life of all nodes, especially who participate in communicating vital signals from one network to another towards the base station is very important. In this work, an energy efficient communication protocol for the wireless body area network (WBAN) is proposed. The essential characteristics of the protocol are: random deployment of nodes, formation of clusters, node with high signal to noise ratio (SNR) as cluster head (CH), random rotation of CHs within each cluster, and so on. The developed algorithm is simulated in MATLAB by varying the number of nodes and networks. Obtained results are compared with some of the recent and most relevant existing works. It is found that there is an enhancement in the network lifetime by 19.5%, throughput by 12.61% and average remaining energy by 57.21%.**

Keywords*—WBAN; energy efficiency; emergency applications; protocol; remote monitoring

I. Introduction

The usage of information and communication technology (ICT) in healthcare sector is not so widely visible even in the pandemic. Enormous challenges in demand of healthcare centre were encountered [1]. People even in urban areas require doctor appointment and physically visit them if they have any health problem. This has been in practice for a long time and continues even now. Significant research, what we call IOTization, has been around in almost every industry such as smart city, smart home, smart agriculture, smart class room, smart governance, smart and intelligent transport, smart eHealth [2-4]. However, to some extent, smartness is observed in the case of home, in the case of manufacturing but smartness in the case of eHealth is still a miles away. It is limited to a very few individuals who can somehow monitors some of his/her body parameters using smart watch, band, and so on. During this pandemic situation, when social distancing has become very important, health industry should have seen significant growth in terms of usage of monitoring devices and techniques. Unfortunately, we are far behind.

Though, there are many tiny, cheap and smart sensors modules available which can be kept on the body of human within and outside or vicinity of body of human. This type of network is well known as body area network (BAN). Furthermore, this BAN is expected to be dynamic and mobile, such system is known as wireless BAN (WBAN). WBAN allows monitoring of health data of a patient or person from remote. The usage of such networks at this time period could have been significantly increased. Despite that a large number of papers, investigations are available on the subject [5], but practical implementation and usage is not promising due to requirements of bandwidth, high storage capabilities and high-power usage. Additionally, huge information transformation and estimation is required at the edge or cloud. Could computing [6] can be used with regular transformation and data analysis. This inquiry can then be used by the clinicians for providing better treatment of patient in healthcare as well as research. In addition, remote monitoring, consultation, counselling can be easily done. Furthermore, when patient is mobile and dynamic, the system complexity and requirement would be different. One of the important problems would be communicating information from one network to other networks towards the patient's parent cloud, that is, routing of information from the source to destination.

WBAN routing protocols are classified into energy aware routing, cross-layer routing, temperature-based routing, cluster-based routing, posture-based routing and quality of service (QoS)-based routing [7]. An opportunistic power-efficient routing with load balancing (OE2-LB) by eliminating the delay caused during the aggregation process algorithm has been proposed in [8]. It helps in avoiding the loops that occur in routing in a more effective way. Authors claimed to have developed a better algorithm with respect to throughput delay, aggregation time, energy, and live nodes count. A power efficient communication protocol for transmitting the data more reliably is proposed by selecting appropriate next hop node [9]. In this, to select next hop node, a maximum benefit function has been defined. It uses parameters like residual power, bandwidth, efficiency in transmission and number of hops to sink. The performance of the proposed protocol is simulated in MATLAB and evaluated with PERA and NEW-ATTEMPT protocols. In [10], a clustering routing protocol for WBANs (CRPBA) is developed for maximizing the network lifetime and minimization of power dissipation for the nodes. The performance of the developed algorithm is evaluated with specifications such as total number of nodes 24; initial energy of each node is 0. 5J and total number of 7000 rounds. The reported results showed the first node death at 3375 rounds. In [11], authors developed iMSIMPLE: improved stable increased-throughput multi-hop link adept communication protocol. A cost function is defined by considering various parameters such as distance to sink and residual power has been used for selecting a new forwarding node. Simulations

were performed with eight nodes having 0.5j of initial energy and radio parameters like DC current for transmitter and receiver, supply voltage. Results showed that the first node death occurred at 5200 rounds with no mobility. Also, it is learnt that when a patient is static i.e., fixed in location, receive signal strength indicator (RSSI) is normally considered [12]. But when the patient is moving, there will be considerable number of noises as channel keeps changing. Hence, signal to noise ratio (SNR) is more appropriate than the RSSI for variable radio conditions.

Hence in this work, we investigated SNR based power adept communication protocol for not only maximizing the network period by reducing the power dissipation, but also increasing the network throughput and reducing delay in the network. For this, we introduced a weightage function for selecting the cluster head. The weightage function is based on two parameters; high SNR among the nodes and lowest distance from the nodes to cluster head. Also, we compared the performance of our developed algorithm with energy efficient low power robust clustering hierarchy (EELEACH), distributed power efficient clustering (DEEC), threshold sensitive power efficient sensor network protocol (TEEN) and clustering based routing protocol for WBAN (CRPBA).

The contents of this paper have the following details. Section II discusses about related works of various routing protocols with respect to performance metrics, goals, and the cost function that has been developed. Simulation parameters with the radio model and energy model of two modes of communication are discussed in Section III. Section IV deals with proposed algorithm with defined cost function. Finally, results and conclusion were discussed in Sections V and VI, respectively.

II. Related Work

Optimized cost effective and energy efficient routing processes (OCER) and Extended-OCER are suggested in [13]. So, as to maximize the network period with the lowest power usage, authors projected a price function with the node's residual energy, path loss, and reliability of link. Nodes owning a less values of price function hold the risks of getting subsequent hop nodes. The functionality of the projected protocol is evaluated using simulations with metrics such as power consumption, throughput and number of packets forwarded. Furthermore, comparison has been made with EPR for indoor medical centres and DMQoS protocols. The outcomes indicate OCER achieves excellent power savings.

When it comes to RK efficient routing protocol [14] has been recommended by the researchers with eight nodes deploying on the human body to keep track of physiological parameters. Nodes are actually split into normal and critical nodes, and that makes use of one hop mechanism and multi hopping respectively. A cost function is identified to pick a forwarder node for multihopping, mostly, based on the least distance to sink and the node's residual power. The functionality of the protocol is then compared and contrasted to ATTEMPT and results suggested that the recommended one outperforms ATTEMPT.

Anwar et. al. developed power aware link effective routing for WBANs (ELR-W) in [15]. Link efficient network model was built to select next hop node with minimum distance to base station and quality link (having very good packet reception ratio). To analyze the efficiency of protocol, parameters such as usage of power, life time of the network, and throughput had been considered for simulations. Very low energy usage, low packet loss i.e., high throughput, and substantial network lifetime were achieved with the ELR-W when in contrast with M-ATTEMPT and iM-SIMPLE.

To have efficient routing approach in WBANs, achieving QoS (throughput, power effectiveness, end-to-end hold off, as well as packet transmission rates) is really very important. As a result, a SDN-enabled and energy-efficient routing algorithm (ESR-W) developed with the usage of the Fuzzy-based Dijkstra method [16] for achieving QoS. So, riverbed modeler simulation software with IEEE 802.15.6 for intra-WBAN, and hubs flow interface protocol for inter-WBAN (SD-WBAN) are used to evaluate the performance of the proposed approach. Simulation results of ESR-W found to be more efficient when compared with AODV, and SDN routing developed for SD-WBAN architecture.

Link-aware and energy efficient pattern for body region networks (LAEEBA) [17] is suggested by the authors to achieve reliable, pathloss efficient and high throughput for WBANs. Besides, on the foundation of higher residual vitality as well as bare minimum distance coming from nodes to BS, a price feature continues to be recommended for choosing the forwarder nodes within the community. The simulations were carried using MATLAB and performance of the LAEEBA is compared with SIMPLE and M-ATTEMPT. A total of 8 nodes with fixed locations, initial energy of 0.5 J with a range of 10 meters is deployed on the body. The results show that LAEEBA protocol improves the stability of network and network life time with first node dying at 5130 rounds when compared with 2147 and 4436 rounds for M-ATTEMPT and SIMPLE protocols, respectively.

A new routing protocol for heterogenous WBANs named Mobility-supporting adaptive threshold-based thermal-aware energy-efficient multi-hop protocol (M-ATTEMPT) [18] has been proposed by the researchers. Apart from this, a system model is built for placing the nodes on human body. To study this protocol, simulations were carried in MATLAB with 10 nodes as randomly placed, initial energy of these nodes 0.5 joules and their transmission range is 10 meters. Results indicate that first node death occurred at 2700 round and total energy of the network lasted till 3500 rounds. Conclusion: proposed algorithm has consumed very less energy and also achieved very good throughput. The same authors in [12] presented reliability enhanced-adaptive threshold based thermal unaware energy-efficient multi-hop protocol (RE-ATTEMPT) for WBANs. The network life time of ATTEMPT and RE-ATTEMPT lasted for 1450 and 1577 rounds, respectively.

Distance aware relaying energy-efficient protocol (DARE) [19] for enhancing the network life time by reducing the energy consumption has been developed. Simulations performed in MATLAB with 56 sensor nodes, initialized with

0.3 joule of energy, 8 body relays with 1 joule. The algorithm is compared with M-ATTEMPT and results indicate that in M-ATTEMPT all the 56 nodes are dead after 714 rounds. While, in the case of DARE, for about 1500 rounds all the nodes are dead.

In [20] stable increased-throughput multihop protocol for link efficiency (SIMPLE) is proposed. Similar simulation parameters discussed in [19] were considered and the performance is evaluated based on the metrics like network period, stability lifetime, residual power, throughput, and pathloss [19]. The first node death occurred at 2010 round with ATTEMPT, while, in case of SIMPLE first node death occurred at 4400 rounds. Also, in other performance metrics SIMPLE algorithm outperformed the state of the art.

As we all know, in line of sight (LOS) and non-line of sight (NLOS) conditions, paths between the nodes are experienced with path loss, multipath fading, shadowing effects and noise effects in both. A cooperative LAEEBA (Co-LAEEBA) [21] routing scheme has been proposed, considering path loss and collaborative learning. A total of 8 nodes have been considered with three as advanced nodes and remaining five as normal nodes. These normal nodes send the information to advanced nodes. The operation of this protocol is divided into many phases like initialization, co-operation and routing, power consumption, relay selection, and path loss phases. Initial energy of each node is given with 0.3 J and 0.1 J for advanced and normal nodes respectively. As we all know that, the performance of the WBANs cab better be understood during simulation studies.

Modified LEACH [22] has been proposed to reduce the energy consumption required for communication. For this, energy and location factors have been considered for electing the CH. The performance evaluation of proposed algorithm, is simulated in MATLAB and then compared with basic LEACH, and LEACH -C. All nodes in LEACH died at 2243, LEACH -C at 2606 and proposed EE-LEACH at 4958 rounds. Moreover, the proposed protocol i.e., EE-LEACH has balanced the network, with less power consumption, improved data throughput, and prolonged network lifetime.

A single protocol cannot meet the multiple quality of service requirements in WBANs. Moreover, the above protocols are based on cost functions considering single parameters such as calculation of residual energy, minimum distance to local base station, RSSI, load balancing, spatial information, links reliability and path loss. Furthermore, it is observed that few of the above works are carried under ideal conditions and didn't consider any radio models. Thus, in this paper we construct a cost function in selecting optimal cluster head by taking number of parameters such as high SNR, average energy and least distance to CH. With this approach, we not only achieve efficient and reliable communication of data, but also improve network lifetime, and energy efficiency of network.

The developed protocol finds usage in sensor networks. Since, the protocol is energy efficient, it can be used in any sensor networks. Other application supported would be Internet of things (IoT). Some of the applications include minimum energy routine protocol and cluster formation of the networks. There are various power-aware routing protocols including nodes route data destined for the base station particularly through the intermediate nodes. To minimize energy usage, the intermediate nodes are chosen in a way that the transmit amplifier energy is greatly minimized. For example, in clustering, the nodes are organized into clusters that enable communication with a local base station. In turn, the local base stations transmit data to the cellular base stations where accessibility by the end user is made possible. This significantly minimizes the distance nodes that are required to transmit their data since the local base station is closer to other nodes in the cluster.

III. Radio System Model and Assumptions

A. Radio Model

A first order radio model defined in [23] is shown in Fig. 1. In this model, the radio will dissipate: total electronic ($T_{elec} = E_{tx} = E_{rx}$) = 50nJ/bit to make the transceiver circuitry up and 100pJ/bit/m^2 for operating transmit power amplifier to achieve desired SNR. Also, we assume that some amount of energy is consumed due to the transmission over channel.

To communicate 'n' bit message to a distance 'd' using the assumed model, radio spends:

$$E_{tx}(n,d) = E_{tx}(n) + E_{tx-amp}(n,d) \quad (1)$$

$$E_{tx}(n,d) = (E_{tx} + E_{rx}) * n + E_{tx-amp} * n * d^2 \quad (2)$$

And, while receiving the same message, the radio dissipates

$$E_{rx}(n) = t_{power} * n \quad (3)$$

where total power is given as tpower = E_{tx} (transmitter power) + E_{rx} (receiver power), and $E_{tx\text{-}amp}$ is power consumed by the transmitter amplifier electronics. Moreover, we also assumed the amount of energy required for sending a packet from wearable sensor 1 to 2 will be same as that of energy required to for transmitting from node 2 to node 1 at a given SNR.

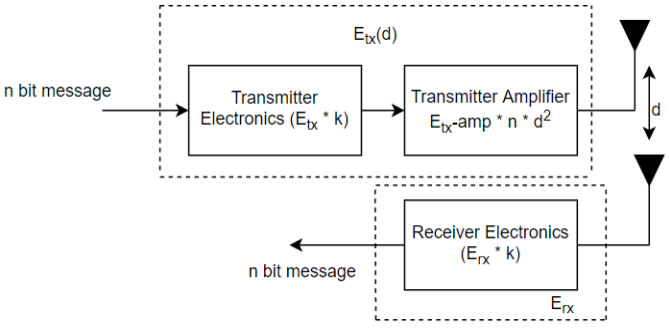


Fig. 1. First Order Radio Model.

B. Energy for Two Modes of Operation

Sensor nodes transmit information to local BS in two modes: direct (single hop) and multi hope. In direct mode, nodes send data straight to BS. If BS is too far from them, nodes transmitter will have to spend more energy to transmit to, as 'd' is greater in eq. 2. With this, nodes batteries drain faster and network lifetime decreases. To demonstrate the above point, Fig. 2 is considered as a simple linear network, where 's' is number of nodes and 'd' is distance between the nodes.

In direct mode, energy spent for transmitting 'n' bit message of a node located at a distance 'sd' from BS is given by:

$$E_{direct} = E_{tx}(n, d = s * d) \quad (4)$$

Equation 1 and 2 can be expressed as:

$$E_{direct} = E_{tx}(n, d = s * d) \quad (5)$$

$$= T_{elec} * n + E_{tx-amp} * n * (sd)^2$$
$$= n(T_{elec} + E_{tx-amp} s^2 d^2 \quad (6)$$

In multihop mode, nodes transmit data to BS through intermediate nodes, acting as routers. These intermediate nodes are selected in such way that their transmitter energy is minimized while sending the data. Consider a linear network shown in Fig. 2 with 3 nodes, if a node 1 wants to transmit to node 3, node 1 should transmit through node 2 only if it satisfies the following condition:

$$E_{tx-amp}(n, d = d_{12}) + E_{tx-amp}(n, d = d_{23}) < E_{tx-amp}(n, d = d_{13}) \quad (7)$$

where d_{12} is distance from node 1 to node 2, d_{23} distance from node 2 to node 3 and d_{13} is distance from node 1 to node 3.

In a multihop mode, nodes near to BS will receive large amount of data, thus these nodes will die soon. Furthermore, nodes far from BS will have to transmit data with more energy causing these nodes to die and eventually the network goes down. Hence, all the nodes can be arranged into clusters, and communicate with local BSs. Moreover, these local BS transmits data to global BS, comprises of personal computer, phones and other smart electronic items [21], whereas, sensor nodes are all kind of biosensors powered by tiny batteries. Thus, clustering seems to be a solution in providing energy efficient communication between the local BS and global BS. However, the local BS should be given with high energy, otherwise this would die soon.

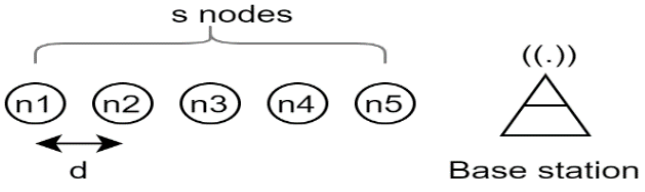


Fig. 2. Simple Network with Linearity.

C. Simulation Parameters

The protocol is based on a model with a person or patient equipped with mMaxBANSize sensor nodes placed outside the body. These nodes will help in monitoring the patient vital signals such as heart rate, body temperature, oxygen levels, blood pressure, blood sugar levels, and respiratory rate, and so on from a remote location. The topology in this network is star with one node acting as a local BS for collecting the data, situated in the centre of the body and others being normal nodes, senses the data and send to local BS in a Multihop or single hop fashion. The reason behind choosing limited number of nodes is due to patient movement in the external environments. All the nodes in the given model are homogeneous and are provided with initial energy E_0 = 0. 5J and considered to have equal computational capabilities. In this work, we assumed a simple first order radio model to run the transmitter and receiver circuitry.

The simulation specifications and radio parameters are shown in Table I.

TABLE I. SHOWS THE SIMULATION PARAMETERS

Parameters	Value
Size of BAN (mMaxBANSize)	6 - 30
Network size	3mx3m
Sink location	Centre
Initial energy (E_0)	0.5 Joules
Number of rounds	6000
Size of each packet	4000 bits
E_{tx} (Transmitter electronics)	50 nJ
E_{rx} (Receiver electronics)	50 nJ
$E_{tx\text{-}amp}$ transmitter amplifier	100 pJ/bit/m^2

IV. PROPOSED METHOD

The protocol is based on communicating the information to BS by avoiding the battery drainage caused due to direct communication i.e., single hop. Disseminating the data to base station, and achieving the easier convergence of network, sensor nodes themselves organize into clusters with one node being cluster head (CH) or local BS. If we chose CH before and fixed for the entire system's lifetime, the CH would die quickly. Thus, there is a necessity to provide random rotation of CHs among all the nodes within the cluster. Moreover, all the remaining wearable sensors within a cluster will transmit data to CH to save energy and further CHs will direct the data to access point by performing data aggregation and segregation. However, the energy consumed by the cluster heads is more when compared with nodes. If it happens continuously, soon the CHs end up with low energy and thus making network down. Thus, CH selection is based on the cost criteria given as: 1) by allocating random signal to noise ratio (SNR) among the nodes, and 2) computing the distance between from nodes to CH. If the node is having high SNR and distance between node and CH is less, then that particular node will become CH in next iteration.

In WBAN scenario, the wearable sensors deployed on a body are fixed with respect to their position after the deployment. All these wearables are given initially an equal amount of power except for the local BSs. Initially all the nodes will be in sleep mode, and gets activated only when its allocated time slot is arrived. Once, the nodes get activated, nodes first check its E_{total} i.e., total energy. Sensor nodes appoint themselves to be CH with certain probability. CH will communicate its status to all wearable sensors in the network. Later nodes in the network identify their neighbors' nodes and forwarder nodes to build a table using eq. (8) and (9).

- Neighbor nodes are identified using:

$$N(n) = \{ j \mid j \in S, d(i,j) < r \} \quad (8)$$

where 'S' is set of all the sensor nodes, r is the communication radius, d(i,j) is distance from ith node to jth node.

- Forward nodes are identified using:

$$FN(i) = \{ j \mid j \in N(i), d(j,BS) < d(i,BS) \} \quad (9)$$

where d(j, BS) is given as distance from jth node to local BS.

Later, each node will decide to which cluster should belong to by choosing CH with minimum energy for communication. Finally, CH will allocate certain schedules to nodes for the transmission of data. With this, node allocated for that particular slot will turn on its radio and transmits the information. Rest all other nodes will be in inactive state.

A. Cluster Head Selection

- Initially when the clusters formation starts, all the nodes within the cluster can participate in CH election. CH election is based on SNR, energy and distance factor of all the nodes.
- Firstly, CH is elected based on evaluation of SNR randomly for all the nodes using temporary variables and rand function. SNR is generally expressed as:

$$SNR = \frac{P_{signal}}{P_{noise}} \quad (10)$$

- If SNR ≥ threshold; only then node will be having higher priority in becoming cluster head. Using this, all the nodes will get an equal opportunity to become CH. Since we have fewer nodes the probability of becoming CHs is high.
- Each node will have to spend certain amount of energy for data transfer, which is different for every node. Later nodes with maximum energy will be participating in CH election process using eq. (11), where E_{node} is node's energy and $E_{residual}$ is remaining energy at current round.

$$E_{node} = \frac{E_{residual}}{E_{total}} \quad (11)$$

- The amount of power dissipated will depend on the distance between source and CH. If distance ≤ do (average distance); nodes estimate their distance with respect to CH and node with minimum distance gets higher probability of becoming next CH. Hence, for the next round, CH is elected using modified equation given in (12):

$$Th = \left[\frac{P}{1-P\left(cr_{mod}\left(\frac{1}{P}\right)\right)} * \left(\frac{SNR_{current}}{SNR_{total}}\right) * \left(\frac{distance_{current}}{distance_{total}}\right) * \left(\frac{E_{current}}{E_{total}}\right) \right] \quad (12)$$

where 'p' represents probability that the node has to become cluster head and 'cr_{mod}' is the current round as shown in algorithm 1.

Algorithm1: Energy efficient communication protocol for WBAN applications

Step-1: Deployment of 'n' sensor nodes, base station (BS)
Step-2: Nodes are initialized with E_o= 0.5 Joules
Step-3: Set-up phase: cluster formation
a) Cluster head selection process begins
Nodes uses LEACH based stochastic algorithm for determining CH for initial round
CH announces CH status to all nodes
Neighbor nodes are identified
$N(n) = \{ j \mid j \in S, d(i,j) < r \}$
Forward nodes are identified
$FN(i) = \{ j \mid j \in N(i), d(j,BS) < d(i,BS) \}$
Wait for Join-request messages
CH creates TDMA schedule and send to all cluster members
b) Steady phase
Nodes transmits the data to CH, based on the slots allocated
for i=1:n
Nodes that have been cluster heads cannot become cluster heads again for P rounds
Each node has a 1/P probability of becoming a cluster head again.
Step-4: Calculation of SNR/ Energy
CH formation is rotated based on
High signal to noise ratio ($SNR = \frac{P_{signal}}{P_{noise}}$), high energy ($E_{node} = \frac{E_{residual}}{E_{total}}$) and lower distance to base station
$$Th = \left[\frac{P}{1 - P\left(crmod\left(\frac{1}{P}\right)\right)} * \left(\frac{SNR_{current}}{SNR_{total}}\right) * \left(\frac{distance_{current}}{distance_{total}}\right) * \left(\frac{E_{current}}{E_{total}}\right) \right]$$
Step-5: end

V. Results and Discussion

Simulations were carried in MATLAB R2019b and performance analysis is compared with various algorithms such as EELEACH, DEEC, TEEN and CRPBA. The following parameters have been considered for performance evaluation: average network life time, throughput, average energy dissipation with varying number of nodes. To have realistic view, we have varied number of nodes between 8 and 30.

Assuming that in a WBANs patient or person, it can be equipped with 8 or 12 or 20 or 30 nodes. For the given number of nodes each protocol is simulated for five times to find exact accuracy of the algorithm and also average number of rounds with 85% of the nodes died.

A. Network Lifetime

In the first iteration, for 8 nodes, after the simulation starts, 85% of all the nodes died after 3100 rounds for EELEACH, 3300 for DEEC, 3900 for TEEN, 4170 for CRPBA and proposed algorithm lasts for 5850 rounds. Similarly, iteration 2, 3, 4, and 5 are carried and their values are shown in Table II. Again, an average has been taken among the results obtained, to get more accurate results (total values of each protocol divided by 5). Finally, for 8 nodes we found that our algorithm has performed 23.58% better than other algorithms. Similarly, for 12 nodes, the average lifetime values are 2159, 2551, 3526, 4060 and 5059 for EELEACH, DEEC, TEEN, CRPBA and Proposed protocol respectively. Later for 12 nodes we found that our algorithm has performed 19.75% better than other algorithms.

Similarly, for 20 nodes, the average lifetime values are 1669, 2055, 2623, 3431 and 4065 for EELEACH, DEEC, TEEN, CRPBA and Proposed protocol, respectively.

Later for 20 nodes we found that our algorithm has performed 16.05% better than other algorithms. Similarly, for 30 nodes, the average lifetime values are 1707, 1880, 2068, 2812 and 3377 for EELEACH, DEEC, TEEN, CRPBA and Proposed protocol respectively. Later for 30 nodes we found that our algorithm has performed 16.7% better than other algorithms. From Fig. 3 it is can be observed that; the network period is decreasing as the number of nodes are increasing. Hence, an optimal number of nodes that can be placed on body are 8 to have very good network lifetime. Reasons: Due to the interference caused by nodes, packet loss, delay, retransmissions increase. With these factors, power of the nodes disseminates faster.

TABLE II. PERFORMANCE EVALUATION OF NETWORK LIFETIME FOR VARIOUS PROTOCOLS WITH 8, 12, 20 AND 30 NODES

Nodes	Network Life Values for five simulations					Analysis
	EELEACH	DEEC	TEEN	CRPBA	Proposed	
8	3100	3300	3900	4170	**5850**	Proposed vs Best 1-(4307/5636) = **23.58%** Better
	3355	3775	3875	4335	**5620**	
	3200	3500	3572	4224	**5575**	
	3757	3558	3550	4550	**5380**	
	3250	3575	3680	4255	**5755**	
Avg Net Life	3332	3542	3715	4307	**5636**	
12	2112	2673	3500	3878	**5200**	Proposed vs Best 1-(4060/5059) = **19.75%** Better
	2220	2527	3425	4258	**5075**	
	2001	2580	3444	3955	**4836**	
	2110	2422	3705	4220	**5160**	
	2350	2555	3556	3988	**5025**	
Avg Net Life	2159	2551	3526	4060	**5059**	
20	1544	1919	2545	3096	**4115**	Proposed vs Best 1-(3431/4065) = **16.05%** Better
	1725	2215	2617	3211	**4256**	
	1780	2007	2528	3857	**3900**	
	1698	2058	2770	3990	**4077**	
	1599	2075	2657	3004	**3975**	
Avg Net Life	1669	2055	2623	3431	**4065**	
30	1705	1902	2002	2770	**3452**	Proposed vs Best 1-(2812/3377) = **16.7%** Better
	1778	1885	2089	2688	**3351**	
	1658	1850	2180	2910	**3317**	
	1595	1897	1998	2805	**3477**	
	1801	1867	2070	2888	**3288**	
Avg Net Life	1707	1880	2068	2812	**3377**	
Average Network Lifetime efficiency= (23.58+23.11++16.05+16.7)/4=19.5%						

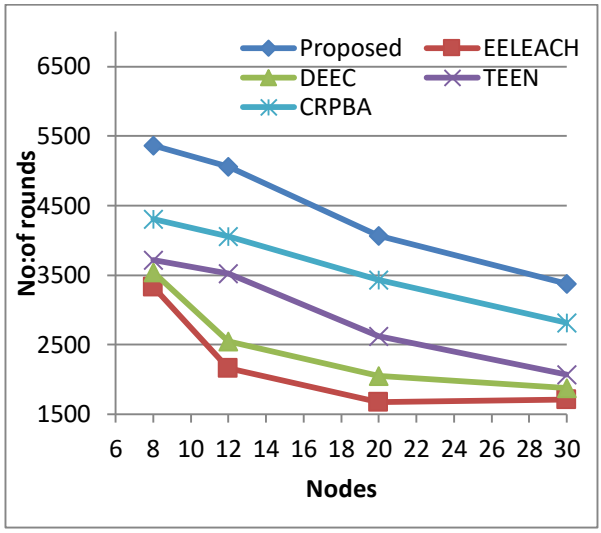


Fig. 3. Number of Rounds over Number of Nodes.

B. Average Remaining Energy Per Node after 85% of Nodes Died

From Fig. 4 and Table III, it is seen that the amount of power dissipated in the network is decreasing as the number of nodes are increasing. For EELEACH, DEEC and TEEN protocols, the even after 85% of the node's death, the amount of energy is decreasing gradually which will lead to network down, whereas in the case of CRPBA and proposed protocol, the curve is flat and leading to efficient network. Hence, an optimal number of nodes that can be placed on body are 8 to have very good network lifetime. Reasons: Due to the interference caused by nodes, packet loss, delay, retransmissions increase. With these factors, power of the nodes disseminates faster.

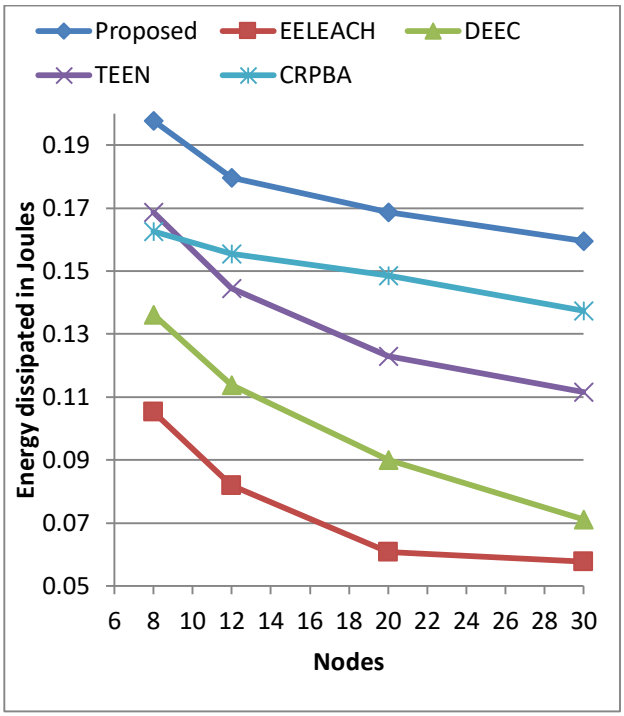


Fig. 4. Number of Nodes over Energy Dissipated.

TABLE III. PERFORMANCE EVALUATION OF AVERAGE REMAINING ENERGY FOR VARIOUS PROTOCOLS WITH 8, 12, 20 AND 30 NODES

Nodes	Average Remaining Energy per Node (Joules)				
	EELEACH	DEEC	TEEN	CRPBA	Proposed
8	0.1052	0.1361	0.1686	0.1627	**0.1977**
12	0.0820	0.1138	0.1445	0.1555	**0.1797**
20	0.0608	0.0901	0.1229	0.1485	**0.1688**
30	0.0577	0.0711	0.1115	0.1375	**0.1595**

C. Throughput

From Fig. 5 and Table IV, BS total numbers of packets reception will depend on how many rounds do the nodes are alive. The more time the alive nodes send more packets to base station and thus increasing the throughput.

Also, it is seen that; the proposed protocol throughput is better than EELEACH, DEEC, TEEN and ERPBA protocols.

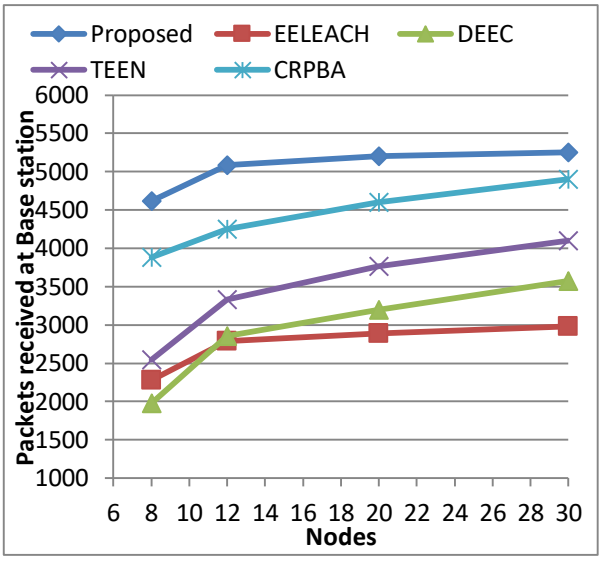


Fig. 5. Number of Nodes over Throughput.

TABLE IV. PERFORMANCE EVALUATION OF THROUGHPUT FOR VARIOUS PROTOCOLS WITH 8, 12, 20 AND 30 NODES

Nodes	Throughput (packets for base station)				
	EELEACH	DEEC	TEEN	CRPBA	Proposed
8	2280	1977	2545	3885	**4615**
12	2788	2857	3328	4254	**5085**
20	2890	3200	3770	4600	**5200**
30	2979	3570	4100	4900	**5257**

VI. CONCLUSION AND FUTURE WORK

Due to limited energy, complicated channel conditions, different body postures and data rates, the communication protocols developed for wireless sensor networks are not appropriate to WBANs. Hence, in this work, SNR based energy efficient communication protocol with a weightage function is developed and analyzed. At the initial phase, number of broadcast flooding's were reduced based on the distance from nodes and CH, then energy saving paths were

selected with respect to location of node and channel conditions. Finally, time slots were allocated for the nodes based on data priority of nodes. From the results, it is observed that developed protocol increases the network lifetime by 19.5%, throughput is increased by 12.61% and average remaining energy by 57.21% for various nodes. The results show that the developed protocol is well suited for time critical applications by reducing the energy consumption while electing the cluster heads and data transmission. It would be appropriate and interesting to implement a prototype and evaluate the performance. Also, it might be challenging to see how it will work in small size of the network.

REFERENCES

[1] Nirupam Bajpai, John Biberman and Yingxin Ye, "ICTs and Public Health in the Context of COVID-19", CSD Working Paper Series: Towards a New Indian Model of Information and Communications Technology-Led Growth and Development, April 2020.

[2] Marcu, Ioana et al. "Arrowhead Technology for Digitalization and Automation Solution: Smart Cities and Smart Agriculture." Sensors (Basel, Switzerland) vol. 20,5 1464. 6 Mar. 2020.

[3] L. U. Khan, I. Yaqoob, N. H. Tran, S. M. A. Kazmi, T. N. Dang and C. S. Hong, "Edge Computing Enabled Smart Cities: A Comprehensive Survey," in IEEE Internet of Things Journal, April 2020.

[4] S. Talari, M. Shafie-khah, P. Siano, V. Loia, A. Tommasetti, and J. Catalão, "A review of smart cities based on the internet of things concept," Energies, vol. 10, no. 4, p. 421, Mar. 2017.

[5] Y. Qu, G. Zheng, H. Ma, X. Wang, B. Ji, and H. Wu, "A survey of routing protocols in WBAN for healthcare applications.," Sensors, vol. 19, no. 7, Apr. 2019.

[6] F. Rismanian, M. Hosseinzadeh, and S. Jabbehdari, "A Review of State-of-the-Art on Wireless Body Area Networks," ijacsa, vol. 8, no. 11, 2017.

[7] M. Alrashidi and N. Nasri, "Wireless body area sensor networks for wearable health monitoring: technology trends and future research opportunities," ijacsa, vol. 12, no. 4, 2021.

[8] A. Sundar Raj and M. Chinnadurai, "Energy efficient routing algorithm in wireless body area networks for smart wearable patches," Comput. Commun., vol. 153, pp. 85–94, Mar. 2020.

[9] Y. Qu, G. Zheng, H. Wu, B. Ji, and H. Ma, "An Energy-Efficient Routing Protocol for Reliable Data Transmission in Wireless Body Area Networks.," Sensors, vol. 19, no. 19, Sep. 2019.

[10] B. Abidi, A. Jilbab, and E. H. Mohamed, "An energy efficiency routing protocol for wireless body area networks.," J. Med. Eng. Technol., vol. 42, no. 4, pp. 290–297, May 2018.

[11] N. Javaid, A. Ahmad, Q. Nadeem, M. Imran, and N. Haider, "iM-SIMPLE: iMproved stable increased-throughput multi-hop link efficient routing protocol for Wireless Body Area Networks," Comput. Human Behav., vol. 51, pp. 1003–1011, Oct. 2015.

[12] A. Ahmad, N. Javaid, U. Qasim, M. Ishfaq, Z. A. Khan, and T. A. Alghamdi, "RE-ATTEMPT: A New Energy-Efficient Routing Protocol for Wireless Body Area Sensor Networks," International Journal of Distributed Sensor Networks, vol. 10, no. 4, p. 464010, Apr. 2014.

[13] N. Kaur and S. Singh, "Optimized cost effective and energy efficient routing protocol for wireless body area networks," Ad Hoc Netw., vol. 61, pp. 65–84, Jun. 2017.

[14] [4] R. A. Khan, Q. Xin, and N. Roshan, "RK-Energy Efficient Routing Protocol for Wireless Body Area Sensor Networks," Wireless Pers. Commun., vol. 116, no. 1, pp. 709–721, Jan. 2021.

[15] M. Anwar et al., "Green communication for wireless body area networks: energy aware link efficient routing approach.," Sensors, vol. 18, no. 10, Sep. 2018.

[16] M. Cicioğlu and A. Çalhan, "Energy-efficient and SDN-enabled routing algorithm for wireless body area networks," Comput. Commun., vol. 160, pp. 228–239, Jul. 2020.

[17] S. Ahmed, N. Javaid, M. Akbar, A. Iqbal, Z. A. Khan, and U. Qasim, "LAEEBA: link aware and energy efficient scheme for body area networks," in 2014 IEEE 28th International Conference on Advanced Information Networking and Applications, pp. 435–440, May 2014.

[18] N. Javaid, Z. Abbas, M. S. Fareed, Z. A. Khan, and N. Alrajeh, "M-ATTEMPT: A New Energy-Efficient Routing Protocol for Wireless Body Area Sensor Networks," Procedia Computer Science, vol. 19, pp. 224–231, 2013.

[19] A. Tauqir, N. Javaid, S. Akram, A. Rao, and S. N. Mohammad, "Distance Aware Relaying Energy-Efficient: DARE to Monitor Patients in Multi-hop Body Area Sensor Networks," in 2013 Eighth International Conference on Broadband and Wireless Computing, Communication and Applications, pp. 206–213, Oct. 2013.

[20] Q. Nadeem, N. Javaid, S. N. Mohammad, M. Y. Khan, S. Sarfraz, and M. Gull, "SIMPLE: Stable Increased-Throughput Multi-hop Protocol for Link Efficiency in Wireless Body Area Networks," in 2013 Eighth International Conference on Broadband and Wireless Computing, Communication and Applications, pp. 221–226, Oct. 2013.

[21] S. Ahmed et al., "Co-LAEEBA: Cooperative link aware and energy efficient protocol for wireless body area networks," Comput. Human Behav., vol. 51, pp. 1205–1215, Oct. 2015.

[22] L. Mao and Y. Zhang, "An energy-efficient LEACH algorithm for wireless sensor networks," in 2017 36th Chinese Control Conference (CCC), pp. 9005–9009, Jul. 2017.

[23] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, p. 10, 2000.

Critical Success Factor of Trusted Elements for Mobile Health Records Management: A Review of Conceptual Models

Fatin Nur Zulkipli[1]
Nurussobah Hussin[2]
Faculty of Information Management
Universiti Teknologi MARA
Puncak Perdana, Selangor, Malaysia

Saiful Farik Mat Yatin[3]
Institute for Big Data Analytics and Artificial Intelligence
Universiti Teknologi MARA
Selangor, Malaysia

Azman Ismail[4]
Universiti Kuala Lumpur Malaysian Institute of Marine Engineering Technology, Lumut, Perak
Malaysia

***Abstract*—Health Information Technology such as Mobile Health Record Management (MHRM) and Electronic Health Record (EHR) depend on each other in maintaining the patients' medical record. For maintaining trust specifically in health information technology development, the relationship among the patients, providers and clinicians needs to be maintained. The present study consists of the understanding of the importance of the trusted elements of mobile health (mHealth) record management implementation in government hospitals. Covid-19 pandemic situation force obeying the technological approach in healthcare delivery. Technology gives a big impact on healthcare industry that deals with confidential data and human life. The increased use of mobile in records management in the wrong way leads the practitioner and communities towards poor quality, security problems, and meaningless data. To fulfil this objective, the conceptual framework has been developed by producing the trust elements for the implementation of mHealth apps in hospitals. Secondary data have been used and analyses to justify the objectives of the study. The findings and discussion have been evolved on correlating the existing literature and the analyses data. Five trusted elements for MHRM have been found: Governance, Professional skills and competency, Mobile Health Records Management (MHRM), Sustainability and, Technological. This paper has evolved the use of electronic health records in the health organizations for the accessibility of trust data and timely access. The involvement success factors of trust elements avoid the petty problem, and inefficient process but giving users convenient and instant access to patients' records.**

***Keywords*—Electronic health record; mobile health record; records management; health information technology; records trust**

I. INTRODUCTION

The advancement in the medical sciences has changed the customers' awareness and transformed the organizations by providing health services in the competitive environment [1]. Sometimes the critical need of the health system cannot be achieved due to the lack of proper documentation of the required papers between the health providers, data processing and timely access [2]. Hence, the need of access to computer-based information has been considered important. The EHR have been executed by hospitals at a rapid scale in recent years across the world [3] and defined as the digitalized format constituting a patient's health-related information that is stored in the computer [4]. This EHR helps to provide easier access of health information such as patient's medical history, their medication, diagnosis, radiology images, lab results and other related medical information to the authorized providers among practitioners. The government has also taken major initiatives concerning the regulation and financial stimulations in different countries such as the USA, UK and Denmark [5] and [6]. The implementation of EHR in hospitals involves the integration and availability of data of the patients by improving the cost-efficiency and establishing good relationships among the patients and doctors. Health care professionals is the primary function in maintaining the nature of the hospital for efficient service [7]. The implementation of EHR helps in providing quality care along with customer satisfaction, cost effectiveness and timely access for complete and precise information [8] and [9]. Although the studies have provided several advantages of EHR, its implementation in the healthcare organizations have been found to be quite complex as its success and productivity depends on various factors [10] and [11]. Dissatisfactions among the patients have caused new exchanges that fulfil their needs by offering them several medical services online. This can be resolved by adopting the mHealth that consists of important information on the patient's medical history. The use of mHealth services has become efficient towards providing medical care. mHealth has been mostly used for disease investigation, treatment assistance, epidemic outbreak tracking and chronic disease management.

The access to mHealth has been considered complex due to the need of infrastructure investment and implementation of several changes in the health system. It also requires the user adaptability for participation in the development of the system [12] and [13]. The success in the implementation of EHR is mainly contributed by factors related to organizational, technological, and human [12]. The human factors include establishing the teamwork for users' participation along with selection of leadership, users gaining training for using the system and providing appropriate support for the promotion and maintenance of the system. Factors related to organization includes the involvement of clinical, administrative and information technology in the healthcare system. The training of knowledge and skills for working with EHR have been provided to the users along with the assurance of the

Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA (UiTM) and Universiti Teknologi MARA Geran Insentif Penyeliaan (600-RMC/GIP 5/3 (018/2021).

confidentiality and security of the data in the healthcare organization. The technological factors involve the usability, infrastructure, interoperability, adaptability, testing and regulation, standards, and policies of the healthcare system.

Almost seven billion individuals are connected to this mobile network, and around 500 million people use personal mHealth apps around the world [14] and [15]. Hence, Hospitals and clinics can use this visibility and usefulness to reach out to patients in the most efficient way possible. The mHealth applications which helps in appropriate management, organization, documents' tracking and maintaining the record for further medical practice has been found to be efficient [16]. Also, the mHealth services maintains the mobile files while managing the workflow using electronic healthcare record management. Mobile files are effective in managing and organizing the paper or system information of the patient's data that helps in treatment procedure by further analysing the history. Trustworthiness of mHealth apps involves the trusted relationship established between the user and the mHealth app to benefit their health which is dependent on mobile health record data. Trustworthiness involves the attributes that are necessary for the individuals [17] such as honesty, competence, and reliability. These are some important trustworthy signals that must be maintained among the users while using the mHealth apps. Therefore, several healthcare models have been developed to cater the need of the increasing healthcare services. Trust elements for mHealth has been one of such models which has been considered as an important model in terms of effective accessibility, affordability, and usability. The use of mHealth in the hospitals have been capable of establishing good communication between the providers and the patients and help in improving the management system of the workflow in hospitals. The mHealth apps have been capable of providing effective communication streamlined between the patients, providers and caregivers allowing 24/7 services as per the condition of the patients.

II. Literature Review

A. Concept of Electronic Medical Records and Mobile Health Records

An EHR document has been considered as an official health document of any individual that is shared among multiple agencies and facilities. It is the digital version of the patient's information, and many consumers are adopting the desire for mobile access to their health records. EHRs are real-time records that available at an instant with security to the authorized users [18]. Health records consist of the medical and treatment history of the patients while an EHR system is the process of going ahead of standardized data collected data in the healthcare organization and is considered to be inclusive with a broader view of patient's care. Therefore, an EHR basically consists of a patient's medical history, diagnosis, treatment strategies, medications, dates for immunization, allergies, images of radiology and laboratory tests [20]. Among all other types of data EHR also involves contact information, information concerning the visit of the patient to healthcare professionals, insurance information, family history, information regarding the patient's condition if suffering from any disease, records of hospitalization and information regarding the surgeries. It also maintains an automated and streamlines provider workflow. One of the important and distinct characteristics of EHR is that all the information related to the health condition of the patient is present in the digitized format with the providers and the authorized providers are capable of managing this document by collaborating with other healthcare providers from multiple organizations [19]. Hence, it can be said that using EHR in the organization helps in building a healthier future for the nation.

The mHealth record has been efficient in improving the quality concerning patient's care and other related care provision [16]. The applications were created utilizing mHealth records, which aims to improve the accessibility and availability of medical records information to make it easier for hospitals and doctors to analyze patients' daily medical activities. The use of android and iOS platforms have contributed to the initiation of the mobile EHR application. The app's mHealth record involves inpatient, outpatient, and emergency patient information; operation schedule; verbal orders; medical consultation; searching patients; on-duty scheduling views; salary views; cafeteria menu; telephone directory; and groupware board [20]. Also, it has been analyzed that the development and adoption of mobile as a new platform in hospitals requires the cooperation of hospital executives.

The program regulations for EHRs and MHRs such as Medicare and Medicaid have been evolved to meet certain criteria. These regulatory programs help EHRs and MHRs to meet specific standards and criteria to ensure the data are trustworthy [21]. The healthcare professionals from the hospitals and healthcare system are assured by using these regulatory programs that EHR and MHR consist of the technologies with technical capabilities, functions, and security for catering the important requirements. However, digitalization has disrupted the healthcare sectors in many ways. Previously, the documented files were used in the form of papers for storage, management and retrieval of data containing patient's information as clinical perspectives. The transformation of a documented-based storage system into EHR has been proved to be efficient in the healthcare system. But this transformation has evolved various issues which have been faced by the providers that need proper resolution. Some major potential issues have evolved that consist of security issues, reduction in data flow, requirement for extra staff training and aspects of slowing down of the system while implementation of the data record of the patient. The findings were categorized under one general themes in Table I, Issues and Challenges of Trusted Mobile Health Records Management in Healthcare. Findings from the review are summarized in Table I and details framework of trust element MHRM are discussed in-depth in the following sections.

Thus, according to [29] and [30], health information management in the healthcare industry is highly critical to develop a comprehensive process for health records management. Consequently, one of the strategies areas on Malaysia Planning - 11 (RMK-11) is to achieve a trusted foundation, integration, and interoperability on mobile application, system transformation, and future-readiness preparation. To achieve the strategies, various elements are

significant to be accounted in planning health records management to ensure data are trustworthy and accurate over the complete records lifecycle. Thus, this article concerns to provide trusted elements in handling patient data to avoid any difficulties and risks on mobile health records towards critical success for MHRM implementation. The analyses also provide a records management framework for readers to understand the sources of the elements.

TABLE I. ISSUES AND CHALLENGES OF TRUSTED MOBILE HEALTH RECORDS MANAGEMENT IN HEALTHCARE

Theme	Types of Issues and Challenges	Explanation	Author(s)
Issues and Challenges of Trusted Mobile Health Records Management in Healthcare	Security	The security of the data has been the major concern. The electronic records of the patient's information can be hacked from the data storage system if the software containing those data is not updated timely and secured [22]. Therefore, a security audit of the healthcare organization is needed for maintaining the security of the patient's record. The electronic health records are designed for streamlining the workflow of the healthcare organization. Hence, these streamlines are created by the developers who are sometimes responsible for creating clunky and difficult work. Also, the staff working with these complex streamlined workflows may require more time for entering the basic information or retrieval of the appropriate record in the system. The transformation of the paper-based information of the patients into a new electronic healthcare record system requires implementation of some additional work. Hence, appropriat instructions need to be given to the staff for handling EHR effectively. Therefore, insufficient knowledge and skills related to the handling of EHR may disrupt the proper functioning of MHRM in the healthcare system due to less efficacy in the workflow.	[22]
	Data privacy	Another major issue involves the data privacy of the patients. The stakeholders are often concerned with the patients' data privacy of getting leaked due to the increase of cybercrime. Lack of appropriate planning and communication proved to be the issues that resulted in the failure of meeting the expectations of the customers adopting MHRM.	[23], [24], [25], [26]
	Lack of technological knowledge	Another major issue involves the lack of technological knowledge among the medical staff members for implementation of EHR [10] and [11]. Also, have several health workers have doubts concerning the value of the EHRs. They may show unwillingness to give up the documentation process. The staff are not aware of the technological advancement which results in delay in EHR implementation. The medical doctor finds it hard to fit with the workflow of the EHR. The software's usability is hampered by design defects or inadequate training.	[10] and [11]
	Technology advancement	The mHealth apps have been evident and it has been confirmed that almost 97,000 mHealth apps have already been used across 62 app stores [27]. One of the reports has revealed that the global mHealth market was predicted to increase from $6.21 billion revenue in 2013 to $23.49 billion by the year 2018 [28]. The potential use of mHealth services in Malaysia has been considered as an alternative delivery channel. This increasing growth of mHealth has led towards the drastic change in the management of the health data with the paradigm shift from mainframe systems situated in the facilities of the healthcare providers towards the apps and other storage systems. It has evolved into a new openness. Previously, the devices were available only in the hospitals, but the development of mHealth apps has evolved flexibility by providing the clinicians with all types of useful resources for the treatment even if located remotely. Hence, this trending use of mHealth services has facing trust data concerns while maintaining the accessibility and cost-efficiency. The population of this new era is now completely dependent on the digitalized technologies and hence require ease for receiving care.	[27] and [28]

III. CONCEPTUAL FRAMEWORK

A healthcare organization initiates mHealth services that majorly involves the clinicians and patients. The major trust elements involved while using mHealth apps by the clinicians includes the adoption of technology, time, awareness, and familiarity with the services. The data of the patients are presented and recorded in EMR and hence the maintenance of privacy and security is the major concern to provide trusted data while implementing mHealth services. Exploring various ideas, frameworks, concepts, and models from connected subject matter is important in selecting the best practise towards the goals of MHRM. The conceptual framework for the given study has been described below:

A. A Trusted Electronic Records Management Framework

Figure 1 explains the credibility and authenticity of electronic records that helps in determining the legacy of the electronic documents along with their effectiveness, preservation value and reproduction history. The activities of electronic record management are ensured by determining the evidence retention and utilization of information. An employment framework of electronic records trusted management based on critical business process management has been constructed by [31]. Yet, this framework is focusing on Government-Controlled Companies not in a healthcare environment as focusing on this study. In consideration of the foregoing, the principles of electronic records in any organization including healthcare, it is depending on the related trusted elements. Hence, the author believes, this model one of the models to consider off.

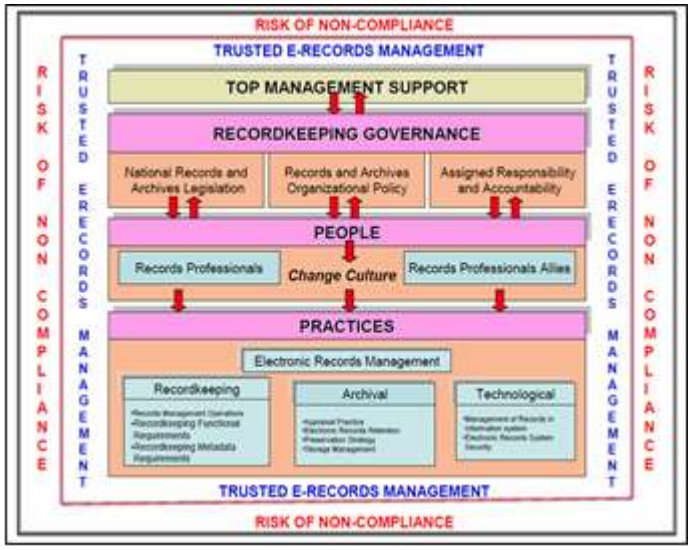


Fig. 1. A Trusted Electronic Records Management Framework [31].

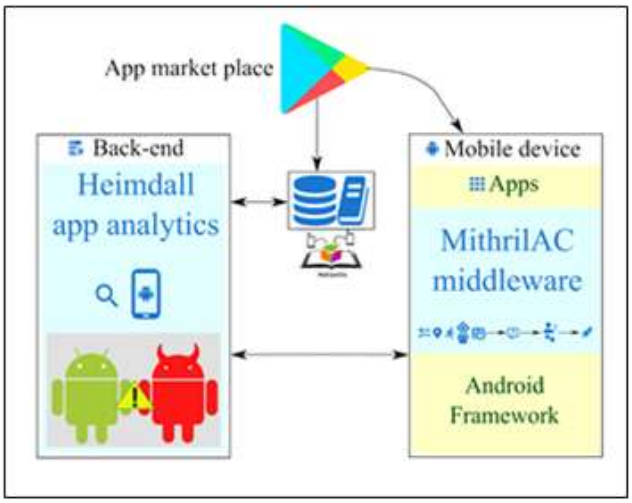


Fig. 2. MITHRIL Conceptual Model [32].

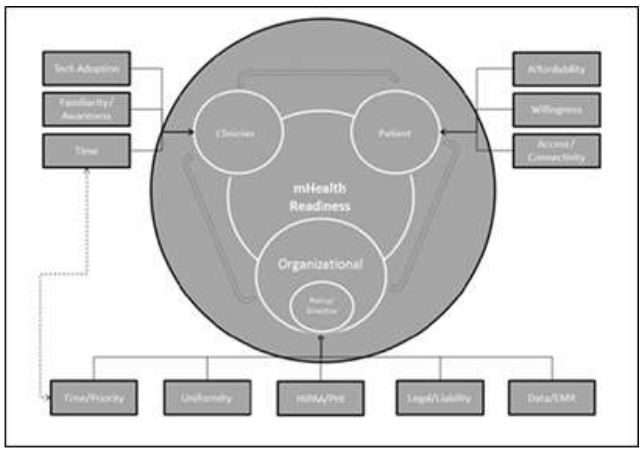


Fig. 3. Conceptual Model for Clinician mHealth Readiness [33].

B. MITHRIL Conceptual Model

Security and privacy are perhaps the most pressing concerns when it comes to the use of mobile platforms in organization. Privacy refers to who has the authority to govern our personal data, whereas security refers to how that authority protects our data. There must have been concerns about security and privacy when mobile devices were first introduced to society. When mobile devices have been introduced to society, there must have been concern about the security and privacy of mobile platforms. [32] in figure 2 highlights, MITHRIL framework is used as a mobile access control framework that permits us to capture access control needs for specific users. This model has described the prevention criteria from the damage of personal data or information. MITHRIL framework which has two major components i.e MithrilAC and Heimdall also contains application analytic system that gathers mobile application metadata.

MithrilAC in figure 2 is the main component of MITHRIL that manage access control on the device. The access and security classification which conduct for security reason and confidential only allows authorized person to access the data in the devices. The second component, Heimdall, detects and prevents infections on mobile devices, tablets, and computers. Using crowd-sourced policy rules such as malware threat for our devices, Heimdall intends to classify applications and generate an initial set of policy rules. Those threats will be sensed and blocked before it can manipulate our personal data. Thus, Heimdall work as a tool that classify the threats which may affect our devices especially the personal data or information.

C. Conceptual Model for Clinician Mobile Health Readiness

MHealth in figure 3 is the practice of providing medical treatment and public well-being through mobile devices. It benefits both the health care practitioner and the patient by increasing productivity and efficiency during the pharmaceutical procedure. The barriers of physicians and related healthcare organizations are assessed for the adoption of mHealth along with the use of EHR data of the patients.

Figure 3 shows, doctors, nurses, and pharmacists rely on mobile health to keep track of their patients. This technological implementation allows them to have consultations with their patients and to save all their patient information in digital form, making it easier for them to track down all the information they require. Simultaneously, this will reduce time in searching and retrieving the data in the mobile device. When their hospital number or identity card number is keyed into the system, all patients who register in the hospital will automatically appear in the system. In addition, mHealth create alertness to the health care professional and patients in dealing with medical treatment. Patients can just check at their smart phones to see what they need to do to improve their quality of health. When the doctors are not there, it serves as a personal advisor for the patients.

D. Conceptual Operating Model for MMAs

Figure 4 is model for Mobile Medical Apps (MMAs) has involved new ways for innovation with low cost healthcare delivery. This model involves the estimation of the MMA interactions with human physiology. Therefore, it is essential to have evidence of their trustworthiness i.e. maintaining privacy of health data, long term operation of wearable sensors and ensuring no harm to the user before actual marketing.

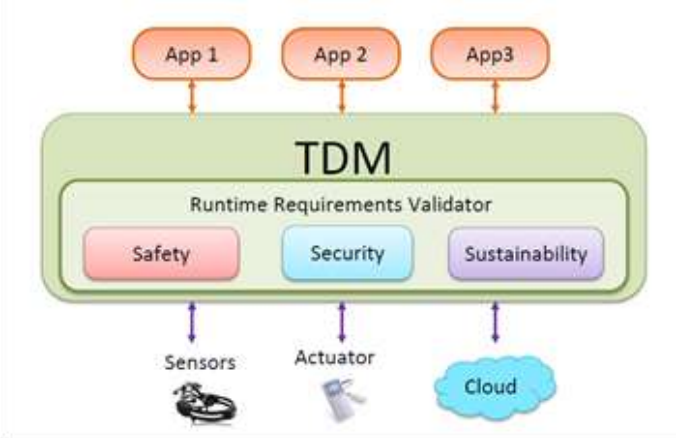


Fig. 4. Conceptual Operating Model for MMAs [34].

E. OAIS Functional Entities Model

The Open Archive Information Systems (OAIS) is a type of archival system that employs both human and technological resources. According to [35] in figure 5, there are six functional entities in an Open Archival Information System, as mentioned by the Consultative Committee for OAIS. OAIS model can provide conceptual framework concerning the preservation and dissemination of digital assets. It defines a set of roles, processes, and functions relevant to long-term preservation. Thus, the OAIS model provides us with a common understanding of what the archives do when they preserve digital information objects. Furthermore, it has laid the groundwork for greater standardisation in the field of digital preservation, including the development of criteria and procedures for analysing and evaluating archival preservation and distribution practises.

Figure 5 reveals the first functional entity is receiving information from producers and packages it for storage. The producer will do the Submission Information Package (SIP), then creates an Archival Information Package (AIP) from the Submission Information Package (SIP) and transfers the newly created Archival Information Package (AIP) to archival storage. Second functional entity is the archival storage function that serves as a repository to stores, maintains, and retrieves archival information. The third functional entity is the data management function that works to coordinates the descriptive information of the archival information package and the system information that supports the archive. The fourth functional entity is the preservation planning to support the archives material which make it accessible to the consumer. The fifth functional entity is the administration function to manage the daily operations of the archives. It handles the users need of the system. It works from the registration, searching, retrieving, and circulating of the archival material. The last functional entity is the access function through registering to the person in charge. This is to make sure the improvement of the security operation increasing to the maximum to keep the archival material stay on its place well.

Preliminary high-level evaluation of aspects of trustworthy recordkeeping addressed in blockchain recordkeeping solutions based on a generic reference architecture and operating model.

Fig. 5. OAIS Functional Entities Model [35].

Blockchain is a database that is controlled by a computer and is not owned by a single person. According to [36], blockchain is an open-source technology that allows for reliable, immutable records of transactions to be stored in decentralised, distributed, and automated ledgers that are publicly accessible. The government and the public sector are beginning to recognise the significance of this technological breakthrough. The blockchain system can help make transaction processes more efficient.

Figure 6 shows the trusted documents, according to [36], must be accurate, reliable, and legitimate. To have the feature of evidence, these three traits must remain unified. Simply said, accuracy means that all the data in the records is correct for the operation.

As a result, to clarify the significance of numerous concepts from diverse matters, the researcher separated the subject matter or theme into five associated subject matter or themes that address problems. This study reviewed five theories and conceptual models as state in figure 1, 2, 3, 4, 5, and 6 related to five elements: Governance, Professional Skills and Competency, Mobile Health Records Management (MHRM), Sustainability, and Technological.

Fig. 6. OAIS Functional Entities Model [36].

These six models detail out the elements of best practices required to effectively manage the mobile electronic health records. As a result, [37] agrees that the lack of a proper standards aspect for mobile health record-keeping in mobile

systems is the primary reason why this topic continues to be debated. According to [3], to tackle this problem, the mobile device system needs precise policies, standardised elements for medical records coordination, and adaptive technology. It details out the elements of best practices required to effectively manage the mobile electronic health records. These attributes are encapsulated in the trusted environment whereby any short fall will implicate the organization and will lead to the risk of non-compliance. In other words, health records management should be recognized as a generic framework to the organization's information strategy, especially in meeting the statutory legal and audit demands to ensure the trustworthiness of the records. Table II shows a list of frameworks/models/concepts/theories for use in this study that are linked to the MHRM's trusted elements.

TABLE II. THE UNDERPINNING MODELS TOWARDS TRUSTED ELEMENTS OF MHRM

Model	Strengths related to the study	Limitations related to the study	Theme/Elements of the study related for this model
1. A Trusted Electronic Records Management Framework [31]	The trusted management and long-term protection of electronic records in agencies.	Records management is not recognized as a generic framework to the organizations' information strategies as well as meeting the statutory legal and audit demands. It is also not recognized as a key role in safeguarding the business continuity of the organizations.	1. Governance 2. Professional skills and competency 3. Mobile Health Records Management (MHRM) 4. Sustainability 5. Technological.
2. Mithril Conceptual Model [32]	MITHRIL assists user to protect their data or information in their devices. The advancement of technology helps us in our daily life which make everything easier.	Although an obvious step towards better access control on mobile devices, it still did not allow dynamic access control.	1. Mobile Health Records Management (MHRM) 2. Sustainability 3. Technological
3. Conceptual Model For Clinician Mhealth Readiness [33]	This technology adoption helps them conduct consultation to their patients. All the personal and medication of their patient will be kept in electronic form which this will be easy for them to trace all the information in need. At the same time, this will safely time in searching all the data about patient when all the data are kept at their mobile.	Based on this model, the access and connectivity are the 10% barriers that were identified from the responses. While conversations touched on this topic at some point, it was not always in a way that access and connectivity was viewed as a barrier. For example, one participant talked about access and connectivity as a barrier. This had limited success.	1. Professional skills and competency 2. Mobile Health Records Management (MHRM) 3. Sustainability 4. Technological.
4. Conceptual Operating Model For MMAS [34]	The efficiency of this tools in the daily shows that this tool cannot be diminish. Patient and health care professional both need Mobile Medical Applications. For example, patient need a guidance by using Mobile Medical Applications meanwhile health care professional needs it as the result to diagnose.	Focus on collecting physiological data for continuity process only and does not provide a model for health records management process.	1. Governance 2. Mobile Health Records Management (MHRM) 3. Sustainability 4. Technological.
5. Preliminary High-Level Evaluation Of Aspects Of Trustworthy Recordkeeping Addressed in Blockchain Recordkeeping Solutions Based on A Generic Reference Architecture and Operating Model [36]	The information, context, and structure of records must all be captured when they are created. If the details are not complete, the transaction cannot be completed by each party. Then there's consistency, which is a component of trustworthiness. To sustain efficiency in daily operations, all processes must be consistent.	Focus on aspects of trustworthy recordkeeping addressed in blockchain recordkeeping solutions based on a generic reference architecture only and does not mentioned and provide an important model of recordkeeping life cycle as essential for records trustworthiness itself.	1. Governance 2. Professional skills and competency 3. Mobile Health Records Management (MHRM) 4. Sustainability 5. Technological.
6. OAIS Functional Entities Model [35]	The process of this framework is stability because of In ingest, the process will be required receiving information from producers and packages it for storage. These entities serve as a repository for archive data, storing, maintaining, and retrieving it. It connects the archival information package's descriptive information with the archive's system information.	Insufficiency of empirical research on the dimensions for health information model/framework.	1. Governance 2. Professional skills and competency 3. Mobile Health Records Management (MHRM) 4. Sustainability 5. Technological.

Thus, this research solely focuses on the needs of trusted elements for the success factor of trusted elements for MHRM adoption of MHRM in healthcare organization. As described by other authors in the literature, the elements are crucial to ensure healthcare organization have a comprehensive rule and regulation of mobile device in term of standard, elements for records coordination, adoptive technologies, policies, and security [35]. Hence, the results are interesting and help to show the important theme/element for the trusted elements of MHRM are Governance, Professional skills and competency, Mobile Health Records Management (MHRM), Sustainability, and Technological. These themes/elements are based on other researchers' findings as a guide to advance the electronic records specifically in adoption into mobile for health records management in Malaysian.

IV. Significance of the Study

The electronic health records and the mHealth records have been found to be effective in maintaining the personal health records of the patients in digitized format ensuring the safety and confidentiality of the data. These electronic health records are now present in the mobile format. mHealth records have enabled the patients to access their health information through internet and telecommunication devices. Hence, it has evolved the initiation of mHealth services for the ease of the patients and clinicians. The mHealth apps have the potential to help the patients and providers in identifying the medical conditions prescriptions accordingly from the different locations that helps in reducing the medical errors and enhances the improvement in health conditions during the emergency condition at the time of treatment with the new providers. The ease of access to the data of patients have been another major important characteristic of the mHealth services. The satisfying result from patients and management based on previous experiences of the implementation as discussed above demanding from the practitioners for enhancement of workflow process by using MHRM since the mobile device is beneficial in the collection, assessment, amendments, and transfer of data proficiently. Now, the medical conditions of the patients are recorded and submitted in digitized format that has helped the clinicians to treat their patients from remote locations. As a result, any adoption of technology for health records management in hospitals complied with the right principle such as trusted elements as suggest by author lead to trust, confidence, and secure feelings for users who involved in the health records process from the creation until disposal. This will be resulted in well-organized records, reduce file storage space, difficulties for updating reports at the workstation, and prevent manual handling of files that delay time for patients and wasted money in recruiting staff. This was because the usage of mobile will be easier, faster, and efficient. The petty problem such as distance workstations, delay time for data updating or inefficient process can be avoided through mobile device management, but also gives users convenient and instant access to patients' records. Besides, the cost efficiency is also maintained as the daily visit to the hospitals have been reduced due to access through mHealth apps in the mobile devices.

V. Limitations

The present study has also involved certain limitations include the unawareness of use of EHR among the patients and medical professionals that results in providing lack of important information while maintaining the patients' record. These issues give a big impact on accuracy, consistency, and integrity of the data. Another major limitation involves the cost as many of the healthcare system cannot bear the high cost software for maintaining the data security and confidentiality of the patients. This may result in the loss of important data from the electronic system.

VI. Conclusion

The implementation of trusted elements of MHRM has led to a positive impact on the health conditions of individuals. The existing knowledge and understanding about the trusted elements of EHR has helped in understanding the basic aspects of MHRM and its effect while adopting it in the government hospitals for enhancing the technology adoption. The literature review of the given study has provided evidence that has implicated the elimination of the traditional approaches in the healthcare system and adopted the technology perspectives for providing care to the patients with accurate health information. The important trusted elements have played a major role in maintaining the ethical consideration of using the mHealth record management in hospitals. These trusted elements in turn affect the healthcare environment and hence, the results implicated the positive impact of using mHealth services. Trust elements as mentioned all play a role in increasing the efficacy of a health organisation while maintaining a trustworthy data and smooth workflow.

Acknowledgment

This article is financially supported by Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA (UiTM) and Universiti Teknologi MARA Geran Insentif Penyeliaan (600-RMC/GIP 5/3 (018/2021).

References

[1] Zaid, A.A., Arqawi, S.M., Mwais, R.M.A., Al Shobaki, M.J. and Abu-Naser, S.S., "The Impact of Total Quality Management and Perceived Service Quality on Patient Satisfaction and Behavior Intention in Palestinian Healthcare Organizations." Technology Reports of Kansai University., vol. 62, no. 3, pp. 221-232, 2020.

[2] Chuah, F.L.H., Tan, S.T., Yeo, J. and Legido-Quigley, H., "Health system responses to the health needs of refugees and asylum-seekers in Malaysia: a qualitative study." International journal of environmental research and public health, vol. 16, no. 9, p.1584, 2019.

[3] Kim, E., Rubinstein, S.M., Nead, K.T., Wojcieszynski, A.P., Gabriel, P.E. and Warner, J.L., "The evolving use of electronic health records (EHR) for research." In Seminars in radiation oncology, vol. 29, No. 4, pp. 354-361, 2019.

[4] Cowie, M.R., Blomster, J.I., Curtis, L.H., Duclaux, S., Ford, I., Fritz, F., Goldman, S., Janmohamed, S., Kreuzer, J., Leenay, M. and Michel, A., "Electronic health records to facilitate clinical research." Clinical Research in Cardiology, vol. 106, no. 1, pp. 1-9, 2017.

[5] Terziev, V., "The good practices in the regulation of social development." International E-Journal of Advances in Social Sciences., vol. 5, no. 14, pp. 568-578, 2019.

[6] Lee, C. and Ma, L., "The Role of Policy Labs in Policy Experiment and Knowledge Transfer: A Comparison across the UK, Denmark, and Singapore." Journal of Comparative Policy Analysis: Research and Practice, vol. 22, no. 4, pp. 281-297, 2020.

[7] Glover, G., 2019. Relationships Between Nursing Resources, Uncompensated Care, Hospital Profitability, and Quality of Care.

[8] Krousel-Wood, et al., "Implementing electronic health records (EHRs): health care provider perceptions before and after transition from a local basic EHR to a commercial comprehensive EHR." Journal of the American Medical Informatics Association, vol. 25, no. 6, pp. 618-626, 2018.

[9] Ommaya, A.K., Cipriano, P.F., Hoyt, D.B., Horvath, K.A., Tang, P., Paz, H.L., DeFrancesco, M.S., Hingle, S.T., Butler, S. and Sinsky, C.A., 2018. Care-centered clinical documentation in the digital environment: Solutions to alleviate burnout. NAM Perspectives.

[10] Eichler, H.G., Bloechl-Daum, B., Broich, K., Kyrle, P.A., Oderkirk, J., Rasi, G., Santos Ivo, R., Schuurman, A., Senderovitz, T., Slawomirski, L. and Wenzl, M., "Data rich, information poor: can we use electronic health records to create a learning healthcare system for pharmaceuticals?." Clinical Pharmacology & Therapeutics, vol. 105, no. 4, pp. 912-922, 2019.

[11] Setia, P., Menon, N. and Srinivasan, S.S., "EHR application portfolio and hospital performance: Effects across hospitals with varying administrative scale and clinical complexity." Information & Management, vol. 57, no. 8, p.103383, 2020.

[12] Y.H. Sidek and J.T. Martins, "Perceived critical success factors of electronic health record system implementation in a dental clinic context: an organisational management perspective." International journal of medical informatics, vol. 107, pp. 88-100, 2017.

[13] Greenhalgh, T., Wherton, J., Shaw, S., Papoutsi, C., Vijayaraghavan, S. and Stones, R., "Infrastructure revisited: an ethnographic case study of how health information infrastructure shapes and constrains technological innovation." Journal of medical Internet research, vol. 21, no. 12, pp. 16093, 2019.

[14] PR Newswire. 2015. The mHealth (mobile healthcare) Ecosystem: 2015-2030 – Opportunities, Strategies and Forecasts. PR News Wire. Retrieved from http://www.prnewswire.com/news-releases/the-mhealth-mobile-healthcare-ecosystem2015--2030--opportunities-cha.

[15] Choi, W., Park, M. A., Hong, E., Kim, S., Ahn, R., Hong, J., Yeo, S., "Development of mobile electronic health records application in a secondary general hospital in Korea," Healthcare Informatics Research, vol. 19, no. 4. pp. 307–313, 2013.

[16] M. Idzwan and M. Salleh, "Evaluating The Performance of Malaysian Health Care Providers With Partial Least Squares Path Modeling Electronic Health Records (EHRs) System," pp. 1–27, 2020.

[17] O'Neill O., "Trust trustworthiness and transparency." Retrieved from https://www.efc.be/human-rights-citizenship-democracy/ trust-trustworthiness-transparency/ (2015, accessed 10 March 2019).

[18] Asulyan, T., 2018. Barriers to the Adoption of Electronic Health Records by Physicians in Hospitals.

[19] Rajkomar, et al., "Scalable and accurate deep learning with electronic health records." NPJ Digital Medicine, vol. 1, no. 1, pp.1-10, 2018.

[20] Jayewardene, D., 2020. Mobile Smartphone Applications for Healthcare Practitioners (Doctoral dissertation, University of Leeds).

[21] Espirito Santo, M. D. F. N., 2017. Evaluation of health outcomes associated with medication in southern Portugal using a novel approach for medication review: ReMeD study (Doctoral dissertation).

[22] K. M. Kuo, C. F. Liu, and C. C. Ma, "An investigation of the effect of nurses' technology readiness on the acceptance of mobile electronic medical record systems," BMC Med. Inform. Decis. Mak., vol. 13, no. 1, pp. 1–14, 2013.

[23] Roehrs, A., da Costa, C. A., da Rosa Righi, R., da Silva, V. F., Goldim, J. R., and Schmidt, D. C., Analysing the performance of a blockchain-based personal health record implementation, Journal of Biomedical Informatics, Vol 92, 2019,103140. https://doi.org/10.1016/j.jbi.2019.103140.

[24] Helmers, R., Doebbeling, B.N., Kaufman, D., Grando, A., Poterack, K., Furniss, S., Burton, M. and Miksch, T., Mayo Clinic Registry of Operational Tasks (ROOT): A Paradigm Shift in Electronic Health Record Implementation Evaluation, Mayo Clinic Proceedings: Innovations, Quality & Outcomes, Vol 3, No 3, 2019, pp.319-326. https://doi.org/10.1016/j.mayocpiqo.2019.06.004.

[25] Choi, E., Xiao, C., Stewart, W.F. and Sun, J., "Mime: Multilevel medical embedding of electronic health records for predictive healthcare," arXiv preprint arXiv:1810.09593, 2018.

[26] Mukamba, N., Beres, L. K., Mwamba, C., Law, J. W., Topp, S. M., Simbeza, S., et al. How might improved estimates of HIV programme outcomes influence practice? A formative study of evidence, dissemination and response, Health Research Policy and Systems, Vol 18, 2020, pp. 1-11. doi: http://dx.doi.org/10.1186/s12961-020-00640-7.

[27] Ronchi, A.M., "e-Health: Background, Today's Implementation and Future Trends." In e-Services, pp. 1-68, 2019. Springer, Cham.

[28] Waldman L, Reed P, Hrynick T. Accountability in Health Systems and the Potential of mHealth.

[29] K. M. Andrews, "Best Practices To Establish Successful Mobile Health Service In A Healthcare Setting" Pepperdine University, 2016.

[30] S. Kim, "A mobile device security implementation model for a national medical center complying with the HIPPA security rule," 2016.

[31] Aliza, I., 2010. Assessing the practice of trusted electronic records management in Malaysia government-controlled companies (Doctoral dissertation, Universiti Teknologi MARA). Retrieved from http://ir.uitm.edu.my/id/eprint/5106.

[32] P. K. Das, "Context-Dependent Privacy and Security Management on Mobile Devices," ProQuest Diss. Theses, p. 147, 2017.

[33] B. P. Weichelt, "Health in Your Hand: Assessment of Clinicians' Readiness To Adopt Mhealth Into Rural Patient Care," ProQuest Diss. Theses, pp. 1–104, 2016.

[34] Priyanka Bagade, "Evidence-based Development of Trustworthy Mobile Medical Apps," ProQuest Diss. Theses, pp. 1–144, 2015.

[35] V. L. Lemieux, "Trusting records: is Blockchain technology the answer?," Rec. Manag. J., vol. 26, no. 2, pp. 110–139, 2016.

[36] V. L. Lemieux, "Blockchain And Distributed Ledgers As Trusted Recordkeeping Systems," November, pp. 41–48, 2017.

[37] A. Murad, "the Impact of Mobile Health Applications on Emergency Medical Services and Patient Information Privacy Phd Dissertation (3-Paper Format) By Claremont Graduate University - Center of Information All Rights Reserved .," Claremont Graduate University, 2014.

Non-linear Multiclass SVM Classification Optimization using Large Datasets of Geometric Motif Image

Fikri Budiman, Edi Sugiarto
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia

***Abstract*—Support Vector Machine (SVM) with Radial Basis Functions (RBF) kernel is one of the methods frequently applied to nonlinear multiclass image classification. To overcome some constraints in the form of a large number of image datasets divided into nonlinear multiclass, there three stages of SVM-RBF classification process carried out i.e. 1) Determining the algorithms of feature extraction and feature value dimensions used, 2) Determining the appropriate kernel and parameter values, and 3) Using correct multiclass method for the training and testing processes. The OaO, OaA, and DAGSVM multi-class methods were tested on a large dataset of batik motif images whose geometric motifs with a variety of patterns and colors in each class and containing similar patterns in the motifs between the classes. DAGSVM has the advantage in classification accuracy value, i.e. 91%, but it takes longer during the training and testing processes.**

Keywords*—Geometric motif; image classification; multiclass; non-linear; large dataset

I. INTRODUCTION

Studies focusing on classification of image recognition have a high level of complexity if it has a large dataset with many different groups or multiple classes. The use of large datasets in question is data with large amounts (Lot of Data) with structured data. Moreover, the boundaries among classes cannot be separated by linear hyperplane (non-linear) due to their high level of image feature similarity. There are several supervised Machine Learning algorithms which can be used for image recognition classification, such as hierarchical-based decision trees, K-Nearest Neighborhood algorithms, partition-based K-means and minimum-distance, and networks based Artificial Neural Network (ANN) like perceptron algorithm, Backpropagation Neural Network (BNN), and Support Vector Machine (SVM). These classification algorithms are categorized as Shallow Learning type since they still require some application of feature extraction algorithms to produce feature image dataset. Feature extraction is a fundamental part in classification as feature dataset obtained from proper feature extraction can maximize the accuracy value of classification results [1-2].

This article elaborates a comparative study and evaluation of multi-class SVM methods which have been carried out. The advantages of classification using SVM are such as the ability to significantly gain good classification accuracy values for image features with high data dimensions [3, 4]. However, it is still necessary to test the classification with several different kinds of feature dimensions to get the right dimensions for use [1-2]. The advantages of SVM than those of Artificial Neural Network (ANN) in classification are that SVM does not involve all vector dataset for training image features in forming hyperplane and margins as class separators, and only vector datasets for contributing image features (support vector) are used for hyperplane formation and margins. In addition, SVM determines the hyperplane by maximizing the distance among classes (maximum margin), so that it has high generalization for the testing dataset. Thus, it is better than ANN in which it searches for a hyperplane by principally minimizing its gradient and depending on the number of parameters used [5-8]. SVM classification works using the principle of Structural Risk Minimization (SRM) to enable it to produce good generalizations with hyperplane fields which can minimize the average error in managing training dataset [7].

Several previous studies showed that classification with SVM can significantly increase the accuracy value as in the classification to recognize the texture of honey pollen images in which the SVM accuracy results were better than Multi-layer Perceptron classification method, Minimum Distance Classifier, and K-Nearest Neighbor [8-9]. SVM classification was tested by two public databases of DNA micro array to classify tumors and non-tumors resulting in a classification accuracy value in which SVM was better than ANN [10]. SVM by using default kernel parameter was applied for non-linear multi-class classification with a dataset of five types of batik textures. The results showed that the classification accuracy value using SVM was better than by using Minimum Distance Method and Backpropagation Neural Network [6, 11].

SVM was initially introduced by Boser, Guyon, and Vapnik in 1992 and only used as binary classifier [12]. SVM has the convenience to maximize non-linear classification patterns since SVM can overcome over-fitting with soft margins by replacing each dot product of testing feature using a non-linear kernel function matrix [12]. The strategies carried out in developing non-linear multi-class SVM classification are still considered to have some weaknesses for large number of dataset samples. To solve such weaknesses of non-linear multi-class SVM classification on large-scale image recognition, there are three experimental stages which can be carried out to optimize the use of SVM method in non-linear multiclass

classification, so that it may maximize classification accuracy value. The first stage requires an experiment focusing on the use of qualified feature extraction by determining feature extraction algorithm and its proper parameters to gain qualified dimensions and feature values [1-2,13], the second is conducting an experiment to determine the best kernel and its parameter values which are fit with the conditions of the dataset used [4,5,7,14], The classification of each new dataset with SVM depends on the kernel function used and the parameters used. And the third requires an experiment on using the right method to handle the training and testing process in multi-class. The method commonly used for multi-class is by using combination approaches of several SVM binary or two classes [15-16].

In obtaining maximum classification accuracy value using SVM, it requires the best results from the three experiments by using the new image dataset. The author has conducted the first and the second experiments for 4 non-linear classes using one-against-all method for geometric pattern image dataset with a high level of feature similarity [1-2,5]. The dataset used Batik images with traditional motifs from Indonesia. The traditional Batik images have very diverse geometric decorations and high similarity of motifs as well as possess multi-scale patterns and multi-color resolution [17-18]. Such batik with traditional Indonesian motifs along with the method of creating this batik has been recognized by UNESCO on 2 October 2009 as a "Representative List of The Intangible Cultural Heritage of Humanity".

The third experiment was as important as the first and the second ones in increasing the maximum accuracy value. Consequently, this article used a study using multi-class SVM methods by combining several binary SVM approaches, such as: one-against-all, one-against-one, and directed acyclic graph SVM from previous studies. Moreover, to find out the ability of these methods in maximizing the value of accuracy and classification time for multi-class non-linear with new datasets, an experimental evaluation of these methods was required using a large dataset of images with geometric motifs in the form of traditional Indonesian Batik images consisting of 4 classes and 7 classes. The results of this experiment can be used as the basis to develop new research methods which are more efficient in training and testing time and have maximum accuracy values in handling non-linear multi-class in SVM classification with large datasets.

II. Related Work

The fundamental first step in images classification is the process of applying feature extraction method to generate feature values. Features are unique characteristics of texture features of each image so that they can be recognized through digital image processing [19]. The results of feature extraction are converted in the form of statistical feature features which is then used for classification in the recognition phase. Maximum accuracy value of image classification depends on whether the value of generated image features is good or not. Thus, the experiment on applying the right feature extraction method can be an interesting stage in the initial image recognition research [1-2,19-20]. Feature extraction methods are commonly divided into three, i.e. structural, statistical, and spectral [20].

Discrete Wavelet Transformation (DWT) which is a spectral method is a feature extraction method often used and applied to 2-dimensional images having multi-resolution spaces with varying scale transformations and can produce features based on differences in image intensity in several sub-band spaces [20-22]. DWT can perform the process of changing signal or wave data which is a combination of time and frequency into a series of wavelet coefficients that are easier to analyze, and can be used for feature extraction in images with geometric decorative motifs that have aperiodic or interrupted signals disconnected and noisy. The non-linear multi-class classification experiment in this study used the value of the image feature based on the author's previous research [1], i.e. using statistical features in the form of energy and standard deviation values obtained from wavelet coefficients by using DWT level 3 method with wavelet type of daubechies 2. For the introduction of batik motifs in the research by Rangkuti, Harjoko, and Putro [22], the use of wavelet daubechies 2 was also better than the haar, coiflets, and biorhogonal types.

Following the experiment on the proper feature extraction method to use, it is also significantly important to determine non-linear kernel function and its parameters which best apply to SVM classification of non-linear multi-class [5, 23]. SVM classification was initially only used to distinguish two classes which could be separated by a hyperplane in the form of a linear line [24]. In the real case, classification in general is more about separation among nonlinear classes, or it will be very difficult to be linearly separated. Furthermore, SVM can be developed into a nonlinear classifier by using a kernel trick. In maximizing non-linear SVM classification pattern with the kernel trick, over fitting is minimized using soft margin concept on the hyperplane by replacing each dot product of feature with a non-linear kernel function matrix to determine the support vector [24-26].

One of the functions of kernel is to solve non-linear problem used to determine the support vector by mapping from the initial training feature data to the new training feature data which has a feature space with higher dimensions without defining the function of input space to the new feature space. [27-28]. Optimization of non-linear SVM classification highly depends on the use of kernel function and its proper parameters. There are several functions of non-linear kernel to replace the mapping to new dimensional feature space, including Polynomial, Gaussian/Radial Basis Function (RBF), sigmoid, Multi Quadratic Inversion, and Additive [7].

The Gaussian RBF kernel is highly recommended to gain maximum non-linear classification results for a new dataset [4, 7, 14, 28] since it has the same performance as the linear kernel on the parameters cost (C) and gamma (γ) / sigma (σ) with a certain value in the optimization of classification. The best parameters combination for C and γ values will be obtained by a hyperplane with the right soft margin so that maximum accuracy of the classification results can be achieved [7, 28]. There is no range requirement for the estimated value of C and γ as test values for the Gaussian RBF kernel parameters. Determination of the RBF-SVM kernel parameters is to obtain a hyperplane and margin with minimal classification errors. The use of correct parameter values of C and γ will also result

in a low measure of the deviations (variance) and a low measure of error contribution (bias). High variance with too high C value and too low γ value can cause over fitting and high bias. On the contrary, if C value is too small and γ is too big, under fitting will occur. In each test, the combination of parameters C and γ can be classified several times with several different training and testing datasets. This is to ensure that there is no excessive over fitting on the tests with different testing data [4-5]. Gaussian RBF kernel can be elaborated as follows:

$$\text{Gaussian: } K(x_i,x_j) = \exp(-\|x_i,x_j\|^2/(2\sigma^2)) \quad (1)$$

$$\gamma = 1/2\sigma^2 \quad (2)$$

$$K(x_i,x_j) = \exp(-\gamma\| x_i,x_j\|^2) \quad (3)$$

$K(x_i,x_j)$ is the value of each kernel matrix element, with x_i and x_j as data point pairs. The value of gamma (γ) as a kernel parameter is used to determine the maximum result in optimization of classification, so it is necessary to estimate a parameter in the form of a constant parameter value for the kernel parameter (γ).

In some previous studies, RBF kernel and its parameter determination were recommended to use to obtain maximum non-linear classification results for the new dataset [5,14,21,29-30]. Determining the value of this parameter is necessary since the function of Gaussian RBF kernel is to substitute dot product mapping from old dimensional features to the new ones depending on the conditions of the image dataset used. Evaluation to optimize the parameters for Gaussian RBF-SVM kernel needs to be done to get the classification with the smallest errors in the image dataset with geometric textures which have many variations and a high level of texture similarity. The parameter values applied to non-linear multi-class experiments using a large number of geometric motif image dataset in this article are from the results of the author's previous research [5]. This study used 4 classes of image datasets with geometric motifs and the maximum classification accuracy value was obtained with a combination of low bias and low variance RBF kernel parameters at the value of $C = 2^7$ and $\gamma = 2^{-15}$ [5].

In addition to having its own complexity in determining the right hyperplane, in multi-class non-linear SMV classification applied to large datasets with image features which have similarities among classes, it is also necessary to evaluate which multiclass method to be applied. The use of non-linear SVM multi-class training and testing methods can be used based on the use of two classes/binary classification, i.e. by combining several binary SVMs. This method is much easier and more practical to apply than by combining all datasets consisting of several classes into the form of optimization problems [16,31]. Some other commonly used methods are: one-against-all (OaA)/winner takes all (WTA), one-against-one (OaO) / Max Wins Voting (MWA), and Directed Acyclic Graph (DAG) [15]. Of all these methods, the superiority of each method in terms of accuracy value, testing and training time duration has not been clearly found out yet, it still depends on the size of the dataset and the number of classes [15,31]. In the methods above, several binary classifications are used in training and testing phases. Of all the other methods, OaA uses the least binary classification as if there are n classes; then n binary classification models are needed. In OaA, there is an unbalanced distribution of the training dataset classes in determining the hyperplane for each binary classification. For example, there are 5 classes, class +1 is class a dataset; then class -1 is a combination of the remaining class datasets, i.e. class dataset b, c, d, and e. Such class imbalance might be problematic when applied to large datasets with many classes since large combined class datasets will cause a slow training phase. Besides, incorrect classification predictions may occur because the training process focuses more on large combined class datasets. The one-against-all method in previous research [5], uses an approach for multi-class SVM for 4 classes by combining kernel parameter functions that can adjust multiple hyperplane and margins in several classifications of two classes, the test stages carried out are as shown in Fig. 1.

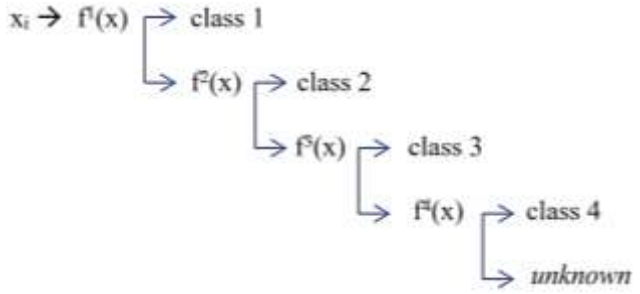


Fig. 1. Stages of Testing One-against-all 4 Classes.

In this study [5] a simplification of the process was carried out by using a combination of several binary SVMs with the one-against-all method (Fig. 1), which was compared using the one-against-one method. The one-againts-one method for 4 classes is done with 6 SVM testing models as shown in Fig. 2. Each test classification model is carried out on data from two classes, namely for data from class i and class j.

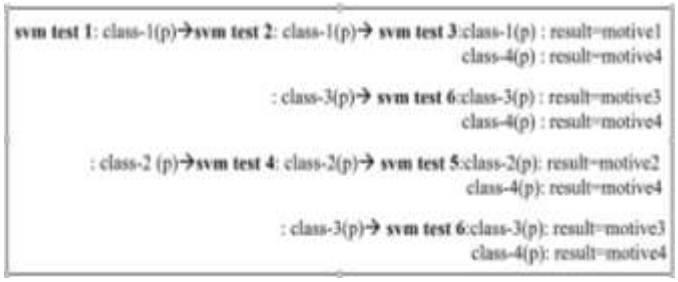


Fig. 2. Stages of Testing One-against-one 4 Classes, p=Positive.

In OaO and DAG there is no class merger, so in the training phase a binary classification model is needed as much as (n(n-1))/2, the value of 'n' is the number of classification class, thus the training phase will be slower as the number of classification classes increases [16]. The difference between OaO and DAG is in their test method, the binary classification model in OaO is first tested and the results are used for the voting method; while in DAG, binary directed acyclic graph method is used. The testing method in DAG is more efficient than OaO, because DAG produces a faster test time complexity for large datasets compared to using OaO, this is because OaO performs two stages in the testing phase [15, 32]. Although the

testing phase is slower, previous research experiments using OAO approach showed that the results of the accuracy value are better than the other multi-class approaches [33-34]. With the advantages and disadvantages of the OaA, OaO, and DAG methods, it is necessary to test the performance of these multiclass methods, the test is carried out in the case of large image datasets, the images used are images with geometric motifs that have high similarity between classes.

III. Proposed Method

This section describes the proposed method to produce maximum accuracy values. In this case, the non-linear multiclass classification optimization method with SVM that uses Large Datasets of Geometric Motif Image can be explained in detail as follows:

1) The feature creation phase is used to optimize the classification accuracy results with feature extraction using Discrete Wavelet Transformation (DWT). Optimization of the use of DWT is done by comparing the use of decomposition level 1 to 5 and wavelet types db 1 to db 5. The features of each sub-band are energy and standard deviations values obtained from the wavelet coefficient values contained in each sub-band.

2) The optimization phase of the SVM classification results with the Grid-Search and Cross Validation processes, to minimize over fitting and obtain a combination of RBF kernel parameters values (C and γ) in the space parameters that produces the maximum classification accuracy value.

3) The results of the first phase are in the form of optimization of the use of wavelet levels and types, and the results of the second phase are the optimal values of the C and γ parameters, used in experiments using DAG SVM, OAA SVM, and OAO SMV, in the classification of 4 classes and 7 classes, with variations in the number of datasets. 200 to 3000 batik image data.

The proposed classification optimization method is as shown in Fig. 3.

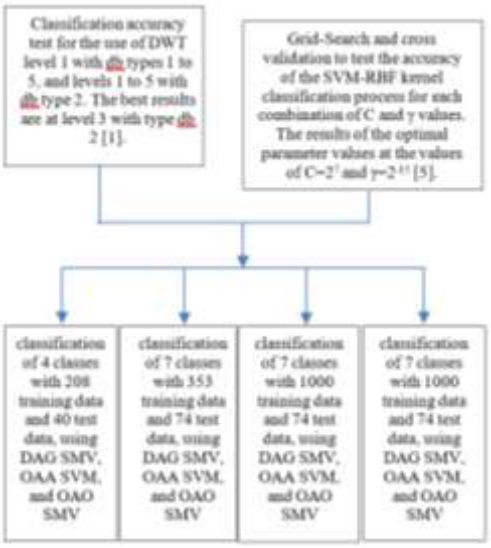


Fig. 3. Proposed Method.

IV. Results and Discussion

The implementations are conducted using a computing platform based on Intel Core i5 1.6 GHz CPU, 8 GB RAM, and Microsoft Windows 10 Professional 64-bit. The development environment is MATLAB programming language.

The datasets used were images of traditional Indonesian Batik which have several classes of decorative patterns with very diverse geometric motifs and high degree of motif patterns similarity among classes. Batik is an original cultural heritage from Indonesia in the form of the beauty of art works on cloth media and contains philosophical meaning of life in each depicted decorative motifs. Batik, which is a work of art on cloth, which is to decorate the surface of the textile by holding the dye. The process of making Batik artwork is by applying color retention popularly known as wax-resist dyeing process. This process of making Indonesian traditional Batik is recognized by UNESCO as a "Representative List of The Intangible Cultural Heritage of Humanity". Each class of Batik motif patterns has a deep meaning of philosophy of life which reflects the noble elements of life. The dataset used in the experiment consisted of 4 classes and 7 classes of Indonesian traditional batik patterns. 4 classes consisted of motifs: Ceplok, Kawung, Nitik, and Parang. In addition, 7 classes consisted of motifs: Ceplok, Kawung, Nitik, Parang, Sidomukti, Lereng, and Slobog (Fig. 4). Total dataset for 4 classes consisted of 208 training data and 40 testing data. Tmeanwhile, 7 classes dataset were tested for 353 training data with 74 testing data, 1000 training data with 74 testing data, and 3000 training data with 74 testing data.

Fig. 4. Seven Classes of Batik Motif.

The level of prediction accuracy of the test dataset image recognition is measured by using the confusion matrix measurement technique which is divided into positive data prediction classes (true positive / TP; false positive / FP) and negative data (true negative / TN; false negative / FN) [22].

$$\text{Accuraracy} = (TP+TN) / (TP+TN+FP+FN) \quad (4)$$

The correctness of the image class prediction from the classification results compared to the actual image class results in the TP, FP, TN, and FN values as shown in table 1.

TABLE I. CONFUSION MATRIX

Classification		Predicted Class	
		Class = yes	Class = no
Actual Class	Class = yes	TP	FN
	Class = no	FP	TN

In this non-linear multi-class classification experiment with this large dataset, the use of SVM classification was maximized through three experimental stages as described in the introduction section. At the first stage, based on the results of the author's previous experiment [1], it used a feature with a statistical characteristic of the energy value and the standard deviation of the Discrete Wavelet coefficients. The calculation of the energy value of each pattern component in the sub-band is with the parameter Es (energy in a certain sub-band calculated by CA / CH / CV / CD), Cs (the matrix coefficient on the sub-band pattern calculated by CA / CH / CV / CD), c (vector coefficient values for all sub-bands), are as follows:

$$E = 100*\Sigma_{x,y}(Cs_{x,y}.\hat{}2) / \Sigma_{n=1:\,m}(c_n.\hat{}2) \quad (5)$$

The standard deviation formula (Std) for each pattern component in the sub-band with the parameter Cs (the matrix coefficient on each component of the sub-band pattern calculated by CA / CH / CV / CD), r is the average value of Cs, and n is the amount of data Cs:

$$Std = (\Sigma_{x,y}(Cs_{x,y} - r)^2 / (n-1))^{½} \quad (6)$$

The level 3 wavelet decomposition is carried out on the sub-band approximation at level 2, as shown in Fig. 5.

Transform level 3 with wavelet type daubechies 2 which has been used in the previous experiment [1], produces 20 values in the feature vector for each image of the training and test data. Each sub-band has 2 (two) feature values, energy and standard deviation, so that the wavelet coefficient vector and the feature vector generated at each level 3 are:

$[E_{CA3}\ Std_{CA3}\ E_{CH3}\ Std_{CH3}\ E_{CV3}\ Std_{CV3}\ E_{CD3}\ Std_{CD3}$

$E_{CH2}\ Std_{CH2}\ E_{CV2}\ Std_{CV2}\ E_{CD2}\ Std_{CD2}$

$E_{CH1}\ Std_{CH1}\ E_{CV1}\ Std_{CV1}\ E_{CD1}\ Std_{CD1}]$

The results of two author's previously conducted experiments [1-2] showed that in addition to the application of the right feature extraction algorithm, the number of values in the feature vector used must also be precisely right, because any values would greatly affect the maximum accuracy value of the classification.

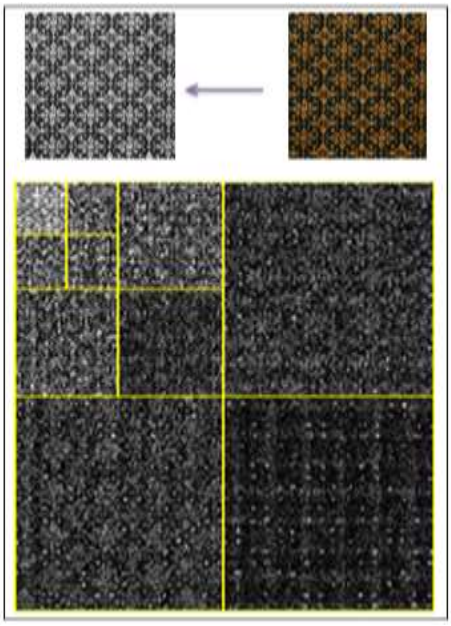


Fig. 5. Level 3 Wavelet Decomposition.

The second stage of the kernel and the parameter values used are the results of the previous author's experiment [5], the determination of the RBF kernel parameter values on the scale tested to obtain the maximum accuracy value in the non-linear multi-class SVM classification method has been successfully identified with the space parameter. In looking for the maximum value of the RBF kernel with low bias and low variance in the parameter range specified in the study, obtained at the optimal parameter value $C=2^7$ and $\gamma=2^{-15}$. In gaining the maximum classification accuracy value, it is necessary to evaluate the use of the correct parameter value, since the optimal parameter value highly depends on the image datasets used. Testing the use of a combination of cost function/ C and gamma/ γ parameters on RBF kernel is better conducted in several times classification with different training and testing datasets. In each test the combination of parameters C and γ is carried out the k-fold Cross Validation (CV) method with 10 classifications using 10 training datasets and different tests (10-fold). In this study [5] the experiment using a small k value less than 10 resulted in more and more image class recognition errors, this is because the training dataset is getting less and less good at representing the hyperplane and margins for each class. Determining the best value for this parameter combination is important, to ensure that there are no over-fitting and under-fitting that could cause sharp differences in accuracy results when using different test datasets. From the test results in the previous study [5] it can be recommended to classify images that have geometric decorative motifs with a non-linear multi-class SVM-RBF kernel, using the Grid-search range $C = \{2^{6.5}, 2^{6.75}, 2^7, 2^{7.25}, 2^{7.5}, 2^{7.75}, 2^8\}$ and $\gamma = \{2^{-14.5}, 2^{-14.75}, 2^{-15}, 2^{-15.25}, 2^{-15.5}, 2^{-15.75}, 2^{-16}\}$, these ranges are to get the optimal parameter combination value.

The third stage was conducted by comparing the methods of training and testing in OaA, OaO, and DAG multiclass using a large dataset of batik images. The use of these three methods was carried out in 4 stages to measure the accuracy value of the

classification and the time required for training and testing process (in seconds). The results of the first stage with 4 classes consisting of 208 training data and 40 test data (table 2) using DAG showed the highest accuracy value. However, it required longer training and testing time than those by using OaA and OaO. The results of the first stage and the second stage (table 3) are obviously the same which used 7 classes with 353 training data and 74 testing data. Furthermore, DAG method resulted in more superior accuracy value with longer training and testing time.

DAG has the best accuracy value than OaA and OaO in the third stage of the experiment (table 4) and in the fourth stage (table 5). However, this is not significant regarding the time required for the training process. There is a significant time leap when the training dataset was increased to 3000 images (fig. 6). By applying the three stages of the process using SVM for this nonlinear multiclass classification, DAG method is considered to be good for achieving the accuracy value of the classification. However, but there are still some constraint as the longer training time needed along with the increasing size of the dataset.

TABLE II. RESULTS OF 4 CLASSES EXPERIMENTS

Method	4 classes					
	Training data	Testing data	Correct classification	Accuracy (%)	Training Time	Testing time
DAG SVM	208	40	37	93	0,834	0,021
OAA SVM	208	40	35	88	0,274	0,016
OAO SVM	208	40	26	65	0,104	0,016

TABLE III. RESULTS OF 7 CLASSES EXPERIMENTS, 353 DATASETS

Method	7 classes					
	Training data	Testing data	Correct classification	Accuracy (%)	Training Time	Testing time
DAG SVM	353	74	62	84	1,863	0,038
OAA SVM	353	74	58	78	0,769	0,024
OAO SVM	353	74	39	53	0,165	0,017

TABLE IV. RESULTS OF 7 CLASSES EXPERIMENTS, 1000 DATASETS

Method	7 classes					
	Training data	Testing data	Correct classification	Accuracy (%)	Training Time	Testing time
DAG SVM	1000	74	67	91	3,755	0,058
OAA SVM	1000	74	63	85	1,439	0,029
OAO SVM	1000	74	42	57	0,477	0,020

TABLE V. RESULTS OF 7 CLASSES EXPERIMENTS, 3000 DATASET

Method	7 classes					
	Training data	Testing data	Correct classification	Accuracy (%)	Training Time	Testing time
DAG SVM	3000	74	67	91	13,726	0,276
OAA SVM	3000	74	63	85	4,550	0,033
OAO SVM	3000	74	42	57	1,544	0,020

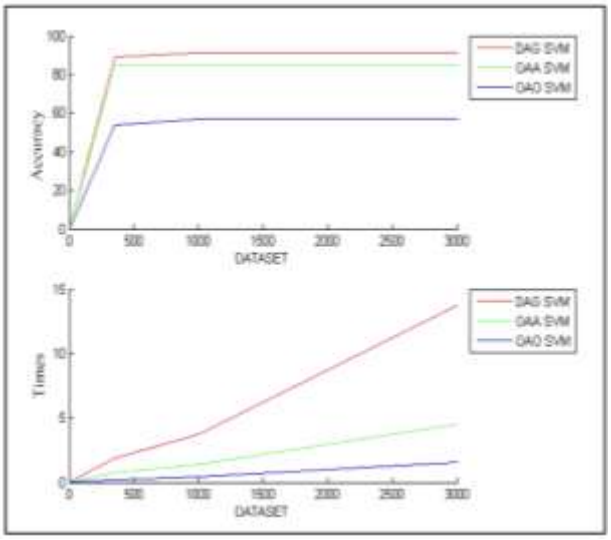


Fig. 6. Comparison of 7 Class Performances.

V. CONCLUSION AND FUTURE WORK

These three stages of non-linear multiclass SVM-RBF classification are still able to produce good accuracy values, for the use of large datasets of traditional Indonesian batik images with highly varying geometric motifs. This good accuracy value depends on the image features, the kernel and its parameters, as well as on the methods used during the training and testing process.

The training and testing methods applied to 1000 and 3000 batik images as the dataset with 7 motif classes showed that DAG can produce a consistent accuracy value of 91%. Nevertheless, with the increase in the number of datasets from 1000 to 3000, it showed that the required training time was increasingly bigger. Thus the SVM classification with large dataset images has problems with the speed of the training process with the increasing number of datasets, thus further research is still needed in the form of developing a combination model of SVM-RBF with Deep Learning to minimize classification time on big data.

REFERENCES

[1] Budiman, F., Suhendra, A., Agushinta, D., and Tarigan, A., ”Wavelet Decomposition Levels Analysis For Indonesia Traditional Batik Classification”, Journal of Theoretical & Applied Information Technology, 92(2):389-394, 2016.

[2] Budiman, F., Sugiarto, E., “Image Feature Extraction of Numbers and Letters Using Matrix Segmentation”, Scientific Visualization, 12(1): 120-131, 2020.

[3] Azhar, Ryfial, Tuwohingide, Desmin, Kamudi, Dasrit, Sarimuddin., dan Suciati, Nanink, ”Batik Image Classification Using SIFT Feature Extraction, Bag of Features and Support Vector Machine”, Procedia Computer Science, Elsevier, 72(1) 24-30, 2015.

[4] Renukadevi, N,T. dan Thangaraj, P., “Performance Evaluation of SVM-RBF Kernel for Medical Image Classification”, Global Journal of Computer Science and Technology Graphics & Vision, Global Journal Inc, USA, 13(4):15-19, 2013.

[5] Budiman, F., “SVM-RBF Parameters Testing Optimization Using Cross Validation and Grid Search to Improve Multiclass Classification ”, Scientific Visualization, 11(1), 80-90, 2019.

[6] Yuan, Qing-Ni, Lu, Jian, Huang, Haisong, dan Pan, Weiji, 2014, ”Research of Batik Image Classification Based On Support Vector Machine”, Computer Modelling & New Technologies 18(12B) 193-196, 2014.

[7] Hofmann, M., Support Vector Machines- Kernels and the Kernel Trick. 1st Edn., Bamberg University, 2006.

[8] Ahmad, A., Hashim, U. K. M., Mohd, O., Abdullah, M. M., Sakidin, H., Rasib, A. W., & Sufahani, S. F., “Comparative analysis of support vector machine, maximum likelihood and neural network classification on multispectral remote sensing data”, International Journal of Advanced Computer Science and Applications (IJACSA), 9(9), 529-537, 2018.

[9] Fernandez-Delgado, M., Carrion, P., Cernadas, E., Galvez, J. F., & Sa-Otero, P., “Improved classification of pollen texture images using SVM and MLP”. In 3rd IASTED International Conference on Visualization, Imaging and Image Processing (VIIP2003) (Vol. 2), 2003.

[10] Ahmad M. Sarhan, “Wavelet-based Feature Extraction For DNA Microarray Classification”, Springer Science+Business Media B.V. Artif Intell Rev, 39:237-249. 2013.

[11] Herulambang, W., Hamidah, M. N., & Setyatama, F., Comparison of SVM And BPNN Methods in The Classification of Batik Patterns Based on Color Histograms And Invariant Moments, In 2020 International Conference on Smart Technology and Applications (ICoSTA), pp. 1-4, IEEE, 2020.

[12] Boser, B.E., Guyon, I.M., and Vapnik, V.N., “A training algorithm for optimal margin classifiers”, In D. Haussler, editor, 5th Annual ACM Workshop on COLT, pp. 144–152, Pittsburgh, PA. ACM Press, 1992.

[13] Makri, E., Rotaru, D., Smart, N. P., and Vercauteren, F., “EPIC: Efficient private image classification (or: Learning from the masters)”. In Cryptographers’ Track at the RSA Conference (pp. 473-492). Springer, Cham, 2019.

[14] Rosales-Perez, Alejandro, Escalante, Hugo Jair, Gonzales, Jesus A., dan Reyes-Garcia, Carlos A., ”Bias and Variance Optimization for SVMs Model Selection”, Procedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference. pp. 136-141, 2013.

[15] Hsu, C. W., & Lin, C. J., “A comparison of methods for multiclass support vector machines”. IEEE transactions on Neural Networks, 13(2), 415-425, 2002.

[16] Groenen, P. J. F., & van den Burg, G. J. J., Multiclass Support Vector Machines With Gensvm. In Computer Data Analysis and Modeling: Stochastics and Data Science, pp. 43-50, 2019.

[17] Riyanto, Rahayu, Slamet, and Pamungkas, Wisnu, Handbook of Indonesian Batik, The Institute For Research and Development of Handicraft and Batik Industrie, Yogyakarta, 1997.

[18] Tirta, I., Batik Sebuah Lakon, 1st Edn., Gaya Favorit Press, Jakarta, ISBN: 9789795154280, pp: 278. 2009.

[19] Nixon, Mark S., dan Aguado, Alberto S., Feature Extraction and Image Processing, Academic Press is an imprint of Elsevier, 2008.

[20] Selvarajah S., Kodituwakku, “Analysis and Comparison of Texture Features Content Based Image Retrieval”, International Journal of Latest Trends in Computing, 2(1):108-113, 2011.

[21] Virmani, Jitendra, Kumar,Vinod, Karla, Naveen, dan Khandelwal, Niranjan, ”SVM-Based Characterization of Liver Ultrasound Images Using Wavelet Packet Texture Descriptors”, Journal of Digital Imaging (2013) 26:530–543, Springer, 2013.

[22] Rangkuti, Abdul Haris, “Content Based Batik Image Classification Using Wavelet Transform And Fuzzy Neural Network”, Journal of Computer Science 10 (4): 604-613, Science Publications, 2014.

[23] Gorunescu, Florin, Data Mining – Concepts, Models, and Techniques, Berlin : Springer-Verlag Berlin Heidelberg, 2011.

[24] Wu, Xindong, Kumar, Vipin (ed.), The top ten algorithms in data mining. CRC press, Taylor & Francis Group, London, New York, 2009.

[25] Boser, B., Guyan, I., dan Vapnik, V., 1992, “A Training Algorithm for Optimal Margin Classfiers”, In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp.144-152, New York:ACM Press.

[26] Ali, Muhammad Asim, and Zain, Ahmed Siddiqui, "Automatic music genres classification using machine learning", International Journal of Advanced Computer Science and Applications (IJACSA), 8(8) : 337-344, 2017.

[27] Schölkopf, Bernhard, and Alexander J. Smola. "Learning with kernels: support vector machines, Regularization." Optimization, and Beyond. MIT press 1.2, 2002.

[28] Hsu, Chih-Wei., Chang, Chih-Chung., dan Lin Chih-Jen., A Practice Guide to Support Vector Classification, Department of Computer Science, National Taiwan University, 2010.

[29] Syarif, Iwan, Prugel-Bernnett, Adam, dan Wills, Gary, “SVM Parameter Optimization Using Grid Search and Genetic Algorithm to Improve Classification Performance”, Telkomnika Journal, 14(4):1502-1509, 2016.

[30] Gaspar, Paulo, Carbonell, Jaime, Oliveira, dan Jose Luis, “On the Parameter Optimization of Support Vector Machines for Binary Classification”, Journal of Integrative Bioinformatics, 9(3):201, 2012

[31] Duan, K. B., & Keerthi, S. S. , “Which is the best multiclass SVM method? An empirical study. In International workshop on multiple classifier systems”, Springer, Berlin, Heidelberg, pp. 278-285, 2005.

[32] Agarwal, N., Balasubramanian, V. N., & Jawahar, C. V., “Improving multiclass classification by deep networks using DAGSVM and Triplet Loss”, Pattern Recognition Letters, 112, 184-190, 2018.

[33] Zheng, H.B., Liao, R.J., Grzybowski, S dan Yang, L.J., “Fault diagnosis of power transformers using multi-class least square support vector machines classifier with particle swarm optimisation”. IET Elect. Power Appl. Vol 5, Iss 9, pp. 691-696, 2011.

[34] Liu, Yang, Jian-Wu Bi, and Zhi-Ping Fan, "A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm", Information Sciences 394, 38-52, 2017.

Recent Progress, Emerging Techniques, and Future Research Prospects of Bangla Machine Translation: A Systematic Review

M. A. H. Akhand[1], Arna Roy[2], Argha Chandra Dhar[3], Md Abdus Samad Kamal[4]*
Dept. of Computer Science and Engineering, Khulna University of Engineering & Technology Khulna-9203, Bangladesh[1, 2, 3]
Graduate School of Science and Technology, Gunma University, Kiryu 376-8515, Japan[4]

***Abstract*—Machine Translation (MT), the way of translating texts or documents from a source language to a target language automatically without human intervention, has gained popularity in the growing information technology-based era of globalization. Bangla is a major language, and several MT studies with different tools and techniques have been investigated in the last two decades. Considering the importance of the Bangla language and its prospects in MT studies, this study provides a comprehensive review of existing Bangla MT studies to meet the timely demand. Specifically, at first, the basic ideas of different MT methods (Rule-based, Example-based, Statistical, Neural, and Hybrid) and performance measures of MT are presented as a background study of the present review. Then an overview of the Bangla language and a brief description of the available Bangla-English corpora are provided. Next, a description of the existing Bangla MT studies is provided categorically following the common strategic fashion to create a valuable reference for current researchers in the field that is also suitable for non-expert users. The achieved performances of individual methods are also compared in a tabular form. Finally, a number of future research prospects are revealed from the studies, encouraging researchers and practitioners to develop a better and comprehensive Bangla MT system.**

Keywords*—Machine Translation (MT); Bangla language; rule-based MT; example-based MT; statistical MT; neural MT; hybrid MT

I. Introduction

The task of content (e.g., voice, speech, texts) translation from one natural language to another has become indispensable in politics, business, research, and other areas, and a human expert usually handles such tasks. Human translators perform a masterful job interpreting conversations between two or more persons (e.g., country chiefs, tourists, business giants) who speak in different languages. Due to rapid globalization, translation becomes essential for ordinary people; translation of web content (e.g., website, document) is also necessary for the era of digitalization and the internet. To handle the translation of such huge contents (especially text, document, and web), machine translation (MT) is a promising research area [1]. In general, MT refers to translating texts or documents from the source language (SL) into the target language (TL) without human intervention.

Individual natural languages advocacy inherently dominated respective MT studies as resources (e.g., corpus) and language-dependent efforts. Specifically, corpus, rules, and other resources of a particular language pair (e.g., English-German) are not usable in MT for another language pair. Moreover, the MT system developed for a particular language pair is not appropriate for other cases as an individual language holds distinct grammar and phrase rules. As the most internationally used language, MT researches are mainly English language concentric. High resource availability and major MT studies with remarkable performance are available for English-French [2], English-German [3], English-Chinese [4] language pairs. Thousands of other natural languages, including several major languages, are remained much behind in MT activities. Specifically, Bangla is one of the most broadly spoken languages, with approximately 228 million native speakers (fifth-most) and 37 million second-language speakers (seventh-most) [5]. However, MT resources and studies are limited for Bangla. Several Bangla-English MT studies are available with different methods, but their achievements are not significant compared to the resource-rich language [6]–[8]. Therefore, it is a timely demand to line up existing Bangla MT studies for the researchers who intend to work in this promising research field and find a motivation to enhance Bangla MT.

This study is a comprehensive review of Bangla MT studies focusing on individual methods and techniques employed, corpus and/or resources used, and performance achieved. As a prerequisite, the fundamentals of various MT methods, the significance of the Bangla language, and MT performance measurements are explained briefly in Section II. Benchmark corpora and resources for Bangla MT; and individual Bangla MT studies are summarized under different MT categories in Section III. The significance of the current review study and prospects of Bangla MT studies are discussed in Section IV and Section V, respectively. Finally, Section VI briefly concludes the present study with few remarks.

II. Background

Various techniques and tools were developed in the last few decades through remarkable research efforts for appropriate MT outcomes in different languages. At the basic level, MT executes the translation of atomic words from one language to another using a dictionary. Nowadays, it is possible to translate whole sentences through corpus techniques, rule-based grammatical techniques, or even idioms

*Corresponding Author

and phrase-based techniques. However, no MT system is available currently that can translate as efficiently as a human translator. Therefore, MT has emerged as a rising research field in Artificial Intelligence.

A. Basic MT Approaches

Existing MT methods are broadly categorized into two approaches: rule-based MT (RBMT) approach and data-driven approach. In the data-driven approach, a parallel corpus is the main element to develop the MT model. Three basic methods in this category are example-based MT (EBMT), statistical MT (SMT), and neural MT (NMT). On the other hand, the hybrid MT (HMT) approach is also available, which combines two or more basic methods. Fig. 1 depicts the classification of the available MT approaches. The following subsections briefly discuss the fundamental points of the five basic MT methods to understand different Bangla MT studies easily.

1) Rule-Based MT (RBMT): Based on linguistic information, RBMT generates translations through human expert-produced grammatical rules regarding verbs, prepositions, inflections, etc. [9]. Dictionaries (unilingual, bilingual or multilingual) and collection of rules covering the main semantic, morphological, and syntactic regularities of source and target languages are the basic requirements of RBMT [10]. Roughly RBMT can be divided into three approaches: Direct MT (DMT), Transfer-based MT, and Interlingua MT [11]. Fig. 2 is the well-known Bernerd Vauquois' pyramid of MT, which shows comparative depths of intermediary representation, interlingua MT at the peak, followed by the transfer-based, then direct translation.

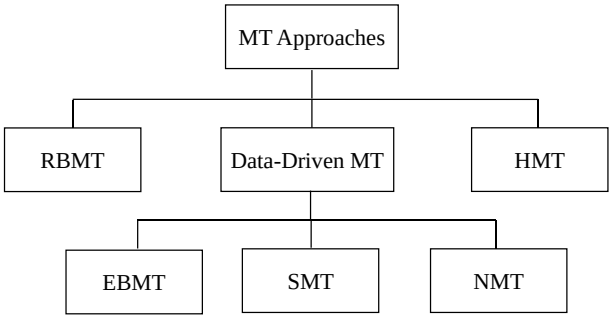


Fig. 1. Classification of MT Methods.

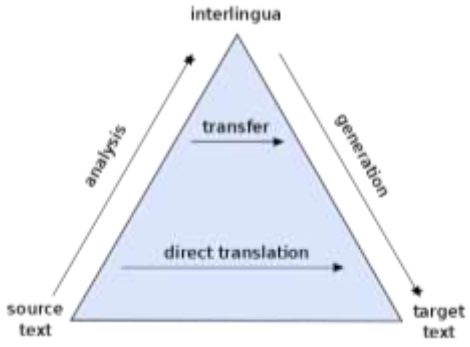


Fig. 2. Bernerd Vauquois' Pyramid Showing Comparative Depths of Intermediary Representation, Interlingua Machine Translation at the Peak, followed by Transfer-based, then Direct Translation.

DMT is the oldest MT approach based on the dictionary that is used in the pioneer Georgetown–IBM public MT demonstration [12]. DMT attempts to match an SL to a TL, i.e., translating word-by-word directly [13]. The method is quite simple, but the translation quality is very poor due to the lack of syntax and semantic analysis of the source language. Then, the RBMT approach with transfer and interlingua [13] is developed to overcome the limitations of DMT. The interlingua approach was proposed to be language-independent [14]. Fig. 3 shows the block diagram of RBMT with interlingua representation. Interlingua is considered an abstract, homogenous, unambiguous, and independent universal language. For translating using interlingua, the source sentence is converted to the interlingua first, and then the interlingua is converted to the target language sentence [9].

2) Example-Based Machine Translation (EBMT): EBMT is a corpus-based data-driven approach based on human language learning process [15]. The main motivation of EBMT is that human does not translate through deep linguistic analysis. Instead, a human translator first properly decomposes input sentences into specific fragmental phrases, then translates these fragmental phrases into other language phrases, and finally correctly composes these fragmental translations into one long sentence. EBMT was introduced for the English-Japanese language pair as RBMT is complicated for English-Japanese and other language pairs due to structural differences [1].

Fig. 4 shows the basic building block of the EBMT model. Sample sentences from SL and TL are stored as examples in a bilingual corpus (i.e., dictionary), a significant component of this model. The SL sentence is fragmented depending on the granularity of the system and followed by a search for (set of) examples from the dictionary that match (or closely matches) the input SL fragment string, and the relevant fragments are picked. The TL fragments corresponding to the relevant fragments are extracted. If the match is exact, the fragments are recombined to form TL output; else, find the TL portion of the relevant match corresponds to a specific portion in SL and align them. Finally, a combination of relevant TL fragments is performed in order to form a legal grammatical target sentence. Further action to translate the untranslated portions (if happen) using a dictionary (called translation memory) has been investigated recently to improve EBMT performance [16].

3) Statistical Machine Translation (SMT): SMT is proposed presuming that language has an inherent logic that might be helpful to treat language mathematically. In SMT, translations are produced based on probability generated through the statistical analysis of bilingual aligned corpora [17]. SMT does not need much knowledge of the SL and TL like RBMT. Fig. 5 shows a simplified block diagram of SMT using decoder, translation model (TM), and language model (LM). The probabilistic TM assigns a score to every possible translation of source text. The language model measures the fluency of the output and assigns each sentence a probability. In the decoding phase, the translation with the best score is

selected [18]. SMT is a corpus-dependent approach, and the requirement of a large human-translated corpus with various linguistic information is its main drawback.

4) Neural Machine Translation (NMT): NMT is the most recent MT technique based on machine learning with a special neural network (NN) framework called Encoder-Decoder architecture. Fig. 6 shows the basic structure of the NMT model. NMT uses vector representations for words and sequence-model of the input sentence to generate TL words sequentially with encoders and decoders in the core [19]. The input words are first encoded in a one-hot vector and passed through an embedded matrix and hidden layers. In the output layer, the decoder outcome is interpreted as a probability distribution. A softmax activation function is used to ensure proper probability distribution [20]. NMT is a data-driven approach where a NN model is trained with a parallel corpus of SL and TL.

NMT has emerged as a hopeful field in the MT system for showing better performance than other MT systems with different NN models. Early NMT models used a feed-forward NN to develop an MT model, which could not provide sufficiently good results [21]. In the recent NMT studies, different deep learning models [22], such as Recurrent NN (RNN) [23], convolutional NN (CNN) [24], multilayered long short-term memory (LSTM), are is used for encoding and decoding purposes [25]. The recently developed transformer model with many encoder-decoder layers is shown better translation performance [26].

Different techniques are also investigated to improve NMT performance. Normally, NMT uses a parallel corpus of SL and TL for training. However, Sennrich et al. [27] have investigated the use of monolingual data effectively applying back-translation. Back-translation is translating back to SL from TL, and it is a way to train the NMT model for better translation quality [28]. However, NMT has still shown lower performance for low resource words and in word alignment [29].

5) Hybrid Machine Translation (HMT): The aim of HMT methods is to achieve better MT performance overcoming distinct constraints of individual MT methods while integrating individual ones. Fig. 7 shows the building block of the HMT system, which may contain several individual MT models for translation. RBMT and EBMT were brought into HMT by pioneer researchers [30]. Bond and Shirai [31] developed an HMT system that uses the EBMT method but allows to use of RBMT where required. Schwenk et al. [32] used a Statistical Post Editing (SPE) system where SMT is used to correct the errors of RBMT. Several recent HMT methods also performed well for MT in different language pairs. Huang et al. [33] developed an HMT by combining NMT and RBMT: the system used the consistency of RBMT to balance the inadequacy of datasets for the NMT system. Banik et al. [34] proposed an HMT system with NMT and SMT for English-Hindi language pairs. Singh et al. [35] used NMT and RBMT to build an HMT system for Sanskrit to Hindi. Beyala et al. [36] tuned the transformer model's output with phrase-based SMT.

Fig. 3. Rule-based MT (RBMT) with Interlingua Representation.

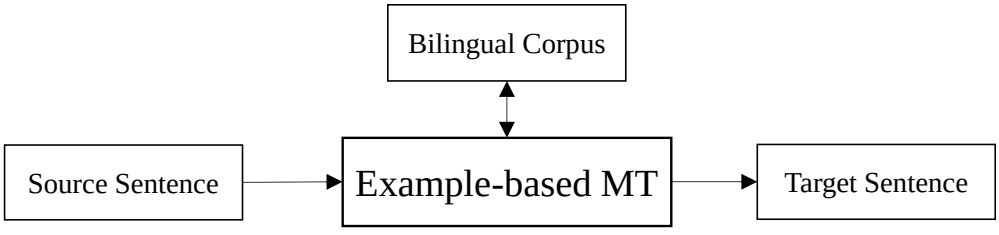


Fig. 4. Basic Example-based MT (EBMT) Method.

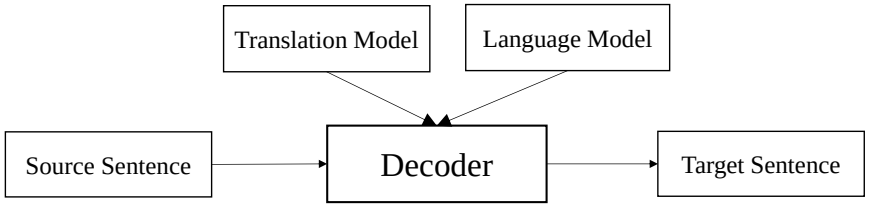


Fig. 5. Basic Statistical MT (SMT) Method.

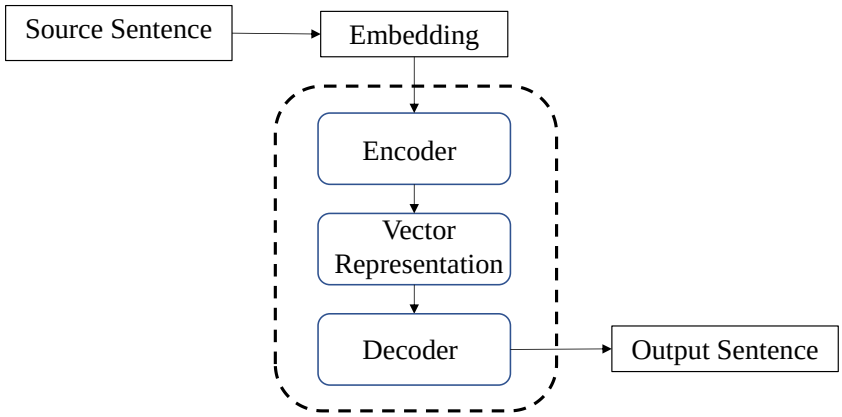


Fig. 6. Basic Neural MT (NMT) Method.

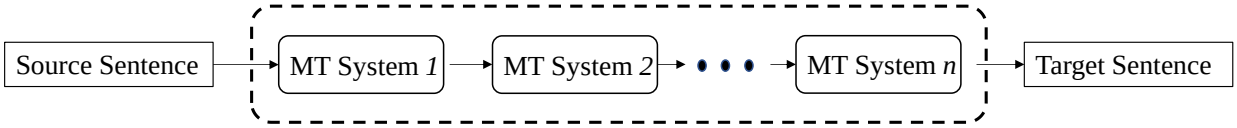


Fig. 7. Basic Hybrid MT (HMT) Method.

B. Bangla Language and Its Significance

Bangla, belong to the Indo-European language family, is a major language in the Indian subcontinent and the main language of Bangladesh; and Bangla is bound by different kinds of languages like Oriya, Assamese, etc. The language has come modern phase through a metamorphosis as the territory was under the rule of various administrations [37] for a long time. The basic sentence structure for the Bangla language is *subject + object + verb* in general. As an example, 'আমি সকালে ঘুম থেকে উঠি' for 'I wake up in the morning'. Whereas, English is in the West-Germanic language family [38]. The basic structure of English is *subject + verb + object*; as an example: I wake up in the morning. In the case of adverb, in Bangla sentences, an adverb comes before the verb like "সে আস্তে দৌড়ায়", which can be translated in English "He runs slowly" where adverbs usually come after verb. In Bangla, both masculine and feminine gender share the same form of pronoun like "সে/তিনি" whereas in English the third person singular number pronoun differs in terms of gender such as he/she, him/her, etc.

The form of verb differs in terms of space, time, and person in Bangla language; examples are "তুই এখান থেকে যা", " আপনি এখান থেকে যান", and "তুমি এখান থেকে যাও". These three sentences are translated as "You get out of here" in English. Sometimes the exact meaning of the English word is not used. For an example, "We are playing in the field" which means in Bangla "আমরা মাঠে খেলছি". Here "are" means "hoy/hoi" which is not used in Bangla. So "are playing" is sometimes considered as a verb phrase. Moreover, instead of the preposition, Bivokti (i.e., inflection) is used, mainly joining a letter with a word to relate with other words. In the previous example, "in the field" means "Maathe" in Bangla. "Field" means "Maath" but the word "in", which is a preposition in English, is considered as "e" in Bangla. Bangla has various kinds of inflection in sentences and varies in phonology, also depending on regions.

C. MT Performance Measurement

Performance measurements in the MT system play an indispensable role in determining the efficacy of the existing system and the requirement of optimization. Regarding MT system evaluation, human evaluation and several other matrices are available [39].

Human evaluation is considered as a baseline for MT evaluation. Adequacy and fluency are the most common methodologies of human evaluation, which are measured on each sentence in the output, allotting points from one to five according to translation quality [40]. Adequacy refers to how much meaning and information have been manifested in the source and target languages. It needs the judge to be bilingual in both source and target languages. Fluency indicates how fluent the translation is, and the judge needs to be fluent in the target language. Human evaluation is a very cumbersome and time-consuming process to judge translation quality sentence by sentence. Therefore, automatic evaluation matric is used nowadays instead of human evaluation, and several such methods are briefly described below.

Bilingual Evaluation Understudy (BLEU) is currently the most popular automatic evaluation metric for MT. BLEU considers multiple references, each of which may use a different word choice to translate the same source word. The base of the BLEU metric is a precision measure [41]. At first, a modified n-gram is calculated by counting the number of n-grams or word sequences in the candidate sentences (i.e., system output) alongside the reference sentences. Then the candidate counts are clipped by their corresponding reference maximum value. These clipped n-grams are then summed and divided by the total number of candidate n-grams [41]. Through this step, the modified precision score (p_n) is calculated.

$$p_n = \frac{\sum_{C\in\{Candidates\}}\sum_{n-gram\in C} Countclip(n-gram)}{\sum_{C'\in\{Candidates\}}\sum_{n-gram'\in C'} Count(n-gram')} \quad (1)$$

This result is multiplied by an exponential brevity penalty factor where a high-scoring candidate translation must now match the reference translations in length, word choice, and word order. Therefore, the next step is to calculate BLEU Brevity Penalty (BP) factor.

$$BP = \begin{cases} 1, if\ c > r \\ e^{1-\frac{r}{c}}, if\ c \leq r \end{cases}, \quad (2)$$

here c is the length of candidate translation, and r is the length of reference translation. Finally, the BLEU score is the geometric mean of the precision scores and is calculated using Eq. (3).

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n\ log\ p_n) \quad (3)$$

The MT score NIST comes from the name National Institute of Standards and Technology, and it is an improved version of BLEU. Where BLEU counts all the n-grams equally, NIST takes into account the informativeness of n-grams on the basis of frequency of occurrence [42]. Besides, NIST uses the arithmetic mean of n-gram counts, but BLEU uses the geometric mean of n-gram count. It also tries to minimize the unwanted effects of the brevity penalty factor by BLEU [42], [43]. However, at first, the information weights are calculated by n-grams counts over a set of reference translations to calculate the NIST score.

$$Info(w_1 \dots w_n) = \log_2 \left(\frac{the\ number\ of\ occurences\ of\ w_1 \dots w_{n-1}}{the\ number\ of\ occurences\ of\ w_1 \dots w_n}\right) \quad (4)$$

Finally, the NIST score is calculated by Eq. (5).

$$NIST = \sum_{n=1}^{N} \left\{\frac{\sum_{all\ w_1 \dots w_2\ that\ co-occur\ Info(w_1 \dots w_n)}}{\sum_{all\ w_1 \dots w_n\ in\ system\ output} 1}\right\} * exp\left\{\beta \log^2 \left[min\left(\frac{L_{sys}}{\overline{L_{ref}}}, 1\right)\right]\right\}, \quad (5)$$

where β is chosen to make the brevity penalty factor = 0.5 when the number of words in the system output is two-thirds of the average number of words in the reference translation.

N=5 means this formula works with five words at a time.

$\overline{L_{ref}}$ denotes the average number of words in a reference translation; averaged over all the reference translations.

L_{sys} denotes the number of words in the translation which have been scored.

Translation Error Rate (TER) is defined as the minimum number of edits required to change a system output to reference translation [44]. It was designed to reduce the human effort for evaluating an MT method [45]. The general equation is given below:

$$TER = \frac{Number\ of\ edits}{Average\ number\ of\ reference\ words} \quad (6)$$

here possible edits can be counted as insertion, deletion, and substitution of single words as well as shifts of word sequences.

III. Review of Bangla MT Systems

Several MT systems developed on the Bangla language in the last two decades. The available Bangla studies are Bangla to English (B2E) or English to Bangla (E2B) with different Bangla-English corpora. A few studies are for both B2E and E2B. Bangla-English corpus is an important element to Bangla MT, and therefore, an overview of Bangla corpora is given first. Then existing Bangla MT systems are described briefly in different MT categories.

A. Bangla Corpus

Several Bangla-English parallel corpora are prepared by different research groups and are publicly available for anyone to use. Table I summarizes prominent Bangla-English corpora mentioning significant attributes of individuals. The corpora are varied in sample sizes. The largest corpus is the Indic Languages Multilingual Parallel corpus (ILMPC) consists of 338500 sentences. On the other hand, the small-sized corpora Penn Treebank (PTB) and AmaderCAT consist of 1313 and 1782 sentences, respectively. In several cases, the available sentences are partitioned into training, validation, and test sets. The training set is to train a model, and the test set is dedicated to the final evaluation of the trained model. The validation set samples may use to evaluate the intermediate performance of a model during training. The last column of Table I referred to several studies that used a particular corpus. Based on recent studies, SUPara corpus [46] is the most popular. The corpus holds quite clean 71861sentences having 244539 words in English and 202866 words in Bengali [46].

B. Review of Bangla RBMT Methods

Using RBMT, based on linguistic information and rule production, diverse techniques for B2E and E2B MT have been investigated for rules generation, including fuzzy rules [47], [48], context-sensitive grammar (CSG) rules [49]–[51], etc. Under the umbrella of RBMT, Rahman et al. [52] utilized morphological analysis in finding the root words from the input Bangla sentences. After matching the Bangla grammar, corresponding English grammar is identified; the input sentence is then rearranged according to it. The final output is the English translation of the corresponding Bangla words with the help of a dictionary. They considered only a few types of sentences. The method seems quite efficient but needs a lot of knowledge about both languages and engagement of the dictionary. Chowdhury [53] projected a system where Bangla sentences are read from left to right, and corresponding English words are generated using a dictionary and the context of the Bangla sentence. In addition to word generation, a set of grammatical rules are used to analyze the source sentence properly.

TABLE I. BENCHMARK BANGLA-ENGLISH CORPORA SUMMARY

Sl.	Corpus Name [Ref.]	Sample Size (Training/Val./ Test Set)	Corpus Data Link	Significance	Study / Works with the Corpus
1	Enabling Minority Language Engineering (EMILLE) [97]	26287 (25287/500/500)	http://catalog.elra.info/en-us/repository/browse/ELRA-W0037/	-	[78]
2	KDE4 [98]	35365 (33365/1000/1000)	https://opus.nlpl.eu/KDE4-v2.php	i) Currently contains words of 60 different languages ii) Already sentence aligned	[78]
3	SUPara [46]	71861 (70861/500/500)	https://ieee-dataport.org/documents/supara08m-balanced-english-bangla-parallel-corpus	i) First free English-Bangla Parallel corpus ii) Balanced and comprehensible	[7], [74], [76], [77], [82], [83]
4	Global Voices [98]	1031725	https://opus.nlpl.eu/GlobalVoices-v2018q4.php	Contains non-printable characters (e.g., Arabic)	[61], [74], [77]
5	Indic Languages Multilingual Parallel Corpus (ILMPC) [99]	338500 (337000/500/1000)	http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/index.2020.html	Consists of 7 parallel languages	[7], [76]
6	Six Indian Parallel Corpora (SIPC) [100]	219140 (20000/914/1000)	https://github.com/joshua-decoder/indian-parallel-corpora	i) Consists of 6 languages ii) Sentences are collected from internet documents	[7], [76]
7	Penn Treebank Bangla-English (PTB) [76]	1313	https://panl10n.net/ [Original source link, not accessible]	Multilingual parallel corpus	[7], [76]
8	AmaderCAT [101]	1782	https://github.com/AridHasan/Data-Collection-System-for-Machine-Translation/tree/master/data	i) A collaborative platform ii) Sentences are collected from newspapers	[7]
9	Linguistic Data Consortium [102]	12600 (11000/600/1000)	https://www.ldc.upenn.edu/	-	[70], [72]

Lexical analysis is important in RBMT as the attributes related to sentences in English and Bangla can be known through it, which are important in the next phases. The customized process can be used in the phase of rule generation. Alam et al. [54] used a bilingual lexicon that stores information and helps place words and error checking. After semantic analysis, they have categorized the sentences according to the subject, object, and verb, leading to the generation of Bangla sentences. However, it cannot identify gerund (i.e., –ing form of a verb), more than one subject or object, and several grammatical issues. Francisca et al. [47] investigated an RBMT process that accepts an English sentence as input where the lexical analyzer is used to generate the class of the sentence by utilizing the information of the word from the dictionary. The generalization of sentences is used to find outmatched fuzzy rules for English sentences. The rules may be matched partially or fully. Later, the dictionary is used to find the corresponding Bangla words, leading to the next step to reconstruct the Bangla sentence, depending on the related rules for Bangla sentences. The process seems compelling, but they have not covered all kinds of sentences. Mukta et al. [48] proposed a model similar to [47], but it is based on tense and phrase. This system emphasizes English grammar, verbs, prepositions, inflection, and other grammatical rules of Bangla. English word translation to Bangla takes place with the help of a dictionary, and a morphological analyzer analyzes these words for the target language. After the reconstruction of Bangla sentences by proper production rules, the system delivers the output. Prepositions (e.g., to, in, etc.) do not have any definite meaning in Bangla, and the auxiliary verbs (e.g., are, is) have some meaning but are implicitly used in Bangla. Therefore, they have assumed preposition and object in one phrase and an auxiliary verb and main verb in one phrase for simplicity. The system seems to work well in comparison with the Google translator.

Anwar et al. [55] used CSG rules in their B2E RBMT system, where after tokenizing, a token is searched in a lexicon, and if found, it is matched by the Bangla grammatical rules. The token is tagged by appropriate parts of speech if matching is found. After that, a parser is used to generate a parse tree for the input string. Finally, the corresponding English sentence is generated by the NLP conversion unit through the help of a corpus. They used the basic bi-gram model as the language model and created basic English sentences by replacing the Bangla words with English words from the lexicon. Using 28 basic production rules, much importance is given to the parts of speech, including simple, complex, and compound sentences. The system shows a remarkable accuracy (over 90 percent) while considered limited sentence types. Muntarina et al. [50] proposed their strategy based on tense-based rules using parse trees. They constructed a parse tree for input English language. Then it was converted into a Bangla parse tree based on production rules for both languages generated by syntactic and morphological analysis. The system uses the NLP conversion technique for conversion and lexicon, which helps simplify the knowledge on both languages and provides the target words for input text. They have considered input and output in the form of tense.

Arefin et al. [56] have designed an MT system that has given much importance to assertive, interrogative, and imperative sentences. They proposed a unique method named

"Transfer" where the conversion from Bangla parse tree to the English parse tree is organized. The system drives the parsing method based on 22 context-free grammar (CFG) rules for Bangla. The transfer method has a separate algorithm that uses 24 CFG rules to generate an English parse tree. The processing of the Bangla parse tree starts from the low-level nodes and goes up to the top level by analyzing. The process continues by creating a subtree of English sentences based on checking the grammatical rule generator module. This system showed higher accuracy than Google translator though the data set is small. Alamgir et al. [51] have depicted a model using the same CFG rule and generation of the parse tree for both languages. 31 CFG rules have been included in a table for Bangla parse tree and another 31 rules have been used to construct the English parse tree. They experimented on imperative, exclamatory, and optative sentences. This system also has higher precision than Google Translator on a limited data set.

Ashrafi et al. [49] used CFG, which helps to replace the tokenized words with the variable. A bilingual dictionary gives apposite information about morphological features along with the meaning of English words. CFG provides grammatical rules according to English and Bangla language structure. An intermittent parse tree is reorganized by Stimulate English Parse Tree module in the form of another parse tree to stimulate computational history. The output is available by substituting the English words with equivalent Bangla meaning as well as reordering the previous tree to get the actual parse tree by Bangla CFG rules. They have used CFG rules for both Bangla and English languages. Example for English, S → NP VP |NP VP ADV | …, NP → N | PN |CN | …, VP → MV | CV| AV CV | etc. For Bangla, S → NP VP | NP ADV VP |…, NP → N | PN | CN … etc. Where, S→ Sentence, NP→ Noun Phrase, VP→ Verb Phrase, ADV→ Adverb, PN → Pronoun, CN→ Complex Noun, MV→ Main Verb, CV→ Complex Verb, AV→ Auxiliary Verb. This architecture is very effective when the sentence falls into the rules made from the morphological analysis. The authors stated the method as Approximate Lexical Meaning Mapping (ALMM).

Anwar et al. [57] focused on structural and syntax analysis to generate grammatical rules in their B2E RBMT system. The system tokenizes Bangla words based on the lexicon and forms groups of the tokenized words according to grammatical rules using a parser. This information helps to create a parse tree to portray the syntactic structure of source sentences. Later they have used fuzzy logic to interpret the input Bangla sentences to convert them into English. Finally, they enumerated the probability of each word (termed as Fuzzy membership) to come first and next in English sentences. They gave much importance to finite verbs, whereas other parts of speech and phrases have contributions to form a sentence. The system needs the help of an aligned bilingual corpus. Fuzzy logic has been used further by the model proposed by Anwar [58] in the conversion phase with a basic RBMT model to interpret the input Bangla sentences to output English sentences. In this model, 28 basic rules have been used to parse a sentence and generate the parse tree. Mainly focused on establishing and using grammatical rules, three main types of sentences, simple, complex and compound, have been used in the experiment.

Rabbani et al. [59] proposed an E2B RBMT approach, which transforms different forms of English sentences (like active, passive, assertive, interrogative, imperative, exclamatory, simple, complex, and compound) into some simplified forms, i.e., subject + verb + object. After identifying the principal verb from the English sentence, it binds the rest of the parts of speech as subject and object. Bangla output sentences are generated by the translation of English words of the newly structured English sentences. Recently, Haque & Hasan [60] proposed an algorithm that takes person, verb root, and tense as arguments and finds what should be the appropriate verb in the sentence, which later applied to E2B RBMT system architecture.

Islam et al. [61] have used the tagging of a token as word, number, person, etc., in their RBMT method to identify the structure of Bangla sentences. Later, it motivates word-by-word translation from Bangla to English and applies necessary suffix and grammatical rules that lead to final output. They investigated three approaches to tackle different forms of verb representation in Bangla sentences. In name identification, they have tried to handle unknown words and names. The names of persons are identified by a method emphasizing with tags.

Table II summarizes the above discussed Bangla RBMT studies mentioning achieved test set accuracies. Notably, most of the studies are related to B2E translation. In few cases, few parameters (e.g., accuracy) are not reported clearly in the corresponding articles mentioned in the comments. Among the B2E studies, Anwar [58] achieved the best accuracy for sample, complex, and compound sentences with 95%, 80%, and 80%, respectively, with their self-prepared dataset. On the other hand, 100% test set accuracy is reported for E2B by Ashrafi et al. [49], although information about the dataset is not provided clearly.

C. Review of Bangla EBMT Methods

Only a few Bangla MT studies are available with EBMT. Dandapat et al. [62] investigated a translation memory (TM) based EBMT architecture. They built two TMs: one is based on phrase pairs alignment (PT), and another is based on word aligned file from source to target language (LT), where these two TMs are used for translation of unmatched parts. At first, the system finds the closest match in the input sentences to be translated and then links with equivalent translation. Later, inapposite fragments are detected, and the main translation is found in the recombination step by adding, substituting, and rearranging fragmented translations. They conducted their experiments on different systems: Basic EBMT, EBMT+ TM (PT) in the recombination step, EBMT+TM (PT+LT), EBMT+SMT in the recombination step, and SMT.

TABLE II. TEST SET PERFORMANCE COMPARISON AMONG BANGLA RBMT METHODS FOR BANGLA TO ENGLISH (B2E) AND/OR ENGLISH TO BANGLA (E2B)

Sl.	Work Ref.: Author, Year [Ref.]	Corpus / Dataset	Test Set Size	Model Used	Accuracy on Test Set	Comments
1	M. M. Anwar et al., 2009 [55]	Self-Prepared	450 (Simple Sentence)	RBMT with context sensitive grammar rules	93.33% (B2E)	
		Self-Prepared	540 (Complex Sentence)		92.6% (B2E)	
		Self-Prepared	420 (Compound Sentence)		91.67% (B2E)	
2	M. Anwar et al., 2010 [57]	Self-Prepared	Less than 900 (Simple Sentence)	RBMT with fuzzy logic	About 90% (B2E)	Data size and outcomes are not mentioned precisely.
		Self-Prepared	Less than 800 (Complex Sentence)		About 80% (B2E)	
		Self-Prepared	About 550 (Compound Sentence)		About 80% (B2E)	
3	Rahman et al., 2010 [52]	Self-Prepared	6	RBMT with Morphological approach	-	Statistical method used for performance measure and accuracy not mentioned
4	Francisca et al., 2011 [47]	Self-Prepared	79/-/ 27	RBMT with fuzzy rules	-	Performance is not mentioned
5	Alam et al., 2011 [54]	-	-	RBMT with modified approach	-	Statistical method used for performance measure and accuracy not mentioned
6	Chowdhury, 2013 [53]	-	-	RBMT with Parts of Speech Tagging	-	Performance is not mentioned
7	Ashrafi et al., 2013 [49]	Self-Prepared	Not Stated	RBMT with Approximate Lexical Meaning Mapping (ALMM)	100% (E2B)	Experiment outcomes are not available
8	Muntarina et al., 2013 [50]	Self-Prepared	600	RBMT with Tense Based Approach	86.16% (E2B)	
9	Arefin et al., 2015 [56]	Self-Prepared	420	RBMT with Context-Sensitive Grammar	83.09% (B2E)	Assertive, Interrogative and Imperative sentences are considered
10	Alamgir et al., 2016 [51]	Self-Prepared	400	RBMT with Context Sensitive Grammar	81.5% (B2E)	Imperative, Optative and Exclamatory sentences are considered
11	M. Anwar, 2018 [58]	Self-Prepared	Less than 900 (Simple Sentence)	RBMT with Fuzzy logic	About 95% (B2E)	Accurate data and result are not shown
		Self-Prepared	Less than 800 Complex Sentence		About 80% (B2E)	
		Self-Prepared	About 550 (Compound Sentence)		About 80% (B2E)	
12	Mukta et al., 2019 [48]	Self-Prepared	1113	Phrase-based RBMT	Mismatch 50 (E2B)	

Khan et al. [63] have proposed an E2B model in EBMT using WordNet [64] and International Phonetic Alphabet (IPA) [65] based transliteration. The system begins with taking English sentences as input and then parsing them into chunks which are similar to tokenization in RBMT. The chunks are matched with an example-based English-Bangla parallel corpus by a matching algorithm whose outcomes are Chunk-String Templates (CSTs) and unknown words. CSTs are the combination of chunks in English and Bangla languages and the information of alignment of words. The translation of unknown words uses a transliteration process, a procedure of converting a text or word from one language to another language. It is useful for people to pronounce foreign words. Lastly, the output is produced with the help of WordNet and the generation rules. Unknown word handling is the specialty of the model. For this purpose, the model first tries to find semantically related words in WordNet and the closest meaning of the words from the dictionary. If the process does not work, the system needs the help of IPA-based transliteration and Akkhor Bangla Software. Overall translation quality of the model seems good but some inconsistencies have been found using WordNet and due to the small corpus. Salam et al. [66] proposed another EBMT method where ontology is used to improve the quality. The model is similar to [63] but some changes made this model unique. The unknown words are searched in WordNet using synonyms, antonyms, and hypernyms, which develop a vast option to increase the quality.

Table III summarizes the above discussed Bangla EBMT studies mentioning achieved test set performance scores. Notably, the three studies mentioned above are related to E2B translation with self-prepared corpora. Based on the achieved BLEU scores, Dandapat et al. [62] are achieved the best among the three mentioned methods with the value of 57.56.

TABLE III. TEST SET PERFORMANCE COMPARISON AMONG BANGLA EBMT METHODS FOR ENGLISH TO BANGLA (E2B). [N.B.: NO EMBT STUDY ON BANGLA TO ENGLISH (B2E)]

Sl.	Work Ref.: Author, Year [Ref.]	Corpus / Dataset	Sample Size: Train./Val./ Test Set	Model Used	Performance Score on Test Set			Comments
					BLEU	NIST	Accuracy	
1	Dandapat et al., 2010 [62]	Self-Prepared Medical data	381/-/41	EBMT with translation memory (Probable Target)	57.47(E2B)	5.92(E2B)	-	
				EBMT with translation memory (Probable Target + Lexical Table)	57.56(E2B)	6.00(E2B)		
				EBMT with SMT	52.01(E2B)	5.51(E2B)		
2	Khan et al., 2013 [63]	Self-Prepared	2000 /-/336	EBMT with unknown word translation mechanism	-	-	41.33%(E2B)	Simple, complex and mixed with various phenomena for testing
3	Salam et al., 2017 [66]	Self-Prepared	2000 /-/336	EBMT with CSTs	-	-	38.69%(E2B)	Simple, complex and mixed with various phenomena for testing
				EBMT with CSTs and unknown word translation mechanism	-	-	36.90%(E2B)	

D. Review of Bangla SMT Methods

SMT is a well-known data-driven approach and SMT models for Bangla-English MT studies are developed in several studies. Uddin et al. [67] have proposed an SMT architecture based on different parameters. Alongside the established parameters like for Bangla and English sentence length, for the various probability of occurrences, etc., they have created new parameters based on few complex sentences: Bi-occurred parameter, Bi-distribution parameter, Absent-Distribution parameter, and Subject-check parameter. The Bi-occurred parameter is for the doubly occurred Bangla verbs. The Bi-distribution parameter works with the Bi-occurred parameter and estimates the appropriate position of English translation for the doubly occurred Bangla verbs. To translate Bangla sentences, sometimes extra words are needed to add in English that are implicit and not connected to any Bangla words. The Absent-Distribution parameter handles this type of problem. The Subject-check parameter handles multiple subjects.

A phrase-based SMT for E2B is proposed by Islam et al. [68], where a 5-gram language model (i.e., five words at a time) has been furnished with different corpora and used in the baseline system along with training data made by the aligner. After finding some English words in the Bangla translation and comparatively low results in the baseline system, they operated a cleaning process of corpora with a sentence alignment process. They achieved improvement in the development after executing this new translation system with an 8-gram language model. They also specialized in preposition handling by assigning inflections to the noun in Bangla (applicable to Bangla corpus) and a transliteration module to identify unknown words. They combined the preposition handling module, transliteration, and new translation system; the combined system outperforms other methods on various dictionaries. They used MOSES, GIZA++, MERT, and SRILM [69] toolkits to construct the whole system.

Roy and Popowich [70] presented a phrase-based B2E SMT with a unique transliteration method. They have designed a module that can handle prepositions and Bangla compound words. Their transliteration module at first finds the untranslated words. Later the best-matched translation is found with the help of a monolingual English dictionary. The preposition handling module, at first, removes the inflections of the Bangla words. Later, appropriate English words with prepositions are applied with the help of the bilingual dictionary. Bangla compound words are handled by a splitting algorithm proposed by Koehn & Knight [71]. In another study, Roy and Popowich [72] applied a different word-reordering approach to the phrase-based SMT model. As an automatic word-reordering approach, they used an algorithm proposed by Crego & Mari [73].

Mumin et al. [74] presented a phrase-based SMT model (called Shu-torjoma) for both B2E and E2B. They used various monolingual, bilingual and parallel corpus to train the model. The preprocessor module processes data into a favorable format at the next step, including punctuation and lexical normalization, tokenization, morphological segmentation, syntactical reordering, etc. The preprocessed resources are then trained and tuned to create various statistical models: 5-gram language model, translation mode using GIZA++, Lexicalized Reordering Model, etc., to refine the system. Then the translated texts are found by MOSES decoder. On the other hand, Rabbani et al. [75] proposed a hybrid phrase-based E2B MT using the concept of RBMT and SMT. The model finds the principal verb from any kind of sentence and then converts it into the simplest form.

Dandapat and Lewis [8] developed an English-Bangla general-purpose domain and worked on both SMT and NMT fields. Using different training sets, they used phrasal (for B2E and vice versa) and Treelet (E2B) translation models and developed a word segmentation model to handle unknown words. They developed a word breaker to handle out of vocabulary words where they have used a linguistic suffix list for partitioning inputs and parallel corpora to rank the

partitioned candidates based on frequency. They also used the transliteration module to transliterate foreign words.

Hasan et al. [76] showed a comparative study between SMT and NMT. They used SRILM as a language model and MOSES decoder to train their SMT system and gather different corpora. They also covered 3-gram and 5-gram language models under training. Al Mumin et al. [77] also depicted a comparative result between SMT and NMT where their preprocessed (correction of spelling, pronunciation normalization, etc.) data has been used in the SMT system using MOSES. The whole architecture of this SMT is much like their Shu-torjoma [74].

Table IV summarizes the above discussed Bangla SMT studies mentioning achieved test set performance scores. Several studies reported performance for both B2E and E2B translations; others are for B2E or E2B. For B2E, Al Mumin et al., 2019 [74] achieved the best accuracy showing a BLEU score of 17.43 with SUPara corpus. On the other hand, the best BLEU score for E2B was 23.30, achieved by Islam et al. [78] with KDE4 corpus. It is also notable from studies with both B2E and E2B that the performance score is slightly different between B2E and E2B.

E. Review of Bangla NMT Methods

Nowadays, NMT is the most studied method with different machine learning and deep learning methods in different languages, and studies with NMT are also popular in Bangla MT. Dandapat and Lewis [8] developed an NMT model combining with an SMT model discussed in the previous section. The NMT system using only conventional bidirectional RNN failed to exceed the score of SMT. They used Phrasal [79] (for B2E and vice versa) and Treelet [80] (for E2B) translation models using different training sets. They also developed a word segmentation model to handle unknown words. Finally, the introduction of early stopping, byte per encoding (BPE) and backpropagated synthetic data enhanced the performance of the NMT model. It outperformed significantly on low-resource data like Bangla.

Hasan et al. [7] used Bidirectional LSTM (BiLSTM) and transformer, the two popular deep learning methods, for B2E NMT. Their preprocessing includes tokenization of English and Bangla sentences, normalization of punctuation, limitation of sentences length and identification of abbreviation, Email, URLs, etc. They have trained their models and created multiple experimental settings on different schemes like using one corpus and multiple corpora. In comparison between the methods, the BiLSTM-based model is found better than the transformer. Hasan et al. [76], in another study, where BiLSTM based methods are compared with SMT. They used different corpora and identified the best performances of each model with a particular corpus. Their results show that the NMT model offers a better result than the SMT model.

Mumin et al. [77] investigated the attention-based model and Byte Pair Encoding (BPE) in their NMT model. They separately examined the basic attention-based model and attention-based model with BPE for both B2E and E2B. It is shown that the attention-based model with BPE gives comparatively better results than other approaches, e.g., SMT.

TABLE IV. TEST SET PERFORMANCE COMPARISON AMONG BANGLA SMT METHODS FOR BANGLA TO ENGLISH (B2E) AND/OR ENGLISH TO BANGLA (E2B)

Sl.	Work Ref.: Author, Year [Ref.]	Corpus / Dataset	Sample Size: Train./Val./ Test Set	Model Used	Performance Score on Test Set			Comments
					BLEU	NIST	TER	
1	Uddin et al., 2005 [67]	Not Stated		Baseline SMT	-	-	-	No experiment is conducted
2	M. Z. Islam et al., 2010 [78]	EMILLE	25287/500/500	SMT with Final combined system	5.70(E2B)	3.16(E2B)	0.83(E2B)	
		KDE4	33365/1000/1000		23.30(E2B)	5.18(E2B)	0.63(E2B)	
		EMILLE+KDE4	58652/1500/1500		11.70(E2B)	4.27(E2B)	0.76(E2B)	
3	Roy & Popowich, 2010 [70]	Linguistic Data Consortium	11000/600/1000	Phrase-based SMT with Transliteration	9.1 (B2E)	-	-	
4	Roy & Popowich, 2010 [72]	Linguistic Data Consortium	11000/600/1000	SMT with Lexicalized reordering	8.2 (B2E)	-	-	
				SMT with Manual reordering	8.4 (B2E)	-	-	
				SMT with Automatic reordering	9.3 (B2E)	-	-	
5	Dandapat & Lewis, 2018 [8]	Websites, Webdunia, WMT	976634/3500/6000	Baseline SMT	16.56(B2E) 7.41(E2B)	-	-	
6	Hasan et al., 2019 [76]	ILMPC, SIPC, PTB, SUPara	346845/500/956	SMT+ 3-gram Language Model	14.61(B2E)	-	-	Training set is merged but ILMPC is used for development and test set
				SMT+ 5-gram Language Model	14.82(B2E)	-	-	
7	Al Mumin et al., 2019 [74]	SUPara, Global Voices	197338/500/500	Phrase-based SMT	17.43(B2E) 15.27(E2B)	5.76(B2E) 5.13(E2B)	67.94(B2E) 71.93(E2B)	Training data sets are combined; SUPara for test and development

Recently, Siddique et al. [6] proposed architecture for E2B MT based on RNN. Their process starts with the preprocessing and tokenization of the English and Bangla sentences according to frequency. Later with the help of a context vector, the English and Bangla sentences are mapped where embedded RNN, both GRU and LSTM are used, which is similar to the attention model. The model calculates the error with loss function to improve the model through backpropagation. They also identified that using a large number of parallel sentences in the corpus may improve the result.

Akter et al. [81] investigated an NMT method using pre-trained embedding and synthetic monolingual data for E2B. They considered two modifications with the baseline NMT: a pre-trained word embedding model for source and target languages and a synthetic monolingual data addition model. NMT with a pre-trained word-embedding model reduces workload, brings outside model information, and decreases the number of parameters. It has shown improvement in the BLEU score. The addition of synthetic monolingual data in the NMT model associated with back translation helps to handle out of vocabulary words. This model showed a relatively better BLEU score for E2B over the other existing methods.

The most recent NMT models for Bangla MT are [82] and [83]. Dhar et al. [82] investigated a transformer-based NMT model for B2E MT where different parameters (especially, number of heads) are tuned for a better outcome. The model is tested on a benchmark of Bangla-English corpus, which outperformed some other MT methods. On the other hand, Roy et al. [83] considered BiLSTM in their MT study for both B2E and E2B. Attention mechanism with BiLSTM model and a special data augmentation mechanism, called Back Translation (BT), are the significant features of the proposed model. The model outperformed the existing prominent models for B2E MT while tested on a benchmark corpus.

Table V summarizes the above discussed Bangla NMT studies mentioning achieved test set performance scores. Notably, among the studies that performed B2E translation, only a few recent studies performed both B2E and E2B translation. For B2E, the most recent study by Roy et al. [83] achieved the best performance showing a BLEU score of 23.12. They used data augmentation with a back-translation mechanism considering GlobalVoices corpus with SUPara corpus. On the other hand, the best BLEU score for E2B 27.46 was achieved by Akter et al. [81] with synthetic monolingual data in the NMT model.

F. Review of Bangla HMT Methods

There are a few Bangla studies with HMT. Among the existing HMT studies, the E2B method called ANUBAAD [84] is the pioneering one which is a hybrid MT system using EBMT and RBMT explicitly. ANUBAAD considered noun phrase, adverbial phrase, and verb phrase. The system morphologically analyzes the input sentences and defines some formal grammars. Noun phrases and adverbial phrases are translated through EMBT with a template matching module, whereas verb phrases are translated using the RBMT approach.

Rabbani et al. [85] investigated the principal verb-based MT (called PVBMT), which is a hybrid of RBMT and SMT, belongs to the HMT paradigm. After passing through lexical analysis, the words that are tagged in the previous step are bound. In the next step, PVBMT determines the verbs in a sentence that works with three types of verbs within a sentence: auxiliary verb (AV), finite verb (FV), and non-finite verb (NV). If a sentence has more than one verb, then PVBMT creates different sets for different types of verbs according to their meanings and positions. Then PVBMT defines the Bangla sentence structure corresponding to the English sentence and generates the output. They transformed different English sentences into the simplest forms, e.g., Subject+Verb+object, and then translated the sentences into Bangla.

Islam et al. [61] recently investigated B2E MT blending RBMT with data-driven MT (i.e., SMT and NMT). Specifically, first, they implemented some basic grammatical rules that identified names as subjects and optimized Bengali verbs in their RBMT. Next, they integrated RBMT with each of SMT and NMT separately using different approaches. Besides, they performed rigorous experiments over several datasets to provide a comparison among the approaches in terms of translation accuracy, time complexity and space complexity. They also discussed how their blending approaches could be reused for other low-resource languages.

Table VI summarizes the above discussed Bangla HMT studies mentioning achieved test set performance scores. Notably, three studies in the table are with self-prepared datasets. Based on the achieved BLEU score, the method by Islam et al. [61] is the best, showing a score of 18.73.

TABLE V. TEST SET PERFORMANCE COMPARISON AMONG BANGLA NMT METHODS FOR BANGLA TO ENGLISH (B2E) AND/OR ENGLISH TO BANGLA (E2B)

Sl.	Work Ref.: Author, Year [Ref.]	Corpus / Dataset	Sample Size (Train./Val./ Test Set)	Model Used	Performance Score on Test Set			Comments
					BLEU	NIST	TER	
1	Dandapat & Lewis, 2018 [8]	Websites, Webdunia, WMT	976634/3500/6000	NMT with synthesis	20.23(B2E) 9.73(E2B)	-	-	
				NMT with BPE	20.64(B2E) 9.80(E2B)	-	-	
2	Hasan et al.,2019 [7]	ILMPC, SIPC, PTB, SUPara, AmaderCAT	419109/500/500	BiLSTM with Bangla and English Embeddings	19.24(B2E)	-	-	
		ILMPC, SIPC, PTB, SUPara, AmaderCAT	419109/500/500	BiLSTM with Bangla Embeddings	19.40(B2E)	-	-	
		ILMPC, SIPC, PTB, SUPara, AmaderCAT	419109/500/500	Transformer	18.99(B2E)	-	-	
		SUPara	70861/500/500	BiLSTM with Bangla and English Embeddings	19.98(B2E)	-	-	
3	Hasan et al., 2019 [76]	ILMPC, SIPC, PTB, SUPara	346845/500/956	BiLSTM with Bangla and English Embeddings	15.62(B2E)	-	-	Training set is merged but ILMPC is used for development and test set
		SUPara	70861/500/500		19.76(B2E)	-	-	
4	Al Mumin et al., 2019 [77]	SUPara, GlobalVoices	197338/500/500	BiGRU with Attention	22.38(B2E) 15.57(E2B)	5.98(B2E) 4.72(E2B)	59.88(B2E) 68.54(E2B)	
				BiGRU with Attention and BPE	22.68(B2E) 16.26(E2B)	6.07(B2E) 5.18(E2B)	60.09(B2E) 68.69(E2B)	
5	Siddique et al., 2020 [6]	Self-Prepared	4000	GRU and LSTM	-	-	-	Performance on test set is not mentioned
6	Akter et al., 2020 [81]	SUPara, Indic parallel, Open subtitles, OPUS Ububtu, OPUS Gnome, OPUS Tanzil	484131/2000/2000	NMT with pre-trained embedding	26.92(E2B)	-	-	
				NMT with synthetic monolingual data	27.46(E2B)	-	-	
7	Dhar et al., 2021 [82]	SUPara	70861/500/500	Transformer with optimal head and BPE	21.33(B2E)			
8	Roy et al., 2021 [83]	SUPara	70861/500/500	BiLSTM with Attention and BPE	22.88(B2E)			
		SUPara, GlobalVoices	115550 /500/500	BiLSTM with Attention,BPE and BT	23.12(B2E)			

TABLE VI. TEST SET PERFORMANCE COMPARISON AMONG BANGLA HMT METHODS FOR BANGLA TO ENGLISH (B2E) AND/OR ENGLISH TO BANGLA (E2B)

Sl.	Work Ref.: Author, Year (Ref.)	Corpus / Dataset	Sample Size: Train./Val./ Test Set	Model Used	Performance Score on Test Set		Comments
					BLEU	Accuracy	
1	Naskar et al., 2004 [84]	Not Stated	-	EBMT and RBMT	-	-	No experiment is conducted
2	Rabbani et al., 2016 [85]	Self-Prepared	9	RBMT with SMT and Principle verb-based approach	-	89.6% (semantic analysis) and 78.3% (syntactic analysis) (E2B)	
3	M. A. Islam et al., 2021[61]	Global Voices	1031725	NMT with RBMT	18.73 (B2E)	-	
				SMT with RBMT	18.02 (B2E)	-	

IV. SIGNIFICANCE OF THE PRESENT STUDY

A comprehensive review on a specific topic is important for the research community to get up-to-date information. Hence, one may get a guideline and/or motivation for further work(s) on it. Due to the language resource dependency, MT studies are scattered on a low resource language (e.g., Bangla), and it is necessary to discuss the studies categorically following a common strategic fashion. Although a few good reviews are available for low-resource languages like Thai [86]; but no such review studies are available for the Bangla language, according to the best of our knowledge. Although several Bangla review studies are available, all are very poor in area and scopes. The pioneer review work by Chowdhury [53] in 2013 considered only B2E RBMT studies emphasizing parts of speech tagging matter. The work by Chopra et al. [87] included only one Bangla SMT in their study. The most recent review by Andrabi and Wahid [88] emphasized Hindi and Urdu, and they considered only a few pioneer Bangla studies. Table VII shows the year-wise projection of Bangla-English MT studies with achieved performance scores summarizing the methods presented in Tables II-VI. It is noticeable from the tables that pioneer Bangla MT studies are with RBMT, and NMT has been explored recently with a relatively better translation score. Considering the importance of the Bangla language and its prospects in MT studies, this comprehensive review on Bangla MT is a timely study with the following significance.

TABLE VII. YEAR-WISE PROJECTION OF BANGLA -ENGLISH MT STUDIES WITH ACHIEVED PERFORMANCE SCORE

Year	RBMT (12)	EBMT (3)	SMT (7)	NMT (8)	HMT (3)
2004					[84] Naskar et al. (E2B)
2005			[67] Uddin et al. (B2E)		
2009	[55] Anwar et al.; Acc. 93.33% (B2E)				
2010	[57] Anwar et al.; Acc. 90% (B2E)	[62] Dandapat et al.; BLEU: 57.56 (E2B)	[78] Islam et al.; BLEU: 23.30 (E2B)		
	[52] Rahman et al; (B2E)		[70] Roy & Popowich; BLEU:9.1 (B2E)		
			[72] Roy & Popowich; BLEU:9.3 (B2E)		
2011	[47] Francisca et al. (E2B)				
	[54] Alam et al. (E2B)				
2013	[49] Ashrafi et al.; Acc. 100% (E2B)	[63] Khan et al.; Acc. 41.33% (E2B)			
	[50] Muntarina et al.; Acc. 86.16% (E2B)				
	[53] Chowdhury (B2E)				
2015	[56] Arefin et al.; Acc. 83.09% (B2E)				
2016	[51] Alamgir et al.; Acc. 81.5% (B2E)				[85] Rabbani et al.; Acc. 89.6% (E2B)
2017		[66] Salam et al. ; Acc. 38.69% (E2B)			
2018	[58] Anwar; Acc. 95% (B2E)		[8] Dandapat & Lewis; BLEU: 16.56 (B2E) & 7.41 (E2B)	[8] Dandapat & Lewis; BLEU: 20.64(B2E) & 9.80(E2B)	
2019	[48] Mukta et al;. (E2B)		[76] Hasan et al.; BLEU:14.82 (B2E)	[7] Hasan et al.; BLEU: 19.98 (B2E)	
			[74] Mumin et al.; BLEU: 7.43 (B2E) & 5.27(E2B)	[76] Hasan et al.; BLEU:19.76 (B2E)	
				[77] Mumin et al.; BLEU: 22.68(B2E) & 16.26(E2B)	
2020				[6] Siddique et al.; (B2E)	
				[81] Akter et al.; 27.46(E2B)	
2021				[82] Dhar et al. ; BLEU:21.33 (B2E)	[61] Islam et al.;BLEU:18.73 (B2E)
				[83] Roy et al.; BLEU: 23.12 (B2E)	

1) Basic ideas of different MT methods (RBMT, EBMT, SMT, and HMT) and performance measures of automatic MT are presented as background studies of the present Bangla MT review.

2) Overview of Bangla language and a brief description of available Bangla-English corpora are given.

3) Bangla MT studies are briefly described categorically; the achieved performances of the individual methods are compared in a tabular form.

V. FUTURE PROSPECTS OF BANGLA MT FROM THIS STUDY

This review streamlines the various aspects, techniques, and resources of Bangla MT studies comprehensively to motivate researchers and pave the way for further investigation in this area. It is observed that corpus-based data-driven approaches, especially, NMTs are shown to outperform other methods. Therefore, recent studies with NMT and hybrid methods with NMT might be a way to improve Bangla MT proficiency further. Resources deficiency, especially lack of rich corpus, is the main lagging to build an appropriate NMT model. Therefore, focus on resource development is necessary, although it requires government and non-government efforts. It is noticeable that the Government of Bangladesh has launched a large national project on Bangla language and corpus development for MT, an important component in MT studies [89]. Such efforts might boost Bangla MT studies; however, investigating

innovative modern techniques is also necessary for better performance. Another observation from the present study is that all the Bangla MT studies involve English (i.e., B2E and/or E2B). It is also timely demand to break the boundary of existing study and develop Bangla MT systems for other major languages (e.g., Arabic, Chinese, Japanese) considering global prospects of Bangla language in the coming future.

Recently developed MT methods that are found to be very effective for English and other languages pairs may also be practical approaches for Bangla, subject to appropriate incorporation of relevant linguistic or other features and tuning of parameters. Hence, investigation on the Bangla MT study may perform in different directions. Multiple attention layers, called deep attention, investigated by Zhang et al. [90] perform well for Chinese/Germany/France-English translation tasks. Incorporating such a mechanism with multiple attention layers, an attention-based Bangla NMT model can be developed to improve its performance efficiently. Gated recurrent unit (GRU) employment of [91] and parts of speech tagging of [92] in attention mechanism might also be useful to employ in Bangla MT. In the line of data augmentation, input denoising plus auxiliary decoder investigated in [93] and self-learning, training with synthetically generated data using monolingual a source language corpus, investigated in [94], are also intuitive to improve MT performance for a low-resource language like Bangla. Multi-source translation, an approach to exploit multiple inputs (e.g., in two different languages) to increase performance, and missing data management investigated by Nishimura et al. [95] might also be a way to achieve better Bangla MT performance. Gated recurrent unit (GRU), an advanced LSTM model, and its updated model [96] might perform well for Bangla MT. Moreover, recently developed HMT techniques, such as [34] [35] [33], might bring good motivation for better Bangla MT system development.

VI. Conclusion

In the global era of digitalization, MT studies are much more important than ever. Considering the limited Bangla MT studies despite being a major language, this paper reviewed prominent Bangla-English MT studies. Specifically, the basic MT methods (i.e., RBMT, EMBT, SMT, NMT, and HMT) are explained in short as background knowledge. Bangla MT studies under individual methods are described briefly, and achieved performances are presented in the tabular form in Tables II-VI. A year-wise projection of all the reviewed methods in Table VII gives a timeline hierarchy Bangla MT study. It is noticeable from the hierarchy view that pioneer Bangla MT studies are with RBMT and SMT methods, and the recently developed NMT methods outperformed the pioneer methods.

This study is expected to be a valuable resource and guideline for researchers interested in the Bangla MT system. The brief description of the available Bangla-English benchmark corpus (Table I) helps develop a new MT model. The prospects of the present study are summarized in a separate section (Section V), mentioning different points. At a glance, NMT has an opportunity to develop a better Bangla MT model with recently developed techniques such as various data augmentations. Moreover, it is time to take Bangla MT studies beyond the involvement of the English language and explore Bangla MT studies involving other languages such as Arabic, Chinese, and Japanese.

References

[1] Garg and M. Agarwal, “Machine translation: A literature review,” arXiv. 2018.

[2] T. Luong, I. Sutskever, Q. Le, O. Vinyals, and W. Zaremba, “Addressing the Rare Word Problem in Neural Machine Translation,” in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 11–19, doi: 10.3115/v1/P15-1002.

[3] S. Jean, K. Cho, R. Memisevic, and Y. Bengio, “On Using Very Large Target Vocabulary for Neural Machine Translation,” in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 1–10, doi: 10.3115/v1/P15-1001.

[4] M. Wang, L. Gong, W. Zhu, J. Xie, and C. Bian, “Tencent Neural Machine Translation Systems for WMT18,” in Proceedings of the Third Conference on Machine Translation (WMT), 2018, pp. 522–527, doi: 10.18653/v1/w18-6429.

[5] “Bengali language,” Wikipedia, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Bengali_language. [Accessed: 08-Jul-2021].

[6] S. Siddique, T. Ahmed, M. Rifayet Azam Talukder, and M. Mohsin Uddin, “English to Bangla Machine Translation Using Recurrent Neural Network,” Int. J. Futur. Comput. Commun., vol. 9, no. 2, pp. 46–51, Jun. 2020, doi: 10.18178/ijfcc.2020.9.2.564.

[7] M. A. Hasan, F. Alam, S. A. Chowdhury, and N. Khan, “Neural Machine Translation for the Bangla-English Language Pair,” in 2019 22nd International Conference on Computer and Information Technology (ICCIT), 2019, pp. 1–6, doi: 10.1109/ICCIT48885.2019.9038381.

[8] S. Dandapat and W. Lewis, “Training deployable general domain MT for a low resource language pair: English–Bangla,” in EAMT 2018 - Proceedings of the 21st Annual Conference of the European Association for Machine Translation, 2018, no. May, pp. 109–117.

[9] P. Bhattacharyya, Machine Translation. CRC Press, 2015.

[10] S. Sreelekha, “Statistical Vs Rule Based Machine Translation; A Case Study on Indian Language Perspective,” arXiv. 2017.

[11] M. D. Okpor, “Machine Translation Approaches: Issues and Challenges,” IJCSI Int. J. Comput. Sci. Issues, vol. 11, no. 5, pp. 159–165, 2014.

[12] L. E. Dostert, “The Georgetown-I.B.M. Experiment,” Mach. Transl. Lang. Fourteen Essays, 1955.

[13] W. J. Hutchins, “Machine Translation: A Brief History,” in Concise History of the Language Sciences, 1995.

[14] J. Slocum, “A survey of machine translation: its history, current status, and future prospects,” Comput. Linguist., vol. 11, no. 1, pp. 1–17, 1985.

[15] M. Nagao, “A Framework of a Mechanical Translation between Japanese and English by Analogy Principle,” in Readings in Machine Translation, The MIT Press, 2003, pp. 1–7.

[16] R. Ehab, E. Amer, and M. Gadallah, “English-Arabic Hybrid Machine Translation System using EBMT and Translation Memory,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, pp. 195–203, 2019, doi: 10.14569/IJACSA.2019.0100126.

[17] A. R. Babhulgaonkar and S. V. Bharad, “Statistical machine translation,” in 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), 2017, pp. 62–67, doi: 10.1109/ICISIM.2017.8122149.

[18] P. Koehn, Statistical Machine Translation. Cambridge: Cambridge University Press, 2012.

[19] N. Kalchbrenner and P. Blunsom, “Recurrent continuous translation models,” in EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2013.

[20] P. Koehn, Neural Machine Translation. Cambridge University Press, 2020.

[21] F. Stahlberg, "Neural machine translation: A review," Journal of Artificial Intelligence Research. 2020, doi: 10.1613/JAIR.1.12007.

[22] M. A. H. Akhand, Deep Learning Fundamentals - A Practical Approach to Understanding Deep Learning Methods. Dhaka: University Grants Commission of Bangladesh, 2021.

[23] A. Esan et al., "Development of a Recurrent Neural Network Model for English to Yorùbá Machine Translation," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 5, pp. 602–609, 2020, doi: 10.14569/IJACSA.2020.0110574.

[24] J. Gehring, M. Auli, D. Grangier, and Y. N. Dauphin, "A convolutional encoder model for neural machine translation," in ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 2017, doi: 10.18653/v1/P17-1012.

[25] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–15, 2015.

[26] X. Liu, K. Duh, L. Liu, and J. Gao, "Very deep transformers for neural machine translation," arXiv. 2020.

[27] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers, 2016, doi: 10.18653/v1/p16-1009.

[28] M. Graca, Y. Kim, J. Schamper, S. Khadivi, and H. Ney, "Generalizing back-translation in neural machine translation," arXiv. 2019, doi: 10.18653/v1/w19-5205.

[29] P. Koehn and R. Knowles, "Six challenges for neural machine translation," arXiv. 2017, doi: 10.18653/v1/w17-3204.

[30] F. Sánchez-Martínez, M. L. Forcada, and A. Way, "Hybrid rule-based--example-based MT: Feeding Apertium with sub-sentential translation units," Proc. 3rd Work. Example-Based Mach. Transl., 2009.

[31] F. Bond and S. Shirai, "A Hybrid Rule and Example-Based Method for Machine Translation," in Recent Advances in Example-Based Machine Translation. Text, Speech and Language Technology, 2003, pp. 211–224.

[32] H. Schwenk, S. Abdul-Rauf, L. Barrault, and J. Senellart, "SMT and SPE machine translation systems for WMT'09," 2009, doi: 10.3115/1626431.1626458.

[33] J.-X. Huang, K.-S. Lee, and Y.-K. Kim, "Hybrid Translation with Classification: Revisiting Rule-Based and Neural Machine Translation," Electronics, vol. 9, no. 2, p. 201, Jan. 2020, doi: 10.3390/electronics9020201.

[34] D. Banik, A. Ekbal, P. Bhattacharyya, and S. Bhattacharyya, "Assembling translations from multi-engine machine translation outputs," Appl. Soft Comput., vol. 78, pp. 230–239, May 2019, doi: 10.1016/j.asoc.2019.02.031.

[35] M. Singh, R. Kumar, and I. Chana, "Hybrid machine translation system using deep learning," ASM Sci. J., vol. 13, no. 2, pp. 31–45, 2020.

[36] V. L. Beyala, M. J., and P. Li, "Factored Phrase-based Statistical Machine Pre-training with Extended Transformers," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 9, pp. 51–59, 2020, doi: 10.14569/IJACSA.2020.0110907.

[37] R. Wagner, "Chatterji, S. K. The origin and development of the Bengali language," Indogermanische Forschungen, 2015, doi: 10.1515/if-1929-0163.

[38] D. Crystal and S. Potter, "English language," 2020. [Online]. Available: https://www.britannica.com/topic/English-language. [Accessed: 10-Feb-2021].

[39] M. S. Ali, A. Alatawi, B. Alsahafi, and N. Noorwali, "QUES: A Quality Estimation System of Arabic to English Translation," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 7, pp. 245–251, 2020, doi: 10.14569/IJACSA.2020.0110732.

[40] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "(Meta-) evaluation of machine translation," in Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07, 2007, pp. 136–158, doi: 10.3115/1626355.1626373.

[41] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, 2001, p. 311, doi: 10.3115/1073083.1073135.

[42] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in Proceedings of the second international conference on Human Language Technology Research -, 2002, p. 138, doi: 10.3115/1289189.1289273.

[43] A. Mauser, S. Hasan, and H. Ney, "Automatic evaluation measures for statistical machine translation system optimization," in Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008, 2008.

[44] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in AMTA 2006 - Proceedings of the 7th Conference of the Association for Machine Translation of the Americas: Visions for the Future of Machine Translation, 2006.

[45] K. Wołk and D. Koržinek, "omparison and adaptation of automatic evaluation metrics for quality assessment of respeaking," Comput. Sci., vol. 18, no. 2, p. 129, 2017, doi: 10.7494/csci.2017.18.2.129.

[46] M. Z. Iqbal, "SUPara : A Balanced English-Bengali Parallel Corpus," SUST J. Sci. Technol., vol. 16, no. 2, pp. 46–51, 2012.

[47] J. Francisca, M. M. MIA, and D. S. M. M. RAHMAN, "Adapting Rule Based Machine Translation From English To Bangla," Indian J. Comput. Sci. Eng., vol. 2, no. 3, 2011.

[48] A. P. Mukta, A. Mamun, C. Basak, S. Nahar, and F. H. Arif, "A Phrase-Based Machine Translation from English to Bangla Using Rule-Based Approach," in 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1–5.

[49] S. S. Ashrafi, H. Kabir, M. Anwar, and A. K. M. Noman, "English to Bangla Machine Translation System Using Context-Free Grammars," IJCSI Int. J. Comput. Sci., vol. 10, no. 3, pp. 144–153, 2013.

[50] K. Muntarina, G. Moazzam, and A. Bhuiyan, "Tense Based English to Bangla Translation Using MT System," Int. J. Eng. Sci. Invent., vol. 2, no. 10, pp. 30–38, 2013.

[51] T. Alamgir, M. S. Arefin, and M. M. Hoque, "An Empirical Machine Translation Framework for Translating Bangla Imperative, Optative and Exclamatory Sentences into English," in 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016, pp. 932–937, doi: 10.1109/ICIEV.2016.7760137.

[52] M. S. Rahman, S. R. Poddar, M. F. Mridha, and M. N. Huda, "Open morphological machine translation: Bangla to English," in 2010 International Conference on Computer Information Systems and Industrial Management Applications, CISIM 2010, 2010, pp. 460–465, doi: 10.1109/CISIM.2010.5643495.

[53] S. A. Chowdhury, "Developing a Bangla to English Machine Translation System Using Parts Of Speech Tagging : A Review," J. Mod. Sci. Technol., vol. 1, no. 1, pp. 113–119, 2013.

[54] M. G. R. Alam, M. M. Islam, and N. Islam, "A New Approach To Develop An English To Bangla Machine Translation System," Daffodil Int. Univ. J. Sci. Technol., vol. 6, no. 1, pp. 36–42, Jan. 2011, doi: 10.3329/diujst.v6i1.9332.

[55] M. M. Anwar, M. Z. Anwar, and M. Al-Amin Bhuiyan, "Syntax Analysis and Machine Translation of Bangla Sentences," IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 9, no. 8, pp. 317–326, 2009.

[56] M. S. Arefin, M. M. Hoque, M. O. Rahman, and M. S. Arefin, "A machine translation framework for translating Bangla assertive, interrogative and imperative sentences into English," in 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2015, no. May, pp. 1–6, doi: 10.1109/ICEEICT.2015.7307534.

[57] M. Anwar, M. Z. Anwar, and A. Bhuiyan, "Structural Analysis of Bangla Sentences for Machine Translation," in International Conference On "Computational Intelligence Applications," 2010, no. April, pp. 3–5.

[58] M. Anwar, "Bangla to English Machine Translation using Fuzzy Logic," Int. J. Comput. Sci. Inf. Secur., vol. 16, no. 11, pp. 156–165, 2018.

[59] M. Rabbani, K. M. R. Alam, and M. Islam, "A new verb based approach for English to Bangla machine translation," in 2014 International

Conference on Informatics, Electronics & Vision (ICIEV), 2014, pp. 1–6, doi: 10.1109/ICIEV.2014.6850684.

[60] M. Haque and M. Hasan, "English to Bengali Machine Translation: An Analysis of Semantically Appropriate Verbs," in 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET), 2018, no. October, pp. 217–221, doi: 10.1109/ICISET.2018.8745626.

[61] M. A. Islam, M. S. H. Anik, and A. B. M. A. Al Islam, "Towards achieving a delicate blending between rule-based translator and neural machine translator," Neural Comput. Appl., vol. 33, no. 18, pp. 12141–12167, Sep. 2021, doi: 10.1007/s00521-021-05895-x.

[62] S. Dandapat, S. Morrissey, S. K. Naskar, and H. Somers, "Statistically motivated example-based machine translation using translation memory," in ICON-2010: 8th International Conference on Natural Language Processing, 2010.

[63] M. A. S. Khan, S. Yamada, and T. Nishino, "How to Translate Unknown Words for English to Bangla Machine Translation Using Transliteration," J. Comput., vol. 8, no. 5, pp. 1167–1174, May 2013, doi: 10.4304/jcp.8.5.1167-1174.

[64] C. Fellbaum, "WordNet and wordnets," in Encyclopedia of Language and Linguistics, 2005.

[65] A. Radford, M. Atkinson, D. Britain, H. Clahsen, and A. Spencer, "The International Phonetic Alphabet," in Linguistics, 2012.

[66] K. M. A. Salam, S. Yamada, and N. Tetsuro, "Improve Example-Based Machine Translation Quality for Low-Resource Language Using Ontology," Int. J. Networked Distrib. Comput., vol. 5, no. 3, p. 176, 2017, doi: 10.2991/ijndc.2017.5.3.6.

[67] M. G. Uddin, M. Murshed, and M. A. Hasan, "A parametric approach to Bangla to English Statistical Machine Translation for complex Bangla sentences -Step 1," in Proceedings of International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh., 2005, no. December 2005, doi: 10.13140/2.1.2720.3526.

[68] M. Z. Islam, J. Tiedemann, and A. Eisele, "English to Bangla phrase-based Machine Translation," in EAMT 2010 - 14th Annual Conference of the European Association for Machine Translation, 2010, no. May.

[69] "SRILM - The SRI Language Modeling Toolkit," 2021. [Online]. Available: https://www.sri.com/case-studies/srilm/. [Accessed: 08-Aug-2021].

[70] M. Roy and F. Popowich, "Phrase-based statistical machine translation for a low-density language pair," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6085 LNAI, 2010, pp. 273–277.

[71] P. Koehn and K. Knight, "Empirical methods for compound splitting," in Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - EACL '03, 2003, vol. 1, p. 187, doi: 10.3115/1067807.1067833.

[72] M. Roy and F. Popowich, "Word Reordering Approaches for Bangla-English Statistical Machine Translation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6085 LNAI, 2010, pp. 282–285.

[73] J. M. Crego and B. Mari, "Syntax-enhanced N -gram-based SMT," MT Summit, 2007.

[74] M. A. Al Mumin, M. H. Seddiqui, M. Z. Iqbal, and M. J. Islam, "shu-torjoma : An English↔Bangla Statistical Machine Translation System," J. Comput. Sci., vol. 15, no. 7, pp. 1022–1039, Jul. 2019, doi: 10.3844/jcssp.2019.1022.1039.

[75] M. Rabbani, K. M. R. Alam, M. Islam, and Y. Morimoto, "PVBMT: A Principal Verb based Approach for English to Bangla Machine Translation," Int. J. Comput. Vis. Signal Process., vol. 6, no. 1, pp. 1–9, 2016.

[76] M. A. Hasan, F. Alam, S. A. Chowdhury, and N. Khan, "Neural vs Statistical Machine Translation: Revisiting the Bangla-English Language Pair," in 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), 2019, no. September, pp. 1–5, doi: 10.1109/ICBSLP47725.2019.201502.

[77] M. A. Al Mumin, M. H. Seddiqui, M. Z. Iqbal, and M. J. Islam, "Neural Machine Translation for Low-resource English-Bangla," J. Comput. Sci., vol. 15, no. 11, pp. 1627–1637, Nov. 2019, doi: 10.3844/jcssp.2019.1627.1637.

[78] M. Z. Islam, J. Tiedemann, and A. Eisele, "English to Bangla phrase-based Machine Translation," in EAMT 2010 - 14th Annual Conference of the European Association for Machine Translation, 2010, no. May.

[79] S. Green, D. Cer, and C. Manning, "Phrasal: A Toolkit for New Directions in Statistical Machine Translation," in Proceedings of the Ninth Workshop on Statistical Machine Translation, 2014, pp. 114–121, doi: 10.3115/v1/W14-3311.

[80] C. Quirk, A. Menezes, and C. Cherry, "Dependency treelet translation: Syntactically informed phrasal SMT," in ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2005, doi: 10.3115/1219840.1219874.

[81] M. Akter, M. Shahidur Rahman, M. Z. Iqbal, and M. R. Selim, "SuVashantor: English to Bangla machine translation systems," J. Comput. Sci., 2020, doi: 10.3844/jcssp.2020.1128.1138.

[82] A. C. Dhar, A. Roy, A. Habib, M. A. H. Akhand, and N. Siddique, "Transformer Deep Learning Model for Bangla-English Machine Translation," in 2nd International Conference on Artificial Intelligence: Advances and Applications (ICAIAA 2021), 2021, pp. 1–10.

[83] A. Roy, A. C. Dhar, M. A. H. Akhand, and M. A. S. Kamal, "Bangla-English Neural Machine Translation with Bidirectional Long Short-Term Memory and Back Translation," Int. J. Comput. Vis. Signal Process., vol. 11, no. 1, pp. 25–31, 2021.

[84] S. Naskar, D. Saha, and S. Bandyopadhyay, "ANUBAAD – A Hybrid Machine Translation System from English to Bangla," in Symposium on Indian Morphology, Phonology & Language Engineering, 2004, pp. 91–92.

[85] M. Rabbani, K. M. R. Alam, M. Islam, and Y. Morimoto, "PVBMT: A Principal Verb based Approach for English to Bangla Machine Translation," Int. J. Comput. Vis. Signal Process., vol. 6, no. 1, pp. 1–9, 2016.

[86] S. Lyons, "A review of Thai–English machine translation," Mach. Transl., vol. 34, no. 2–3, pp. 197–230, Sep. 2020, doi: 10.1007/s10590-020-09248-8.

[87] D. Chopra, N. Joshi, and I. Mathur, "A Review on Machine Translation in Indian Languages," Eng. Technol. Appl. Sci. Res., vol. 8, no. 5, pp. 3475–3478, 2018, doi: 10.48084/etasr.2288.

[88] S. A. B. Andrabi and A. Wahid, "A Review of Machine Translation for South Asian Low Resource Languages," Turkish J. Comput. Math. Educ., vol. 12, no. 5, pp. 1134–1147, Apr. 2021, doi: 10.17762/turcomat.v12i5.1777.

[89] B. Ministry of ICT, "Enhancement of Bangla Language in ICT through Research & Development," Ministry of ICT, Bangladesh, 2020. [Online]. Available: http://eblict.gov.bd/site/page/07289180-0932-4a40-b5ad-8a163780e16f. [Accessed: 01-Jan-2021].

[90] B. Zhang, D. Xiong, and J. Su, "Neural Machine Translation with Deep Attention," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 1, pp. 154–163, 2020, doi: 10.1109/TPAMI.2018.2876404.

[91] B. Zhang, D. Xiong, J. Xie, and J. Su, "Neural Machine Translation With GRU-Gated Attention Model," IEEE Trans. Neural Networks Learn. Syst., vol. 31, no. 11, pp. 4688–4698, Nov. 2020, doi: 10.1109/TNNLS.2019.2957276.

[92] L. H. B. Nguyen, H. D. Minh, D. Dinh, and T. Le Manh, "Improving neural machine translation with POS tags," ICIC Express Lett. Part B Appl., vol. 12, no. 1, pp. 91–98, 2021, doi: 10.24507/icicelb.12.01.91.

[93] B. Pan, Y. Yang, Z. Zhao, Y. Zhuang, and D. Cai, "Bi-Decoder Augmented Network for Neural Machine Translation," Neurocomputing, vol. 387, pp. 188–194, Apr. 2020, doi: 10.1016/j.neucom.2020.01.003.

[94] Y. Li, X. Li, Y. Yang, and R. Dong, "A Diverse Data Augmentation Strategy for Low-Resource Neural Machine Translation," Information, vol. 11, no. 5, p. 255, May 2020, doi: 10.3390/info11050255.

[95] Y. Nishimura, K. Sudoh, G. Neubig, and S. Nakamura, "Multi-Source Neural Machine Translation With Missing Data," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 28, pp. 569–580, 2020, doi: 10.1109/TASLP.2019.2959224.

[96] X. Shi, H. Huang, P. Jian, and Y.-K. Tang, "Improving neural machine translation with sentence alignment learning," Neurocomputing, vol. 420, pp. 15–26, Jan. 2021, doi: 10.1016/j.neucom.2020.05.104.

[97] R. Z. Xiao, a. M. McEnery, J. P. Baker, and A. Hardie, "Developing Asian language corpora: standards and practice," in The Fourth Workshop on Asian Language Resources, 2004, pp. 1–8.

[98] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, 2012, pp. 2214–2218.

[99] T. Nakazawa et al., "Overview of the 5th Workshop on Asian Translation," in Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation, 2018, pp. 904–944.

[100] M. Post, C. Callison-Burch, and M. Osborne, "Constructing parallel corpora for six Indian languages via crowdsourcing," 7th Work. Stat. Mach. Transl., pp. 401–409, 2012.

[101] A. Kids, I. N. Learning, B. Mathematics, E. A. R. A. Eva, and R. Rion, "A Collaborative Platform to Collect Data for Developing Machine Translation System," 2018.

[102] LDC, "Linguistic Data Consortium, Philadelphia, PA," 2021. [Online]. Available: https://www.ldc.upenn.edu/. [Accessed: 20-Jul-2021].

Classification of Breast Cancer Cell Images using Multiple Convolution Neural Network Architectures

Zarrin Tasnim[1], F. M. Javed Mehedi Shamrat[2], Md Saidul Islam[3]
Md.Tareq Rahman[4], Biraj Saha Aronya[5], Jannatun Naeem Muna[6], Md. Masum Billah[7]
Department of Software Engineering, Daffodil International University, Bangladesh[1, 2, 5, 7]
School of Computer and Software, Nanjing University of Information Science and Technology, China[3]
Department of Computer Science and Engineering, Daffodil International University, Bangladesh[4]
Department of Computer Science and Engineering, United International University, Bangladesh[6]

***Abstract*—Breast cancer is a malignant tumor that affects women. It is the most prevalent cancer in women, affecting about 10% of all women at any point in their lives. The development of breast cancer begins in the lobules or ducts of the cells. Early detection and prevention are the best ways to stop this cancer from spreading. In this study, five Convolution Neural Network (CNN) models are used to process image data of breast cells. AlexNet, InceptionV3, GoogLeNet, VGG19 and Xception models are used for the classification of Invasive Ductal Carcinoma, IDC and Non-Invasive Ductal Carcinoma (Non-IDC) cells. The models are trained and tested at different epochs to record the learning rate. It is observed from the study that with higher epochs, the data loss decreases and accuracy increases. The accuracy of InceptionV3 and Xception is 92.48% and 90.72% respectively. Likewise, VGG19 and AlexNet have fairly close accuracy of 94.83% and 96.74%. However, GoogLeNet dominates over the other implemented models with the highest accuracy of 97.80%. The GoogLeNet model performs with high accuracy and precision in detecting IDC cells responsible for breast cancer.**

***Keywords*—Breast cancer; IDC; non-IDC; AlexNet; VGG19; Inception sV3; GoogLeNet; Xecption; accuracy**

I. Introduction

Cancer, also known as a malignant neoplasm, is a group of more than a hundred diseases marked by irregular cell development with the ability to spread to the body's underlying tissues. IDC is a kind of breast cancer that started in the ducts of the breast and has progressed to fibrous or fatty tissue outside of the duct. IDC is the most prevalent kind of breast cancer, accounting for 80% of all occurrences. Breast cancer is the most common kind of cancer in women worldwide [1]. Many imaging techniques have been developed to aid in the early diagnosis and treatment of breast cancer, as well as the reduction of breast cancer-related mortality. To improve diagnostic precision and accuracy, many assisted breast cancer diagnosis methods have been employed [2-4]. Fig. 1 shows breast cancer cases around the world.

To classify and predict breast cancer, machine learning algorithms with image processing have become quite famous for their accuracy in detecting the disease at an early stage. Ciresan et al. [5] classified each pixel into mitotic and non-mitotic groups using an 11-layered CNN. The predictions were made using likelihood ratings allocated to each pixel depending on its distance from the mitosis centroid. A related study [6] used Transfer Learning in CNNs to identify and segment brain and colon cancer images, and the findings were cutting-edge. It used AlexNet (pre-trained on ImageNet) to train a Support Vector Machine with the features extracted from the last FC layer Support Vector Machine (SVM). Gao et al. used CNN to identify interstitial lung infections [7] and discovered that a pre-trained model converged categorization faster than a randomly initialized network. It is possible to automate cell counting in microscope pictures. Weidi et al. [8] took a regression approach to the issue, which eliminates the need for previous identification or segmentation. They regressed a density surface generated by the superposition of Gaussians using completely convolutional regression networks (FCRNs). The dot annotations of each cell given as the ground truth for the training set are expressed by these Gaussians. To identify the best-supervised learning classifier, Vikas Chaurasia and Saurabh Pal [9] evaluate the performance criteria of Naive Bayes, SVM-RBF kernel, RBF neural networks, Decision trees, and basic CART in breast cancer datasets. The experimental results indicate that the SVM-RBF kernel outperforms other classifiers, scoring 96.84% accuracy in the Wisconsin Breast Cancer (original) datasets. Djebbari et al. [10] investigate the impact of an ensemble of machine learning approaches on breast cancer survival period prediction. When compared to prior results, their methodology is more accurate on their breast cancer data collection. S. Aruna and L. V Nandakishore [11] compare the findings of C4.5, Nave Bayes, Support Vector Machine (SVM), and K-Nearest Neighbor to find the appropriate classifier in WBC (K-NN). SVM is the most accurate classifier, with a 96.99% accuracy rate. Angeline Christobel. Y. [12] use a decision tree classifier (CART) to obtain an accuracy of 69.23% in breast cancer datasets. The accuracy of data mining algorithms SVM, IBK, and BF Tree is compared by A. Pradesh [13]. SMO outperforms other classifiers in terms of performance. T.Joachims [14] uses neuron fuzzy methods to reach a precision of 95.06 % by utilizing Wisconsin Breast Cancer (original) datasets. In this study, a hybrid method is proposed to increase the classification accuracy of Wisconsin Breast Cancer (original) datasets using 10-fold cross-validation. Liu Ya-Qin, W. Cheng, and Z. Lu [15] used the C5 algorithm with picking to produce additional data for training from the initial array using variations of repetitions to yield multisets of the same scale as the original data to predict breast cancer survivability. Delen et al. Lu [16] pre-classified 202,932 breast

cancer medical records into two groups: those who "survived" (93,273) and those who "didn't" (93,272). (109,659). The precision of the prediction of survivability was in the region of 93%.

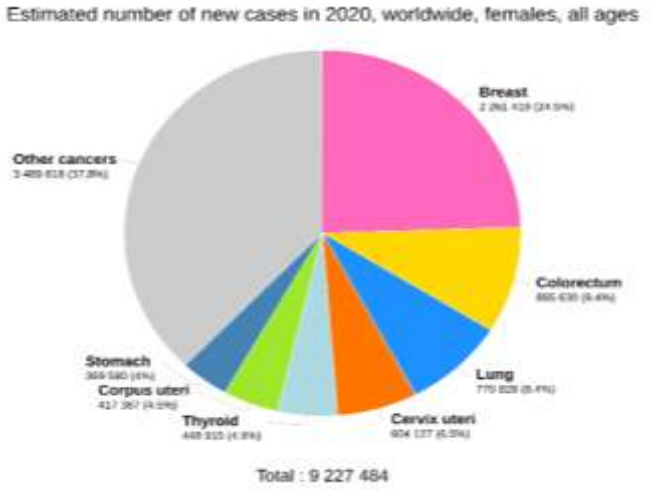


Fig. 1. The Distribution of the Confirmed Breast Cancer Cases around the World (2020). (Source: World Health Organization).

In this work, to detect cancer cells in time for fast treatment, CNN models are used to classify the IDC cells in the breast to determine breast cancer. The CNN models proposed for the study are AlexNet, GoogLeNet, VGG19, Xception and InceptionV3. The aim is to classify IDC from Non-IDC breast cell images from the dataset. Furthermore, the accuracy of the models is compared with each other to determine which model performs best.

The remainder of this paper is organized as follows. Section II contains the overall view of the system. It gives an idea of how the study is conducted in each step. Section III comprises the materials and the methods in detail. It also explains the criteria under which the performance of the implemented models will be evaluated. Section IV summarizes the experimental studies and the obtained results. Section V provides a comparison of the proposed system with the existing studies to show that the proposed system has superior performance over others. Finally, Section VI presents the conclusion of the study.

II. Proposed System

Early detection of breast cancer is a critical field on which researchers are working since it may improve the rate of diagnosis, care, and recovery of affected women. Early identification is the most important measure in reducing this condition's clinical and social risks, given the high expense of care and the high incidence of the disease among women worldwide. There are several approaches and techniques for detecting this form of cancer, each with its own set of benefits and drawbacks. When cancer has spread through the later phases, it is usually identified and diagnosed. This is especially bad since cancer risks have metastasized by the time it is discovered are large, leaving the chances of treating it very low. Self-testing is rarely done, which tends to cancer detection in its latter stages. A lump or mass on the breast, self-examination, or mammography is the most common way to diagnose breast cancer.

In the proposed system, image data of breast cells are used to predict breast cancer. For that, the images are classified to identify which cells cause cancer. The models used for the prediction are AlexNet, IceptionV3, VGG19, Xception and GoogLeNet. An overall flow diagram of the study is presented in Fig. 2.

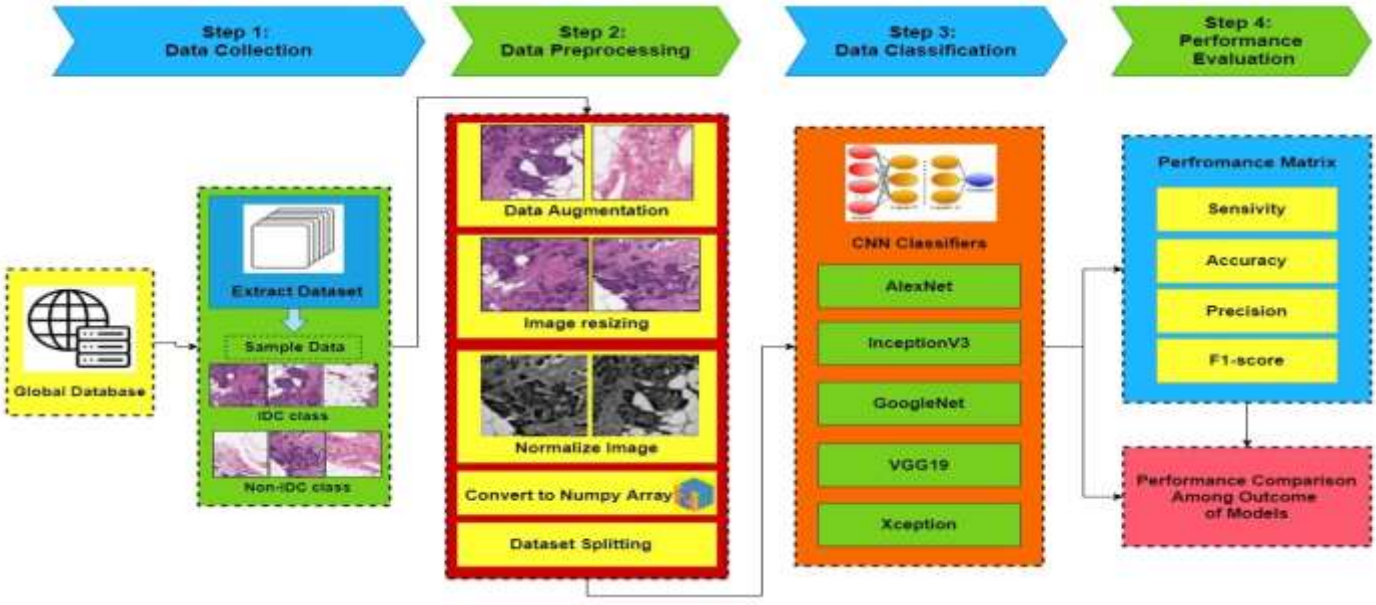


Fig. 2. System Flow Diagram of the Proposed Study.

III. System Operation

A. Data Description

The image used in this study is of breast cells to diagnose Breast cancer. The dataset was retrieved from "https://www.kaggle.com/paultimothymooney/breast-histopathology-images/discussion/130203". For training and testing the machine learning models, 27800 image data are used. The images are categorized into two categories,

- Non-IDC: categorized as "class0".
- IDC: categorized as "class1".

It will help specify IDC (Malignant (cells are abnormal and grow uncontrollably)) and Non-IDC (Benign (if the cells are normal just overgrown)).

Several IDC and non-IDC image data were added for better training of the models to enrich the dataset. Fig. 3 and Fig. 4 contain some of the IDC and non-IDC class images that were added to the dataset, respectively.

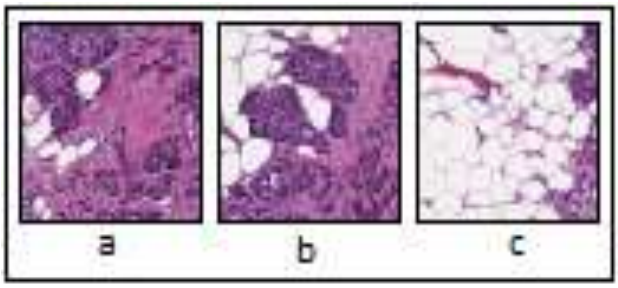


Fig. 3. Image Data of IDC Class.

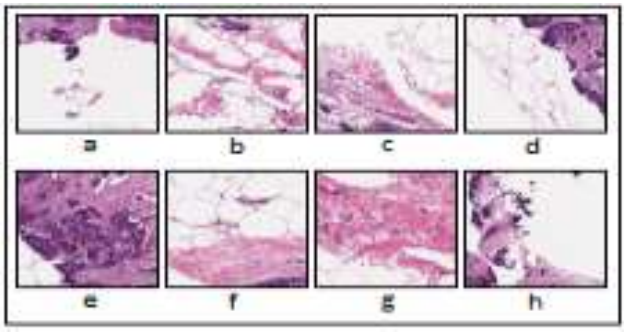


Fig. 4. Image Data of non-IDC Class.

B. Data Pre-processing

The dataset contains raw image data that are not fit for models to be trained and tested. To bring all the raw data into the same scale, data pre-processing is done on the dataset used in the work. The steps of the pre-processing stage are as follows.

- Generating augmented image: to train any machine learning algorithms, a large dataset is required. To increase the dataset volume, augmented images are generated using the ImageDataGenerator class in the Keras library. Table I holds the attributes used for image generation.
- Resize image: Since all images in the dataset are not in the same size, all the images are reshaped to 128 X 128 pixels.
- Normalize image: since all the images are in RBG, converting to greyscale images is being divided by 255 for normalization.
- Convert to NumPy array: image data a converted to NumPy array for faster computation.
- Dataset Splitting: the dataset is split into 80:20 for training and testing the models.

TABLE I. IMAGEDATAGENERATOR ATTRIBUTES

Shear Range	0.3
Zoom_Range	0.2
Horizontal_Flip	True
Vertical_Flip	True
Rescale	1/255

C. CNN Models of Classification

1) AlexNet: AlexNet [17] is an 8-layered network with 5 convolutional layers and 3 Max Pooling layers [18]. ReLU activation is used. 96 filter sizes with a stride of 4 give the first Convolution sheet. After that, the 3X3 inputs go into the Max Pooling Sheet, with a stride of 2. Then, the data is sent to the second convolution layer with stride one and padding two, with a total of 256 5.x/2. The data is then followed by a second datasheet, where the stride is 2 and the filter size is 3. Three Convolution layers with 384, 384, 384, and 256 kernels are then applied to the input results, followed by an Activation layer with 3 X 3 kernels, followed by a Reshape layer with 512 kernels and a padding value of 1. With pool size 3, the final MaxPooling is implemented. If all the operations have been performed, the results are transferred to three connected layers, which are eventually converted into totally connected layers. In Fig. 5, the architecture is seen.

2) InceptionV3: InceptionV3 [19] is the third iteration of Google's Inception Convolutional Neural Network, which was first shown at the ImageNet Recognition Challenge. It includes Label Smoothing, Factorized 7X7 Convolution, RMSProp Optimizer, BatchNorm in the Auxillary Classifiers and a downscaling classifier to identify and add information from smoothed label sequences. This is shown in Fig. 6.

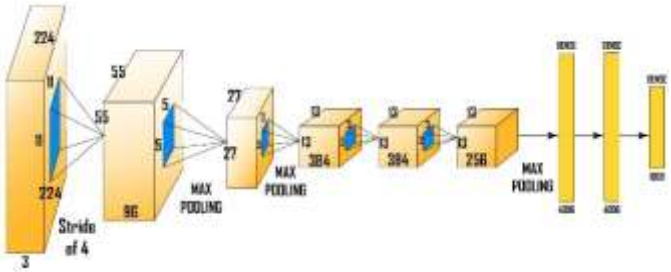


Fig. 5. AlexNet Architecture.

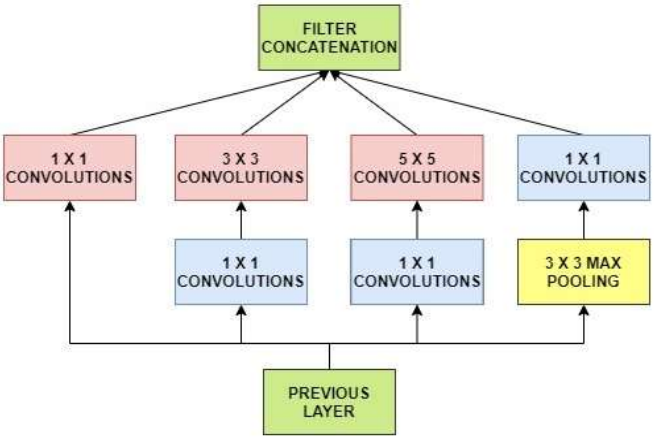


Fig. 6. Implemented InceptionV3 Structure.

3) GoogLeNet: Since 'InceptionV1' [20] is sometimes referred to as GoogLeNet [21]. There are 47 stages of aggregation and several pooling layers in GoogLeNet. A result is that, to sum up, the nine Inception modules are lined up one after another. In the case of GoogLeNet, the stochastic descent algorithm is employed. The following Fig. 7 is an example.

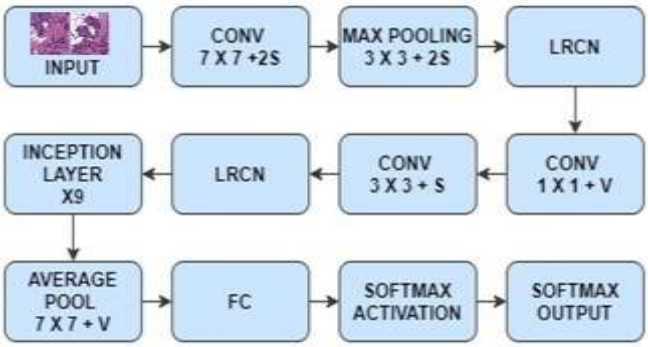


Fig. 7. GoogLeNet Model.

4) VGG19: VGG19 is a version of the VGG model that includes 16 convolutional layers, three fully connected layers, five MaxPool layers, and one SoftMax layer as shown in Fig. 8. A fixed-size RGB picture was used as the input to this network, which also has a matrix of the same size. Max pooling was done using sride 2 across a 2 * 2 pixel window. This was followed by the Rectified linear unit (ReLu) to add non-linearity into the model in order to enhance classification and computing speed. Three completely linked layers were implemented. And finally, a softmax function is used as the last layer.

5) Xception: The Xception model's base layer is initially frozen with the command (include top=False), followed by the trainable layer as shown in Fig. 9. The trainable layer employs images that have undergone the Average Pooling procedure. The Average Pooling pool size is (7,7), and there are 128 hidden nodes accessible in this layer. The Adam Stochastic gradient descent method is utilized for optimization, and the ReLu activation function is employed in that layer. Following that, in the output layer, the Softmax activation function is utilized to identify IDC cells using two nodes. A learning rate of 0.01 is specified for backpropagation.

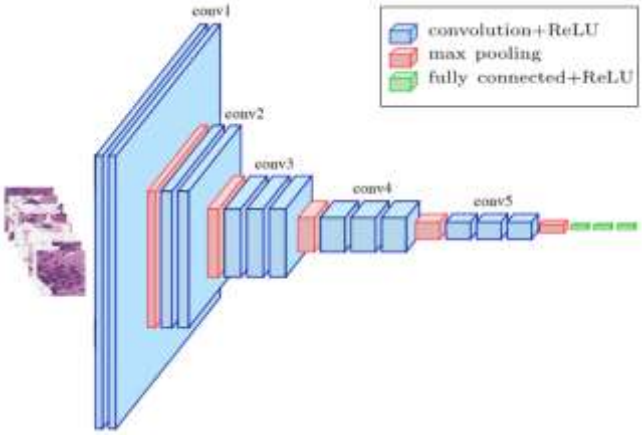


Fig. 8. Applied InceptionV3 Structure.

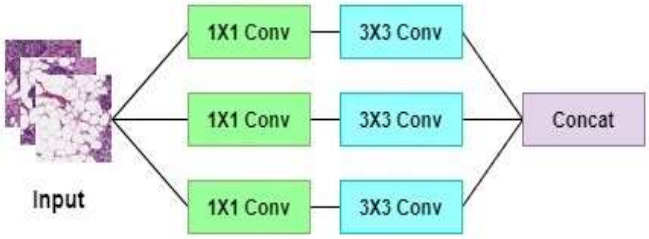


Fig. 9. Implemented Xception Model.

D. Performance Evaluation

After completing the training and testing process, the performance of the models is calculated [22, 23]. The evaluation criteria are precision, recall, f1-score, and accuracy, as described in Eq. 1, 2, 3, and 4.

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

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

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

$$F1-Score = \frac{Recall*Precision}{Recall+Precision} \quad (4)$$

IV. Result and Analysis

Using the CNN models Vgg19, Xception, AlexNet, InceptionV3, and GoogLeNet, breast cancer prediction can be done successfully. The models can efficiently classify IDC cells and non-IDC cells. However, not all the models perform the same. Some show more accuracy than others.

The outcome was recorded for up to 10 epochs. Table II illustrates the recorded data for the InceptionV3 model where the highest accuracy is 92.48%.

Table III contains the record of accuracy and data loss of the GoogLeNet model for the epoch. It can be observed from the table that with every new epoch, the rate of data loss gradually decreases and the accuracy increase for both the training and test set.

TABLE II. TRAINING AND TESTING OUTCOME FOR INCEPTION V3

Epochs	Training Data Loss	Training accuracy in %	Testing Data loss	Testing accuracy in %
1	13.29	77.34	10.21	83.42
2	11.02	81.93	6.46	89.94
3	7.76	85.67	6.39	90.24
4	7.39	86.76	6.23	90.73
5	7.56	88.34	5.93	91.23
6	7.26	87.99	5.74	91.47
7	6.98	89.23	5.46	91.78
8	7.23	88.80	5.28	92.12
9	7.34	88.23	4.97	92.48
10	7.03	88.57	5.01	92.10

TABLE III. TRAINING AND TESTING OUTCOME FOR GOOGLENET

Epochs	Training Data Loss	Training accuracy in %	Testing Data loss	Testing accuracy in %
1	9.39	93.53	7.14	94.54
2	5.28	95.39	4.24	96.83
3	5.19	95.73	4.19	97.23
4	5.03	95.93	3.95	97.53
5	4.83	96.23	3.69	97.46
6	4.90	96.15	3.53	97.94
7	4.85	96.39	3.45	97.45
8	4.72	96.32	3.33	97.23
9	4.79	96.83	3.49	97.42
10	4.67	96.45	3.19	97.80

The record of outcome accuracy of AlexNet has stated in Table IV with data loss in every epoch. The highest rate of accuracy rate of AlexNet model is 96.74% in the test set with a data loss rate of 9.59% at the 10^{th} epoch. At the same epoch, it achieved the highest accuracy rate on the training set as well with 96.34%.

TABLE IV. TRAINING AND TESTING OUTCOME FOR ALEXNET

Epochs	Training Data Loss	Training accuracy in %	Testing Data loss	Testing accuracy in %
1	29.26	89.93	25.77	91.01
2	25.18	91.37	20.54	92.80
3	22.06	92.32	17.28	93.59
4	20.62	93.34	15.66	94.11
5	17.97	93.73	15.23	94.21
6	14.49	95.06	14.98	95.87
7	14.44	95.45	10.10	96.23
8	10.29	96.11	9.89	96.19
9	9.87	96.72	9.48	96.33
10	9.93	96.34	9.59	96.74

Table V contains the record of accuracy and data loss of VGG19 model with respect to epoch. The rate of data loss gradually decreases as the accuracy increase for both the training and test set. The model gives an accuracy of 94.83% with 5.1 data loss.

Table VI contains the record of outcome accuracy of the Xception model implemented in the dataset, along with data loss for each epoch. Xception model achieves an accuracy rate of 90.72% in the test set, with a data loss rate of 7.21 at the 8th epoch.

From the recorded data is can be observed that on the 10th epoch, all the models show the highest accuracy and lowest data loss for both train and testing data. A graphical comparison of the accuracy for training (a) and testing (b) data for the models are depicted in Fig. 10 as well.

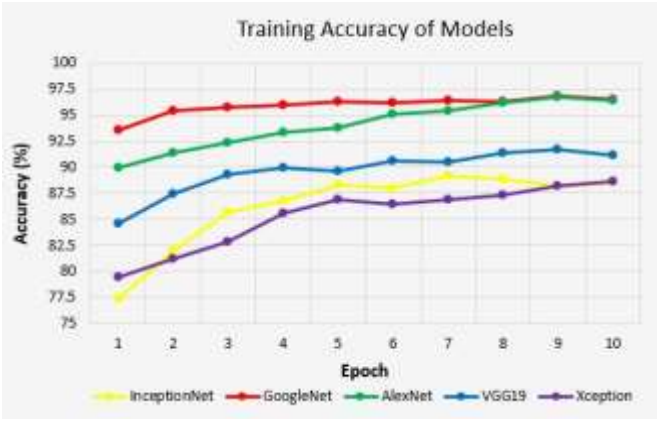
The data loss rate for the models decreases for both training and test set with every increasing epoch for all the models. The graphs in Fig. 11 show the rate of data loss obtained in each epoch as the model learns from the training (a) and testing (b), the less data it losses.

TABLE V. TRAINING AND TESTING OUTCOME FOR VGG19

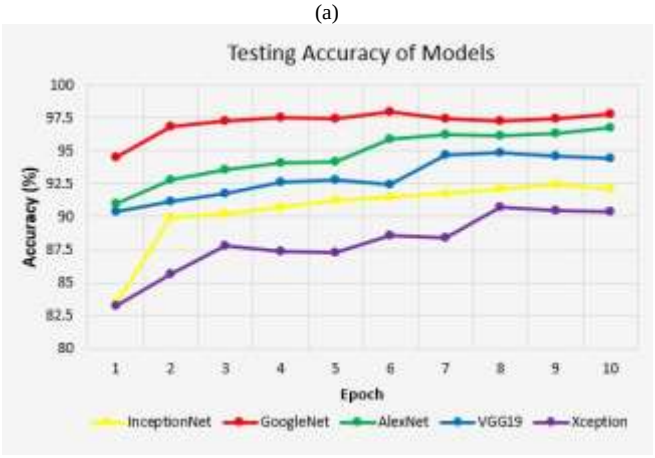
Epochs	Training Data Loss	Training accuracy in %	Testing Data loss	Testing accuracy in %
1	13.85	84.63	9.85	90.36
2	11.53	87.44	9.37	91.18
3	10.55	89.27	8.77	91.74
4	9.45	89.94	7.23	92.58
5	9.46	89.66	7.19	92.83
6	8.57	90.57	7.85	92.47
7	9.35	90.48	5.55	94.66
8	8.35	91.38	5.10	94.83
9	8.02	91.65	6.02	94.57
10	8.33	91.19	6.16	94.46

TABLE VI. TRAINING AND TESTING OUTCOME FOR XCEPTION

Epochs	Training Data Loss	Training accuracy in %	Testing Data loss	Testing accuracy in %
1	16.46	79.46	12.34	83.26
2	15.33	81.24	10.73	85.63
3	14.63	82.84	9.27	87.78
4	10.53	85.59	9.74	87.38
5	10.12	86.94	9.49	87.26
6	10.11	86.47	8.74	88.57
7	9.73	86.88	8.37	88.39
8	9.48	87.37	7.21	90.72
9	9.46	88.24	7.64	90.48
10	9.53	88.63	7.48	90.38

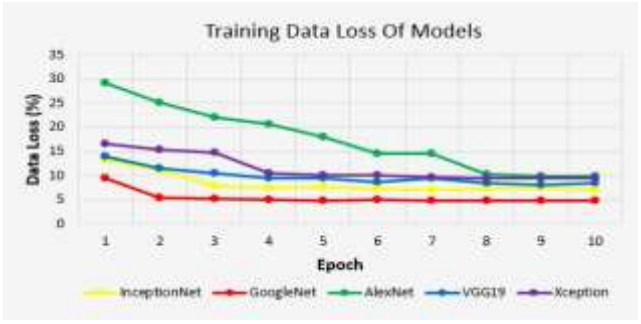


(a)

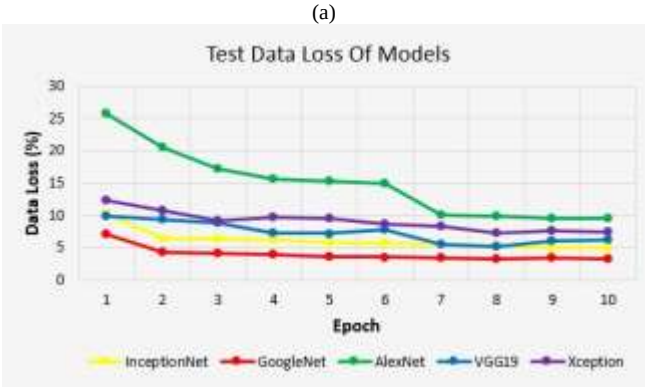


(b)

Fig. 10. Accuracy Comparison for (a) Training Data and (b) Testing Data of the Models.



(a)



(b)

Fig. 11. Data Loss Comparison for (a) Training Data and (b) Testing Data of the Models.

From the comparison, it is understood that GoogLeNet demonstrates much higher accuracy in predicting breast cancer and classifying cancer cells over other models. For further clarity, and overall performance measurement of the models is calculated in Table VII.

From Fig. 12, it can be deduced that GoogLeNet shows a higher rate for precision, recall and F1-score (97.34%, 96.46%, and 96.46%) along with the highest accuracy in predicting breast cancer.

GoogLeNet shows much superior classification performance over InceptionV3 and AlexNet. The models can successfully classify IDC and Non-IDC cells with 97.8% accuracy. A classification outcome of GoogLeNet for both classes is demonstrated in Fig. 13.

TABLE VII. CLASSIFICATION REPORT OF TEST DATA

Models	Performance Measures			
	Precision	*Recall*	*Accuracy*	*F1-score*
Inception v3	89.84%	90.12%	92.48%	89.54%
GoogLeNet	97.34%	96.46%	97.80%	96.46%
Alexnet	95.12%	93.54%	96.74%	94.65%
VGG19	93.63%	92.47%	94.83%	91.36%
Xception	89.49%	91.33%	90.72%	90.03%

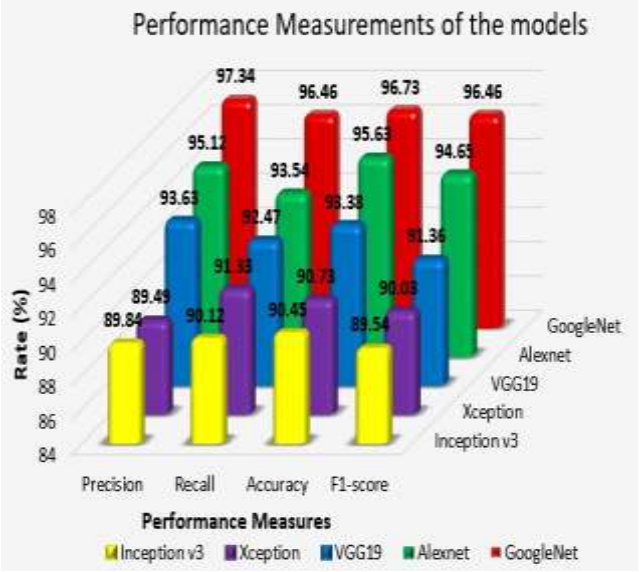


Fig. 12. Performance Evaluation of the Models.

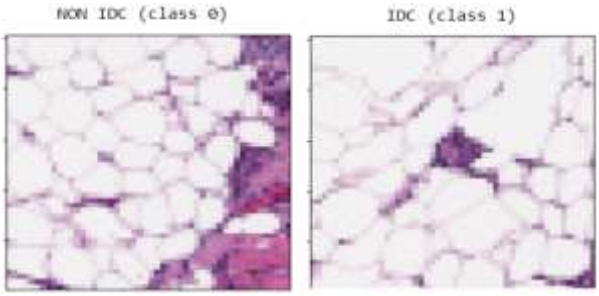


Fig. 13. Classified Classes by GoogLeNet.

V. Comparative Analysis

As mentioned before, many studies have been performed to predict Breast cancer using machine learning, deep learning, and other techniques based on different imaging data of cancer cells. The following table VIII compares the suggested technique to many other methods used by researchers on various datasets. Despite the fact that previous studies have shown a high degree of accuracy in predicting breast cancer, the proposed method surpassed prediction accuracy.

TABLE VIII. Comparison of Previous Methods and the Proposed Method

Studies	Datasets	Models	Accuracy
Proposed study	Breast histopathology images from kaggle	AlexNet	96.74%
		VGG19	94.83%
		Xception	90.72%
		InceptionV3	92.48%
		GoogLeNet	**97.80%**
[9]	Wisconsin Breast Cancer (original) datasets	SVM-RBF	96.84%
[11]	-	SVM	96.99%
[12]	Breast cancer datasets	Decision tree	69.23%
[14]	Wisconsin Breast Cancer (original) datasets	Neuron fuzzy methods	95.06%
[24]	Wisconsin Breast Cancer (original) datasets	SVM	97.13%
[25]	Image data from Mayo Clinic	SD-CNN	90%
[26]	Kaggle 162 H&E	CNN	87%

VI. Conclusion

Breast cancer is a deadly disease that has claimed the lives of many people in both emerging and industrialized countries around the world. Breast cancer is the second leading cause of cancer death among females in industrialized countries and the first among females in developing countries. According to recent estimates, one of every eight people in Bangladesh will grow breast cancer over their lifetime. As a result, the fight against cancer is far from over. The main focus of this paper is to detect breast cancer that an early stage using the images of IDC cells in the breast. To complete the work, CNN models are used to classify the image data. AlexNet, InceptionV3, VGG19, Xception and GoogLeNet are the algorithms used in the classification process. The algorithms separately are successful in predicting the disease. However, they vary in accuracy of the prediction. It is seen that GoogLeNet has a much higher accuracy rate of 97.80% compared to the AlexNet and InceptionV3 with 96.74% and 92.48%. And VGG19 and Xception with 94.83 % and 90.72 %. GoogLeNet has higher Precision, Recall, and F1 scores than the other two models. However the models are only tested for the data images available in the Breast histopathology images dataset. The models will further be studied on more dataset to ensure that they can be universally used to detect breast cancer. Using the same process, it is also to be studied that other diseases such as liver cancer, colon cancer, ovarian cancer, etc., can be diagnosed efficiently and with high accuracy at a very early stage.

References

[1] Sharma, G. N., Dave, R., Sanadya, J., Sharma, P., & Sharma, K. K. (2010). Various types and management of breast cancer: an overview. *Journal of advanced pharmaceutical technology & research*, *1*(2), 109–126.

[2] P. Ghosh et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," in IEEE Access, vol. 9, pp. 19304-19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[3] M. A. Islam, S. Akter, M. S. Hossen, S. A. Keya, S. A. Tisha and S. Hossain, "Risk Factor Prediction of Chronic Kidney Disease based on Machine Learning Algorithms," *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 2020, pp. 952-957, doi: 10.1109/ICISS49785.2020.9315878.

[4] F.M. Javed Mehedi Shamrat, Md. Asaduzzaman, A.K.M. Sazzadur Rahman, Raja Tariqul Hasan Tusher, Zarrin Tasnim "A Comparative Analysis of Parkinson Disease Prediction Using Machine Learning Approaches" International Journal of Scientific & Technology Research, Volume 8, Issue 11, November 2019, ISSN: 2277-8616, pp: 2576-2580. [Asif].

[5] Cireşan D C, Giusti A, Gambardella L M, and Schmidhuber J (2013) Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks, pp 411–418 (Springer Berlin Heidelberg, Berlin, Heidelberg). doi:10.1007/978-3-642-40763-5_51. Available: https://link.springer.com/chapter/10.1007/978-3-642-40763-5_51.

[6] Xu Y, Jia Z, Wang L-B, Ai Y, Zhang F, Lai M, Eric I, and Chang C (2017) Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinform. 18: 281. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1685-x.

[7] Gao M, Bagci U, Lu L, Wu A, Buty M, Shin H-C, Roth H, Papadakis G Z, Depeursinge A, Summers R M, Xu Z, and Mollura D J (2018) Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks. Comput. Methods Biomech. Biomed. Eng. Imaging Vis. 6: 1–6. doi:10.1080/21681163.2015.1124249. Available: https://pubmed.ncbi.nlm.nih.gov/29623248/.

[8] Xie W, Noble J A, and Zisserman A. Microscopy cell counting and detection with fully convolutional regression networks. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2016, pp. 1–10. Available: https://www.tandfonline.com/doi/abs/10.1080/21681163.2016.1149104.

[9] Chaurasia, Vikas and Pal, Saurabh, Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability (June 29, 2017). International Journal of Computer Science and Mobile Computing IJCSMC, Vol. 3, Issue. 1, January 2014, pg.10 – 22, Available at SSRN: https://ssrn.com/abstract=2994925.

[10] Djebbari A, Liu Z, Phan S, Famili F. An ensemble machine learning approach to predict survival in breast cancer. International Journal of Computational Biology and Drug Design. 2008 ;1(3):275-294. DOI: 10.1504/ijcbdd.2008.021422. Available: https://europepmc.org/article/med/20054993.

[11] S. Aruna and L. V Nandakishore, "KNOWLEDGE BASED A NALYSIS OF VARIOUS STATISTICAL TOOLS IN DETECTING BREAST CANCER", Computer Science & Information Technology (CS & IT), pp. 37–45, 2011.

[12] Christobel, A. "An Empirical Comparison of Data Mining Classification Methods," vol. 3, no. 2, pp. 24–28, 2011. Available: https://www.semanticscholar.org/paper/An-Empirical-Comparison-of-Data-Mining-Methods-Christobel/ba873f0723d244c6f47c58d931747108cf26abc9.

[13] A. Pradesh, "Analysis of Feature Selection with Classification: Breast Cancer Datasets," Indian J. Comput. Sci. Eng., vol. 2, no. 5, pp. 756–763, 2011.

[14] Thorsten J. Transductive Inference for Text Classification Using Support Vector Machines. Icml. 1999;99:200-209. doi:10.4218/etrij.10.0109.0425.

[15] L. Ya-qin, W. Cheng, and Z. Lu, "Decision tree based predictive models for breast cancer survivability on imbalanced data," pp. 1–4, 2009. Available: https://www.infona.pl/resource/bwmeta1.element.ieee-art-000005162571/tab/summary.

[16] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," Artif. Intell. Med., vol. 34, pp. 113–127, 2005. Available: https://www.sciencedirect.com/science/article/pii/S0933365704001010.

[17] Hosny, K. M., Kassem, M. A., & Fouad, M. M. (2020). Classification of skin lesions into seven classes using transfer learning with AlexNet. Journal of digital imaging, 33(5), 1325-1334.

[18] Zarrin Tasnim, Sovon Chakraborty, F. M. Javed Mehedi Shamrat, Ali Newaz Chowdhury, Humaira Alam Nuha, Asif Karim, Sabrina Binte Zahir and Md. Masum Billah, "Deep Learning Predictive Model for Colon Cancer Patient using CNN-based Classification" International Journal of Advanced Computer Science and Applications(IJACSA), 12(8), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120880.

[19] Dong, N., Zhao, L., Wu, C. H., & Chang, J. F. (2020). Inception v3 based cervical cell classification combined with artificially extracted features. Applied Soft Computing, 93, 106311.

[20] Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). An overview of early vision in inceptionv1. Distill, 5(4), e00024-002.

[21] Hirano, G., Nemoto, M., Kimura, Y., Kiyohara, Y., Koga, H., Yamazaki, N., ... & Nagaoka, T. (2020). Automatic diagnosis of melanoma using hyperspectral data and GoogLeNet. Skin Research and Technology, 26(6), 891-897.

[22] F. M. Javed Mehedi Shamrat, P. Ghosh, M. H. Sadek, M. A. Kazi and S. Shultana, "Implementation of Machine Learning Algorithms to Detect the Prognosis Rate of Kidney Disease," 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 2020, pp. 1-7, doi: 10.1109/INOCON50539.2020.9298026.

[23] P. Ghosh, S. Azam, K. M. Hasib, A. Karim, M. Jonkman, A. Anwar, "A Performance Based Study on Deep Learning Algorithms in the Effective Prediction of Breast Cancer," International Joint Conference on Neural Networks (IJCNN 2021), 2021.

[24] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", Procedia Computer Science, Volume 83, 2016, Pages 1064-1069, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.04.224.

[25] Gao, F., Wu, T., Li, J., Zheng, B., Ruan, L., Shang, D., & Patel, B. (2018). SD-CNN: A shallow-deep CNN for improved breast cancer diagnosis. Computerized Medical Imaging and Graphics, 70, 53-62.

[26] Alanazi, S. A., Kamruzzaman, M. M., Islam Sarker, M. N., Alruwaili, M., Alhwaiti, Y., Alshammari, N., & Siddiqi, M. H. (2021). Boosting breast cancer detection using convolutional neural network. Journal of Healthcare Engineering, 2021.

A Multi-dimensional Credibility Assessment for Arabic News Sources

Amira M. Gaber[1], Mohamed Nour El-din[2], Hanan Moussa[3]
Information System department, Faculty of Computer and Artificial Intelligent, Cairo University, Giza, Egypt[1, 2, 3]
Higher Institute of Computer Science and Information Systems, Culture and Science City Academy, 6 October-Giza-Egypt[1]

***Abstract*—Due to the advances in social media, it has become the most popular means of the propagation of news. Many news items are published on social media like Facebook, Twitter, Instagram, etc. Facebook is a huge source for spreading and consuming daily news, but it is an unstructured way of producing news about domains (Art, Health, Education, Sport, Politics, etc.). Thus, this paper will present a model to assess the credibility of news sources over the social context in a particular domain through a particular period of time from a multidimensional perspective. Based on these dimensions of credibility, this model will be designed, evaluated, and implemented by using machine learning algorithms and Arabic NLP approaches to assess the credibility score for Arabic news sources on Facebook. In addition, the study will visualize their scores at different data analysis levels to make the assessment more precise and trustworthy. The proposed model has been implemented and tested over some real Arabic news sources for specific domains and over a period of time to produce a credibility score for each one, whereas the user can display these scores and choose the most credible news sources. The credibility assessment model will be more specific and accurate for a specific domain and time with an accuracy of 98%.**

***Keywords*—Information credibility; social media; machine learning; Arabic Natural Language Processing (ANLP)**

I. INTRODUCTION

Recently, online social media has become a great way to connect people with others. Users of social media share news, communicate with other people, and create more posts and tweets related to the news than they consume. Consequently, a huge amount of incredible news is created and propagated through social media, which has a serious impact on society and individuals. Though this impact can make political gains, increase advertising revenue, and attract attention, it can harm the reputation of businesses. The news spreads rapidly on social networks and may lead to many problems. Various domains on social media such as politics, education, finance, art, sports, and health are not categorized for the news. The objective of the research community in social media now is to reach out the authenticity of the information from the trusted sources to build trusted news displayed on the internet in different domains. In this regard, information credibility has become an important indicator for citing particular news sources over others in different domains at a particular time.

Here, it is worth mentioning that there are many problems on setting and defining credibility**.** Firstly, the lack of assessing credibility in multiple domains at the same time, such as arts, sports, and education, etc. Secondly, there are a few studies work on the Arabic language for the purpose of assessing credibility because its structural essentials are complex. Finally, there is a strong need to assess the credibility of the news sources automatically on Facebook social media which has become the most popular mean to spread the news rabidly.

Information credibility is a topic that has become very important for assessing credibility of information to avoid the conflict from huge number of information due to the growth of technology and internet. In this section will present the main types of information credibility, Level of assessment and approaches used in the assessment process.

There are two main types of information credibility, offline and online. Offline information credibility the news spreads through means of 'traditional media that do not require the use of digital technologies like magazines, journals, etc., while online information credibility concerns the news will spread through digital technologies such as social media or the web. Online information credibility has two types of information credibility are content or social information credibility. The content-based information credibility involves the news collected from various websites, web blogs through a search engine, or fact-checking websites while the social-based information credibility the news published over social media like Facebook, Twitter, Instagram, etc. [1].

Moreover, there are several levels of credibility that can be used to assess the new source these levels are, Post-Level which the assessment of credibility occurs over one post or tweet, and the assessment includes accurate information about a certain topic. It's semantic and characteristic of the message either image, videos, or audio may be considered and sentiment features will be considered also, while topic –Level which the assessment of the credibility can be calculated for a certain topic hashtag or trending topics that attract any users and the users start to comment or retweet about this topic, and member- level applies the assessment on each user account over the social media and measures the reliability of a user by the following features: age group, sex, school degrees, the number of followers, friends, tweets, and retweets, finally hybrid-level (post - topic - member) the credibility assessment contains (topic, post, and user) assessment [2].

There are approaches for assessing the information credibility, automated-based approaches, human – based approaches, or hybrid approaches. The automated approaches consist of several algorithms that can be followed, these algorithms are weighted and information retrieval algorithms

which the algorithms for measuring credibility for a certain claim collected from different credible independent sources and provides evidence for the similarity about the claim from different sources or knowledge graph algorithms: which used to build an information graph of facts. The graph has made up of relations to all derived information and their origin, and finally, machine learning algorithms which can be supervised techniques, unsupervised techniques, semi-supervised techniques, or reinforcement techniques while Human-Based approaches can be cognitive and perception or crowdsourcing: or voting approaches, and hybrid approaches can be automated or human using experts who are specialized in a specific domain [2].

Motivated by the above, there are many contributions of the research are:

- Build a model based on social-context credibility assessment using a hybrid level (content and member) assessment executing over three dimensions: calculating a credibility score of the news source in a specific domain at a specific time to enhance the credibility assessment and give more precise and accurate results.
- Visualize the results over different data analysis.

This paper is organized as follows. Section 2 introduces the related work for the existing studies accessing credibility. Section 3 presents the text categorization methodology. Section 4. The details of suggested model, Section 5 displays the experiment results, and finally, Section 6 concludes this works.

II. Related Work

Some studies have developed approaches attempt to assess the credibility in Social Media. Popat, Kashyap, et al. [3], proposed approaches to leverage the stance, reliability and trend of sources of evidence and counter evidence for credibility assessment of textual claims. And provide explanations for the credibility verdict in the form of informative snippets from articles published by reliable sources that can be easily interpreted by the users. Another one, Ahmad, Faraz, and Syed Afzal Murtaza Rizvi [4], presented a survey approaches for detection of rumors on social networks. They present the data collection methods, extract the features which responsible for finding and estimating credibility using different machine learning techniques. And, Pasi, Gabriella, and Marco Viviani [5], Present Three important issues in the assessment information credibility: in the review sites, it detects the opinion spam, detect the fake news in microblogging, and the credibility assessment of online health-related information. They present a concise survey of the approaches and methodologies.

There are some frameworks assess the credibility, Jaho, Eva, et al. [6], proposed an Alethiometer framework to improve trustworthiness and the validity of contents in the midst of overloads of information. This paradigm was utilized for evaluating the veracity of news consumed on social media. It measures the content by deriving a single metric that takes into account the quality of the contributor and the content, in a unified manner, providing a few preliminary statistical data from the examination of 10 million Twitter users, which offers useful insights into social media data characteristics. Maps of features value on a qualitative scale manner.

Moturu, Sai T., and Huan Liu [7], proposed also, a two-step unsupervised, feature-driven framework for health content, and proposed a various unsupervised scoring models for the user content based.

Mahmood, Saba, et al. [8] present a framework which predict the credibility of web information for a typical user in order to improve decision-making. This framework decides the content credible if based on expert' reviewer. It uses the past behavior of the entities and opinion of others as reputation for expert ranking. Reputation based credibility assessment computing the credibility of the content by evaluating the reputation of the expert' reviewer.

Some researches presented algorithms which assess the credibility, Gupta, M., et al. [9], enhanced the Basic Credibility Analyzer (BasicCA) for Twitter algorithm, which automatically assesses the credibility of Twitter events by performing event graph optimization (EventOptCA). The EventOptCA better than the Basic CA because it depends on stronger event associations inferred from shared unigrams (event similarity) which gives more accurate performance for credibility scores. Abbasi, Mohammad-Ali, and Huan Liu [10], propose a method to measure user credibility in social media. They assess the credibility of the content and user (source of the information) based on the user's profile by using the *CredRank* algorithm, this algorithm measures the user credibility in social media and analyzes social media users' online behavior to measure their credibility. Gupta, Aditi, et al. [11], present a semi-supervised ranking model for scoring tweets according to their credibility" TweetCred" it is available as a browser plug-in and score credibility on tweets user's timeline. Mitra, Tanushree, and Eric Gilbert [12], present CREDBANK algorithm combining machine and human computation on tweets, topics, events and associated human credibility judgements.

There are tools assessing the credibility, Saikaew, Kanda Runapongsa, and Chaluemwut Noyunsan [13], developed a tool which is a chrome extension of fb credibility for Facebook users to evaluate the credibility of by manual human's labelling by using Support Vector Machine (SVM). Horne, Benjamin D. et al. [14], introduced a tool called News Landscape (NELA)Toolkit is an open source toolkit for the systematic exploration of the news landscape. NELA allows users to assess the credibility of news articles using the content based features, filter and sort article predictions based on the user needs. The NELA allows users to visualize the media landscape at different time slices using a variety of features computed at the source level. Also, Saez-Trumper and Diego [15], presented a web application called "the Fake Tweet Buster (FTB)", which identifies fake tweets' images and users that are consistently uploading and/or promoting fake information on Twitter.by reverse image searching, user analysis and a crowd sourcing approach to detect that kind of malicious users on Twitter. Finally, Lorek, Krzysztof, et al. [16], presented a tool called "TwitterBOT" which able to

score submitted tweets by using an automated credibility assessment on Twitter. Using random forest algorithm as an automatic classifier.

Today, several studies have developed models to assess the credibility in Social Media. Podobnik, Vedran, et al. [17], proposed a model on a Facebook application named *"Closest Friends"*, which calculates the user's closest friends on the Facebook social network to calculate who-trusts-whom and implement that knowledge in the social recommender, it record the detail of all possible relations between users in the social network and evaluate those relations properly.

Bauskar, Shubham, et al. [18], present a novel machine learning model based on Natural Language Processing (NLP) techniques for the detection of 'fake News' by using both content-based features and social features of news.

Some researches assess credibility for Arabic news it is difficult and have a few papers talk about it, Jardaneh, Ghaith, et al. [19], utilized machine learning algorithms to identify fake news Arabic tweets based by content and user features supervised classification model.by extract important features for classification purposes and employ sentiment analysis to generate new features for the detection of fake Arabic news. Sentiment analysis led to improving the accuracy of the prediction process.

Chadi Helwe et al. [20], developed methods for assessing the authenticity of Arabic blogs using deep co-learning, a semi-supervised end-to-end approach, and deep learning. This deep co-learning technique is based on co-training, which employs several classifiers that learn from one another using distinct data characteristics. The classifiers are trained using a small train set in a totally supervised way initially. The small, fully-labeled dataset trains two deep learning models for assessing the credibility of Arabic blog posts. The two classifiers are based on a convolutional neural network (CNN) architecture. The first model uses continuous bag of words (CBOW) word embeddings as features, while the second uses character level embeddings.

Al-Eidan, Rasha M. BinSultan et al. [21], proposed a system to measure information credibility of Arabic web content automatically focused on weblogs and need to enhance the selected credibility features by using SVM machine learning technique to be suitable for the Arabic language.

Floos, Ahmad Yahya M. [22], the author illustrates the difficulty of Arabic rumor identification in twitter social platform by studying the impact based on Arabic tweet content. And explains how these content features are too influential in measuring the credibility of those Arabic tweets.

Mouty, Rabeaa, and Achraf Gazdar [23], focus on discover the credibility of Twitter publishing in Arabic by integrating the social media mining techniques with the Natural Language Processing (NLP) by using the random forest classifier.

El Ballouli, R., et al. [24], presented a Credibility Analysis of Arabic Content on Twitter (CAT) model which uses a binary classifier that classifies a given tweet as either credible or not and uses both content based and user-based features. The authors [25] enhanced in "TweetCred" algorithm, but there is difference between "TweetCred" algorithm and CAT algorithm which the first one relies in its classification on real-time features only, such as, count of re-tweets, and count of friends. However, CAT utilizes the tweeter's history for any clues that might be helpful in deciding on the credibility of the tweet.

Al Zaatari, A., et al. [24], creating two corpora for credibility analysis and validate their usefulness for Arabic blogs, the two corpora manually labeled as credible, fairly credible or non-credible by a number of human judges.

Finally, Alrubaian, Majed, et al. [2], introduced a comprehensive study about the information credibility assessment at different levels of features. It addressed a new taxonomy of credibility analysis and assessment techniques. And a cross-referencing of literature review and suggested a new topic for future studies of credibility assessment in social media context. This research suggested the credibility researchers commonly use text analysis tasks. However, analysis of multimedia (images, audio, and video) must be explored further, Text analysis has been employed effectively; nonetheless, semantic analysis of text content has not been explored, the feature levels of credibility assessments require further investigation, especially in terms of relative importance. In some cases, certain features are more important than others, leading to misjudgment of the trustworthiness of the content and source, expand the hybrid models to formulate automation relevant to social media content credibility judgment.

Different from the above work, this paper proposes an efficient assessment model to enhance the accuracy of assessing credibility for news source in online social network specially over the Facebook social media. Online users can check the veracity of this news from credible sources specialized in field or domain at a certain time. Although all the above work assesses the credibility of users or post but didn't assess the credibility of a news source in a specific domain and a time to trust all news from the most credible sources to preserve the authenticity of the news to spread the news through the network correct. Therefore, this research paper will be introduced a Multi-Dimensional credibility assessment to assess the credibility of the news source in a specific domain and time.

III. Multi-Dimensional Credibility Assessment Model

The proposed model assesses the credibility score for different Arabic news sources from multi-dimensional manner. The dimensions are: The credibility Score C_i, for a specific domain (Art, Health, Sport, Education and Politics) D and at a specific epoch of Time T using a hybrid – level (Topic- Post- Member) at a social-based information credibility type. The proposed model will pass on three main phases shown in Fig. 1. The first is the preparation of the dataset, the second is feature extraction phase, and the last one is measuring the credibility of the news source phase. These phases will be explained in details in the next sections.

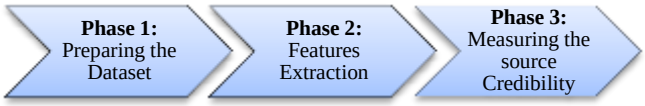


Fig. 1. The Proposed Model Phases.

A. Preparing the Dataset Phase

In this phase, the proposed model will use the "SMAD" dataset which is a new Arabic social media corpus collected and labeled for credibility from Facebook social media for the credibility assessment purpose. it consists of 15,240 textual topic of news collected from several websites (El-youm7, BBC Arabic, El-watan, etc.) scrapped for five domains Sports, Art, Education, Health and Political with size 2MB.

It is trained on 4200 textual topics in different domain in the training phase and about approximately 120 textual topics news item extracted from Facebook social media for testing phase. It preprocessed using Arabic Natural Language Processing (ANLP) and evaluated and gives accuracy of 98% using the TF-IDF classifier algorithm.

B. Features Extraction Phase

There are four types of Social–Based Information Credibility assessment (Post –Topic –Member –Hybrid levels) which can assess the credibility of them. This paper will use the hybrid level to give robust and more accurate information. The following tables, Table I and Table II show the features extracted at each level to be used in the calculation of the credibility scores.

1) Post features: The credibility of the content is the level of the believability of the content and the source which produces the news. The degree assigned to the post is an indication of the post believability. Table I illustrates four features of Facebook posts used for computing credibility. The reason for choosing these features is that they are the indicator for the user interaction of that post.

2) Member features: Table II illustrates four features of Facebook's user account used for computing credibility at the member level. The values of these features can be accessed via the puppeteer library.

C. Credibility Measurement Phase

Unfortunately, social network users are unable to directly observe how well someone is trusted in a particular domain as a source of credibility. The credibility will be assessed based on content trust and source trust in a particular domain of interest to maximize relevance, credibility, and the quality of the information received to consume the news.

Our model will use a uniform distribution to derive the significance of the news source page on Facebook. The calculating credibility score is based on the Simple Aggregated Score presented in [6] [7] [16] [19], where each score for each feature in a specific domain at the period of time for each news source will be added to provide an aggregated score. The credibility score will be calculated based on features and is rated on a discrete 3-point scale. The rating of the credibility of the news source is based on threshold values a0, a1, a2 with the following mapping: [0, a0) → Highly Credible (HC), [a0, a1) → Neutral (N), [a1, a2) → Highly Non-Credible (HNC).

Each news source NS has several features F. The significance will be calculated for each feature on Facebook posts in a specific domain D in an epoch of time T. Then commutate them to calculate the total significance. The features are like (e.g. number of likes, number of comments, number of shares, if the news source is verified or not, and the number of followers).

To calculate the significance of each feature S there are some procedures that must be applied. The first is rescaling the feature values using a min-max scaling algorithm is the process of transforming the extracted values to be in the range of [0,1]. Denoted by Eq. (1) on each feature.

$$S(F_L) = \frac{ki - kmin}{kmax - kmin} \quad (1)$$

Where k values will be collected for each feature for a specific domain post N at an epoch of time T, so the values are $\{k_1, \ldots .. \ldots, k_N\}$, k_i ($i \in \{1, \ldots, N\}$).

In the same manner, the calculation will be done in all features to get the significance of news sources. The values are $S(F_L)$, $S(F_C)$, $S(F_S)$. The second is calculating the total significance value as well as calculate the weighted significance of each feature.

To calculate the weight, first, calculate the dispersion of each feature value around its mean (standardization). Standardization is the transformation of each feature and it has a mean of 0 and a standard deviation of 1. The idea is that the closest a value is to the mean; the more reliable it is, whereas the farthest values are outliers. Standardization is applied by using scikitlearn's and StandardScaler packages. For a set of values $\{k_1, \ldots, k_i, \ldots, k_j, \ldots, k_N\}$ The dispersion of features can be denoted by the sample average by $k\mu$ and the sample standard deviation by s_k in Eq. (2).

TABLE I. POST AND TOPIC FEATURES

Feature Name	Feature Description
Originality	Has the same content been used in the past?
Majority	the target data will be analyzed by searching for the same type of sources to analyze how supportive of data in a set of the collected ones.
likes_count	The number of likes
comments_count	The number of comments
shares_count	The number of shared posts

TABLE II. MEMBER-LEVEL FEATURES

Feature Name	Feature Description
Has a verification badge	if verified = 1 if unverified = 0
Posts_count	The number of posts in a specific domain and time
Followers_count	The number of followers

$$d(F_L)= |\frac{k_i - k\mu}{s_k}| \quad (2)$$

The weight of (#of Likes) W feature F_L will be calculated by Eq. (3).

$$W(F_L) = \frac{(1 - d(F_l))}{(d(F_l) + d(F‹)+d(F_s))} \quad (3)$$

while that the weight of (#of Comments) W feature F_C will be will be calculated by Eq. (4).

$$W(F_C) = \frac{(1 - d(F‹))}{(d(F_l) + d(F‹)+d(F_s))} \quad (4)$$

and the weight of (#of Shares) W feature F_S will be calculated by Eq. (5).

$$W(F_S) = \frac{(1 - d(F_s))}{(d(F_l)+ d(F‹)+d(F_s))} \quad (5)$$

Finally, the total significance of the news source i (S_i) will be calculated by Eq. (6).

$$S_i = W(F_L)*S(F_L) +W(F_C)*S(F_C) + W(F_S)*S(F_S)+ F_{VERIFIED} + F_{NO\ OF\ FOLLOWERS} \quad (6)$$

The following algorithm illustrates how to calculate the credibility score for each news source NS in a specific domain D in time T.

Algorithm 1: CREDIBILITY SCORE Algorithm

Input: a set of values $\{k_1, \ldots, k_i, \ldots, k_j, \ldots, k_N\}$ where N is several posts on Facebook, k is a feature for a post for one source in a specific domain at the epoch of time

Output: Get the significance of source i and source j

1 For each feature K, I,.. for a specific news source i

2 Get k_{min} , k_{max}

3 Calculate the significance of $\bar{K}$, σ_k

4 Calculate the significance of k by S (Ki) = $\frac{k_i - k_{min}}{k_{max} - k_{min}}$

5 Calculate the dispersion of k by $d(k_i)= |\frac{ki - \bar{\bar{K}}}{s_k}|$

6 Calculate the weight of source i by $w(k_i) = \frac{(1-d(k_i))}{(d(k_i)+d(I_i))}$

7 End for

8 Calculate the total significant of news source i by Si = w(ki) ∗ S(ki) + w(li) ∗ S(li)

9 RETURN S_i

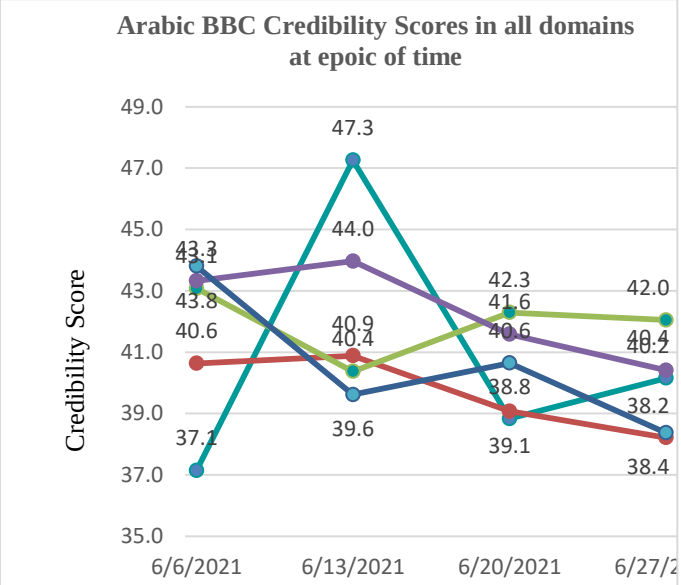
IV. Experimental Results

Now, to apply the proposed assessment model we scrap posts from Facebook along "June" month for five news sources " (El-youm7) اليوم السابع, (Arabic BBC) BBC عربى, الوطن(El-watan), (Akhbar El-youm) اخبار اليوم and سكاى نيوز (Sky news) "as a sample of news sources in five domains (Education, Art, Health, Sport, and Political). For (Akhbar El-youm) اخبار اليوم new source, we scrap using pupputeer library about 2200 post in all domains to calculate its credibility. In "Arabic BBC " news source , we scrap about 1200 post. In (El-youm7) اليوم السابع new source, we scrap about 2250 post. In الوطن(El-watan) news source, we scrap 2510. In سكاى نيوز (Sky news) news source, we scrap about 2400 post. For testing the result for classification posts into domains using TF-IDF and KNN algorithms, the experiment results show the algorithm accuracy is 0.98%. The following figures will visualize the credibility scores over different data analysis for different levels and investigate the credibility score at different domains and visualize these scores into graphs.

In the first level of data analysis, the credibility score will be calculated for a specific news source in all different domains at any period of time. Fig. 2 depicts the credibility score for the "Arabic BBC" page on Facebook social media through the four weeks of June, in all domains (Education, Art, Health, Sport, and Political).

Also, Fig. 3 shows the credibilty score at a point of time applied on اليوم السابع (El-youm7) page.

The second level of data analysis, the credibility scores Cxi for all news sources NSi in all different domains Di at a point of time T. Fig. 4 depicts the credibility scores for the five news sources " (El-youm7) اليوم السابع, (Arabic BBC) BBC عربى, الوطن(El-watan), (Akhbar El-youm) اخبار اليوم and سكاى نيوز (Sky news) "as a sample of news sources in five domains (Education, Art, Health, Sport, and Political) to show the highest and lowest news source in each domain.



	6/6/2021	6/13/2021	6/20/2021	6/27/2021
Art	37.1	47.3	38.8	40.2
Education	40.6	40.9	39.1	38.2
Medicine	43.1	40.4	42.3	42.0
Political	43.3	44.0	41.6	40.4
Sports	43.8	39.6	40.6	38.4

Fig. 2. Credibility Score of a Specific News Source.

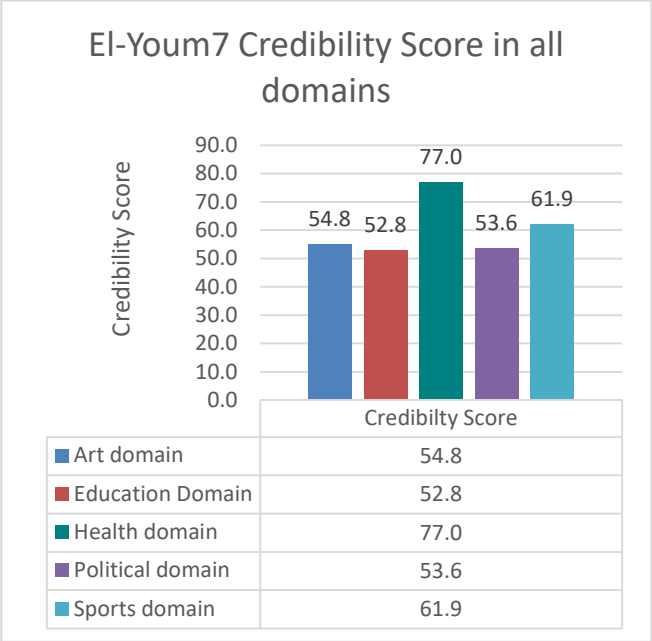


Fig. 3. Credibility Score for a News Source at Point of Time.

The third level of data analysis, the credibility scores for all different news sources at a certain period of time in specific domain. The following figures depict the variation of credibility scores over time for five news sources such as اليوم السابع (El-Youm7), (Arabic BBC) BBC عربى, (El-Watan) الوطن, اخبار اليوم (Akhbar El-youm) and (Sky News) سكاى نيوز in same domain.

Fig. 5 shows the investigation for all news sources in June in the Art domain. The news source (El-Watan) "الوطن" has the highest credibility score in the first week. While the news source (El-Youm7) "اليوم السابع" has the highest credibility score in the second week. In the third week, the news source "اليوم السابع" (El-Youm7) is the highest. Finally, in the last week of the month, the highest one is "اليوم السابع" (El-Youm7). Thus, the overall investigation of data analysis of Fig. 5 the most credible source in the Art domain over the last four weeks is the "اليوم السابع" (El-Youm7) news source. The least credible source in the Art domain over the last four weeks is the (Akhbar El-youm) "أخبار اليوم" news source.

Fig. 6 shows the credibility scores of all news sources over a period of time in the Education domain. The highest credible score is the (El-Youm7) "اليوم السابع" news source. Therefore, it is the most credible one during this period. The lowest credible source is the "أخبار اليوم" (Akhbar El-youm) news source.

Therefore, it is the least credible Fig. 7 shows the credibility scores of all news sources over a period of time in the political domain, while Fig. 8 shows the credibility score in the health domain.

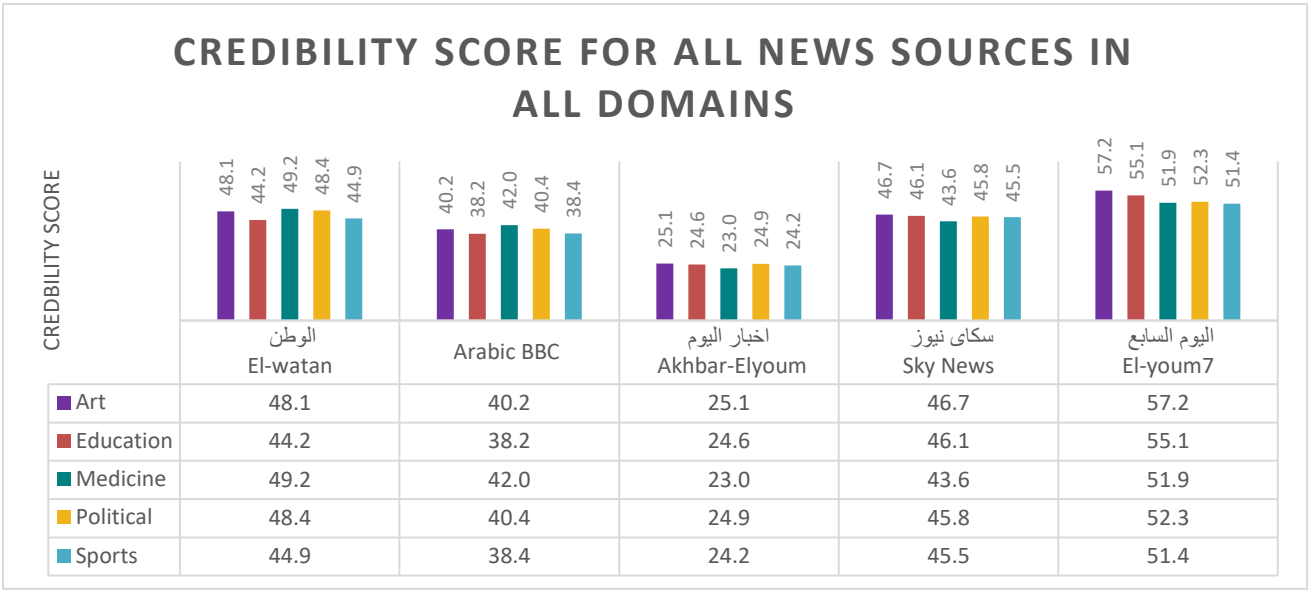
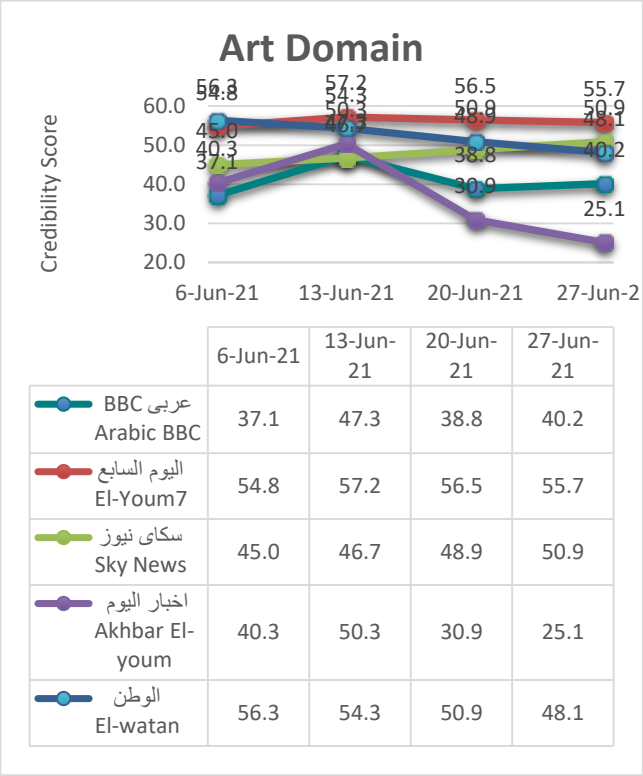
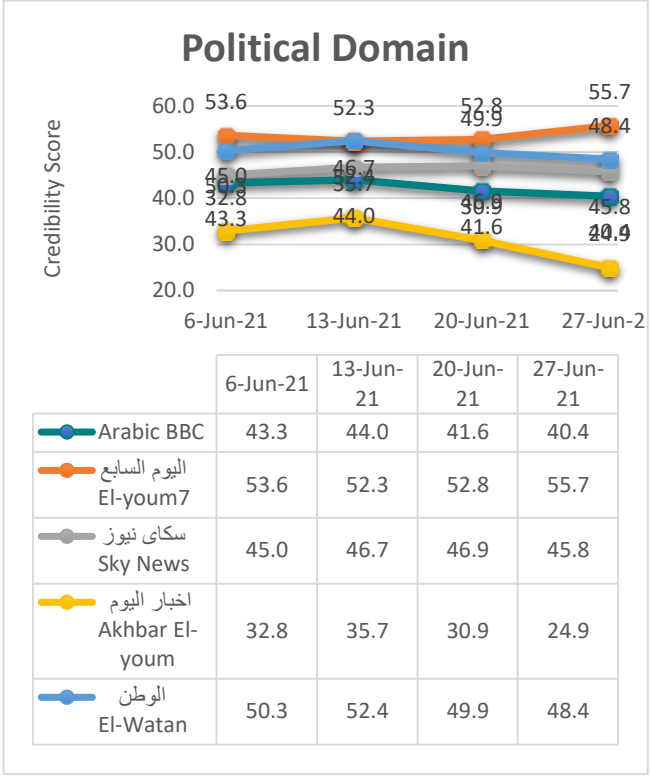


Fig. 4. Credibility Scores for News Sources in All Domains at a Point of Time.



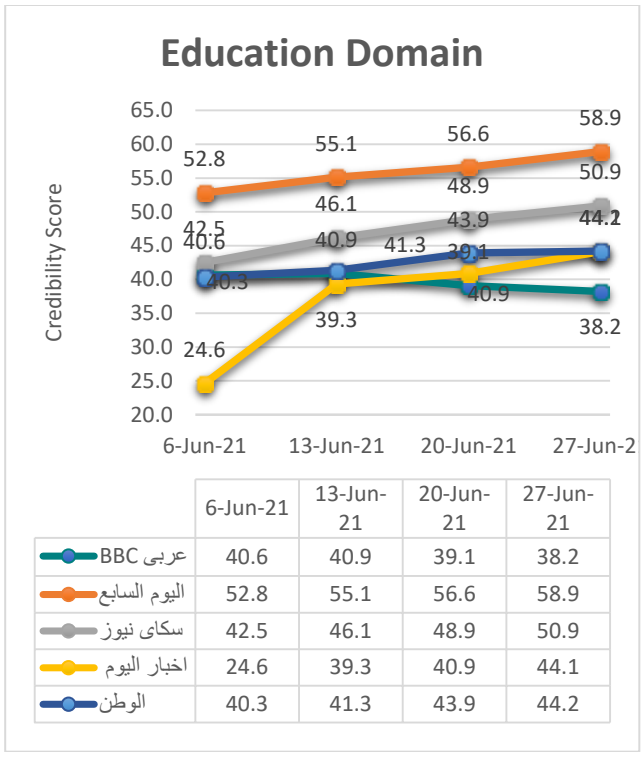
	6-Jun-21	13-Jun-21	20-Jun-21	27-Jun-21
BBC عربى Arabic BBC	37.1	47.3	38.8	40.2
اليوم السابع El-Youm7	54.8	57.2	56.5	55.7
سكاى نيوز Sky News	45.0	46.7	48.9	50.9
اخبار اليوم Akhbar El-youm	40.3	50.3	30.9	25.1
الوطن El-watan	56.3	54.3	50.9	48.1

Fig. 5. The Credibility Score for All News Source in Art Domain at Epoch Time.



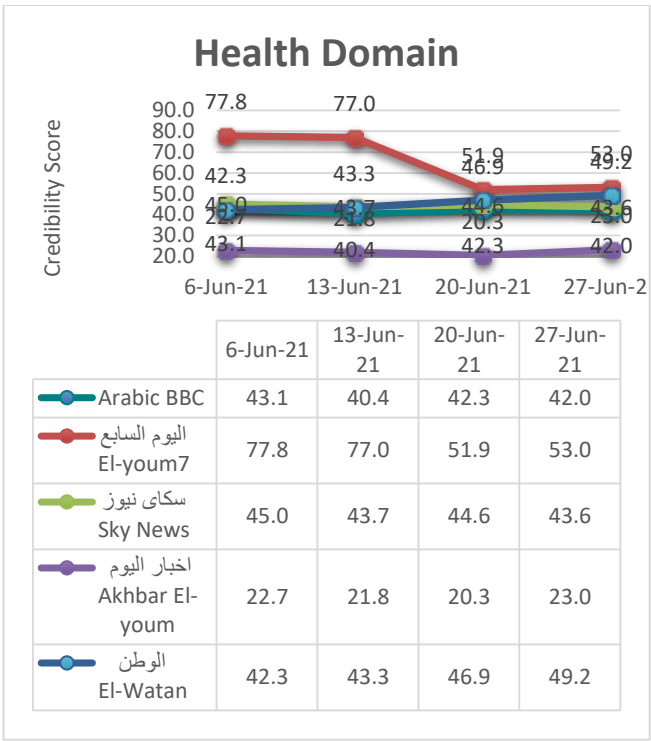
	6-Jun-21	13-Jun-21	20-Jun-21	27-Jun-21
Arabic BBC	43.3	44.0	41.6	40.4
اليوم السابع El-youm7	53.6	52.3	52.8	55.7
سكاى نيوز Sky News	45.0	46.7	46.9	45.8
اخبار اليوم Akhbar El-youm	32.8	35.7	30.9	24.9
الوطن El-Watan	50.3	52.4	49.9	48.4

Fig. 7. The Credibility Score for All News Source in Political Domain at Epoch Time.



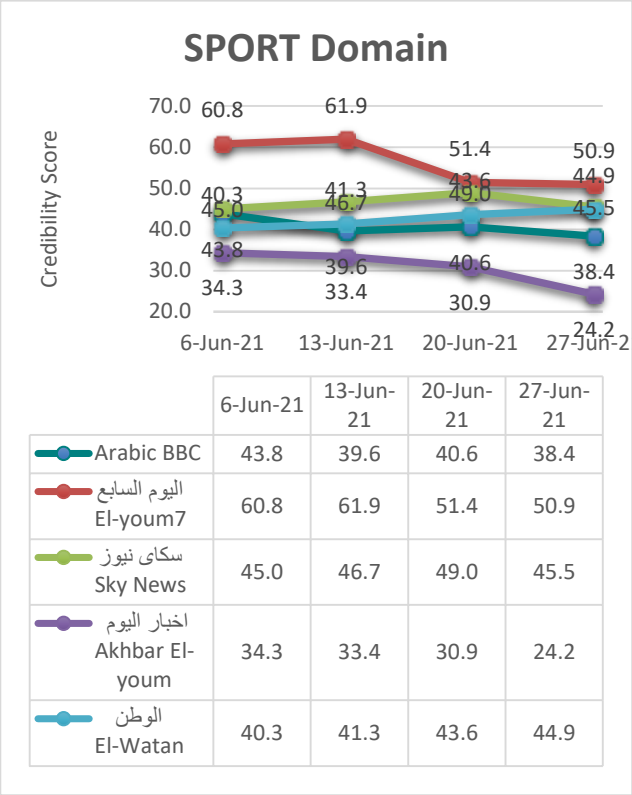
	6-Jun-21	13-Jun-21	20-Jun-21	27-Jun-21
BBC عربى	40.6	40.9	39.1	38.2
اليوم السابع	52.8	55.1	56.6	58.9
سكاى نيوز	42.5	46.1	48.9	50.9
اخبار اليوم	24.6	39.3	40.9	44.1
الوطن	40.3	41.3	43.9	44.2

Fig. 6. The Credibility Score for All News Source in Education Domain at Epoch Time.



	6-Jun-21	13-Jun-21	20-Jun-21	27-Jun-21
Arabic BBC	43.1	40.4	42.3	42.0
اليوم السابع El-youm7	77.8	77.0	51.9	53.0
سكاى نيوز Sky News	45.0	43.7	44.6	43.6
اخبار اليوم Akhbar El-youm	22.7	21.8	20.3	23.0
الوطن El-Watan	42.3	43.3	46.9	49.2

Fig. 8. The Credibility Score for All News Source in Health Domain at Epoch Time.



	6-Jun-21	13-Jun-21	20-Jun-21	27-Jun-21
Arabic BBC	43.8	39.6	40.6	38.4
اليوم السابع El-youm7	60.8	61.9	51.4	50.9
سكاى نيوز Sky News	45.0	46.7	49.0	45.5
اخبار اليوم Akhbar El-youm	34.3	33.4	30.9	24.2
الوطن El-Watan	40.3	41.3	43.6	44.9

Fig. 9. The Credibility Score for All News Source in Sport Domain at Epoch Time.

Finally, Fig. 9 despites the credibility score of the Sports domain for all the news sources over one month.

V. Discussion

Prior studies [3-24] have presented the importance of assessing the credibility of the news source and posts. Most of the mentioned work, the most popular social media platform used in the information credibility assessment is Twitter, due to its ease of use and because it has a 140-280 characters' limit in tweets. The second one is the web content or web sources' information credibility assessment, the third is microblogging and weblogs, and the last one is Facebook. While Facebook is the most popular social network worldwide, where there are 2.89 billion monthly active users, However, there is one tool that works on Facebook to assess the credibility assessment [13]. It assesses the credibility at the post-level and uses crowdsourcing to justify its results.

Different from the above, our model assesses the credibility of Arabic news sources using hybrid levels (Post-Topic-Member) and works especially on the Arabic language due to the complexity of the essential structure of the Arabic language used as a challenge. Our model assesses credibility along three dimensions. The first dimension involves assessing the credibility of Arabic news sources automatically. The second dimension, the credibility of the news sources will be calculated in a specific domain (Art-Health-Education-Politics-Sports) to be more accurate and help the audience or consumers of the news to support these sources or not and to be guidance and trustable to them to rely on. The final dimension that should be taken into consideration is that news sources can vary their credibility scores within a period of time. Thus, our model assesses the credibility of news sources in a specific domain at an epoch of time with an accuracy of 98%. Finally, our model also, created a new dataset called "SMAD" that contains 5000 Arabic news items classified from Facebook social media. The SMAD corpus gives accuracy of about 98% in five domains. Each domain has been estimated and compared with other different datasets according to quality measurement metrics (precision, recall, F-measure and accuracy). In the sports domain, its precision is 0.95, recall is 1 and the F-measure is 0.98. In the education domain, its precision is 0.98, recall is 1 and the F-measure is 0.99. In the arts domain, its precision is 1, recall is 0.96 and F-measure is 0.98. In the health domain, its precision is 0.99, recall is 0.97 and F-measure is 0.98. Finally, the political domain precision is 0.98, recall is 0.97, and F-measure is 0.98.

VI. Conclusion and Future Work

To sum up, the present study constructs a new model assessing the credibility of Arabic news sources on Facebook social media in a multidimensional manner for news sources, in different domains, and at different epochs of time. This model will be visualizing this data in different ways from the existing benchmark dataset for the news to calculate the credibility scores quickly related to topics from social media posts. This model visualizes the results at different data analysis levels: visualizes the credibility assessment scores of El-youm7 at a specific epoch of time in different domains (Art, health, Education, Political and sports), visualizes the five Arabic news sources in all domains at a specific epoch of time and finally visualizes each domain with respect to the scores of five Arabic news sources at a specific epoch of time.

There are number of open issues can be considered in future: First one, construct an Arabic Fact-checking website to visualize the credibility scores to be readable for online users and help them to retrieve the most credible news sources to rely on them. Second, use Semantic similarity analysis methods to detect near-duplicated and similar news retrieved from different sources about a certain claim in a specific domain and a specific period of time should be handled using Natural Language Processing (NLP) methods. Third, sentiment analysis can be taken into consideration to analyze positive and negative comments to get the credibility scores more accurate and precise. Finally, use the Blockchain technology to assess the credibility of each news source in its domain at a time and retrieve all the true news from the most credible sources, which will join into the creditable blockchain network to preserve the authenticity of the news and spread it through the network correctly.

References

[1] G. Pasi, G and M. Viviani, "Information credibility in the social web: Contexts, approaches, and open issues", 2020, *arXiv preprint arXiv:2001.09473*.

[2] M. Alrubaian, M. Al-Qurishi, A. Alamri, M. Al-Rakhami, M.M., Hassan and G. Fortino, "Credibility in online social networks: A survey". "IEEE Access", Vol: 7, pp.2828-2855, 2018.

[3] K. Popat, S. Mukherjee, J. Strötgen and G. Weikum, "Where the truth lies: Explaining the credibility of emerging claims on the web and social

media", In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 1003-1012, 2017, April.

[4] F. Ahmad and S. A. M. Rizvi," Information Credibility on 2 Using Machine Learning Techniques"," Journal: Futuristic Trends in Networks and Computing Technologies Communications in Computer and Information Science", pp. 371-381, 2020.

[5] G. Pasi and M. Viviani, "Information credibility in the social web: Contexts, approaches, and open issues", 2020, arXiv preprint arXiv:2001.09473.

[6] E. Jaho, E. Tzoannos, A. Papadopoulos and N. Sarris," Alethiometer: a framework for assessing trustworthiness and content validity in social media", In: Proceedings of the 23rd International Conference on World Wide Web, (pp. 749-752), 2014, April.

[7] S. T. Moturu and H. Liu," Quantifying the trustworthiness of social media content"," Distributed and Parallel Databases", Vol. 29, No. 3, pp. 239-260, 2011.

[8] S. Mahmood, A. Ghani, A. Daud and S. Shamshirband, "Reputation-Based Approach Toward Web Content Credibility Analysis", "IEEE Access", Vol.:7, pp. 139957-139969, 2019.

[9] M. Gupta, P. Zhao and J. Han," Evaluating event credibility on twitter", In: Proceedings of the 2012 SIAM International Conference on Data Mining Society for Industrial and Applied Mathematics, pp. 153-164, 2012, April.

[10] M. A. Abbasi and H. Liu," Measuring user credibility in social media", In: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, Springer, Berlin, Heidelberg, pp. 441-448, 2013, April.

[11] A. Gupta, P. Kumaraguru, C. Castillo and P. Meier,"Tweetcred: Real-time credibility assessment of content on twitter", In: *International conference on social informatics*, Springer, Cham, pp. 228-243, 2014, November.

[12] T. Mitra and E. Gilbert, "Credbank: A large-scale social media corpus with associated credibility annotations", In: *Ninth international AAAI conference on web and social media*, 2015, April.

[13] K. R. Saikaew and C. Noyunsan, "Features for measuring credibility on Facebook information. "International Scholarly and Scientific Research & Innovation", vol.: *9, No.* 1, pp.174-177 ,2015

[14] B. D. Horne, W. Dron, S. Khedr, and S. Adali, "Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news", In: Companion *Proceedings of the The Web Conference 2018*, pp. 235-238, 2018, April.

[15] D. Saez-Trumper," Fake tweet buster: a webtool to identify users promoting fake news ontwitter", In: *Proceedings of the 25th ACM conference on Hypertext and social media*, pp. 316-317, 2014, September.

[16] K. Lorek, J. Suehiro-Wiciński, M. Jankowski-Lorek and A. Gupta," Automated credibility assessment on twitter". "*Computer Science"*, Vol.:16, No.2, pp. 157-168, 2015.

[17] V. Podobnik, D. Striga, A. Jandras and I. Lovrek, "How to calculate trust between social network users?", In: *SoftCOM 2012, 20th International Conference on Software, Telecommunications and Computer Networks*, IEEE, pp. 1-6, 2012, September.

[18] S. Bauskar, V. Badole, P. Jain and M. Chawla," Natural language processing based hybrid model for detecting fake news using content-based features and social features"," International Journal of Information Engineering and Electronic Business", Vol.11, No.4, p. 1,2019.

[19] G. Jardaneh, H. Abdelhaq, M. Buzz and D. Johnson, "Classifying Arabic tweets based on credibility using content and user features", In :*2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), IEEE,* pp. 596-601, 2019, April.

[20] C. Helwe, S. Elbassuoni, A. Al Zaatari and W. El-Hajj," Assessing Arabic weblog credibility via deep co-learning" In: Proceedings of the Fourth Arabic Natural Language Processing Workshop", pp. 130-136,2019, August.

[21] R. M. B. Al-Eidan, H. S. Al-Khalifa and A. S. Al-Salman, "Towards the measurement of arabic weblogs credibility automatically", In : *Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services*, pp. 618-622, 2009, December.

[22] A. Y. M. Floos, "Arabic rumours identification by measuring the credibility of arabic tweet content", In: *Media Controversy: Breakthroughs in Research and Practice*, IGI Global, pp. 236-248, 2020.

[23] R. Mouty and A. Gazdar," The effect of the similarity between the two names of twitter users on the credibility of their publications" , In "*2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)* , IEEE, pp. 196-201, 2019, May.

[24] A. Al Zaatari, R. El Ballouli, S. ELbassouni, W. El-Hajj, H. Hajj, K. Shaban , ... and E. Yahya," Arabic corpora for credibility analysis", In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* , pp. 4396-4401 , 2016, May.

[25] R. El Ballouli, W. El-Hajj, A. Ghandour, S. Elbassuoni, H. Hajj and K. Shaban, "Cat: Credibility analysis of arabic content on twitter", In: *Proceedings of the Third Arabic Natural Language Processing Workshop*, pp. 62-71, 2017, April.

Online Programming Semantic Error Feedback using Dynamic Template Matching

Razali M.K.A, S. Suhailan, Mohamed M.A, M.D. M. Sufian
Faculty of Informatics and Computing
Universiti Sultan Zainal Abidin (UniSZA)
Terengganu, Malaysia

***Abstract*—Many of automated computer programming feedback is generated based on static template matching that need to be provided by the experts. This research is focusing on developing an automated online programming semantic error feedback by using dynamic template matching models based on students' correct answers submission. Currently, there is a lack of research using dynamic template matching model due to their complexity and varies in terms of programming structure. To solve the formulation of the dynamic templates, a new automated feedback model using front and rear n-gram sequence as the matching technique was developed to provide feedback to students based on the missing structure of the best-matched template. We have tested 60 student's Java programming answers on 3 different types of programming questions using all the dynamic templates randomly chosen for each student. An expert was assigned to manually match the student's answer with the 3 randomly chosen templates. The result shows that 80% of the best-matched templates for each student using the technique were similarly chosen by the expert. Based on the matched template, the student will be given feedback notifying the possible next programming instruction that can be included in the answer to get it correct as was achieved by the template. This model can contribute to automatically assist students in answering computational programming exercises.**

***Keywords*—Dynamic; feedback; online programming; semantic error; template matching**

I. Introduction

Computer science is a discipline that involves the understanding and design of computers and computational processes, including their theory, analysis, software and hardware design, efficiency, implementation, and application and effect on society [1]. In other words, computer science is an emergent, scientific and practical method, which deals with the theoretical basis of information and computation, and combines its realization and application technology [2]. Computer programming is one of the core subjects that every computer science student must be competent to become a good programmer. Therefore, to obtain the programming skill, lots of programming exercises need to be completed [3].

Students need to develop programming logic and thinking skills to understand and solve the tasks especially on code writing. Students also need to solve any encountered programming errors in their coding regarding the syntax, semantic, and also question requirements. From there, students will get the knowledge and experiences on how to encounter any common programming errors or mistakes. Learning through practice is the best way to learn computer programming and attract novice students [4].

Unfortunately, most computer science students face difficulties in learning computer programming especially in writing the programming scripts [5]. Despite the importance of computer science, there is a high percentage of failures and dropout rates in introductory programming courses recorded by most educational institutions around the world [2]. Lecturers must also be responsible for assisting and providing some feedback to their students to resolve students' misunderstandings or mistakes. Helping a large number of students in providing personalized feedback during programming exercises will be a difficult role for teachers [6].

Furthermore, there are a lot of automated programming assessment tools with automated feedback that have been continuously developed to help students practice programming and build up logic skills and also programming syntax [7]. By using any automated programming tools, a student can submit a computer program on a problem-solving exercise while the tool will promptly produce automated feedback to highlight any encountered errors or mistakes during the compilation or implementation of the program [8]. The error is produced by the compiler known as Syntax Error. The compiler will highlight which lines that contain errors. However, for a beginner student, the syntax error j does not explain on how to fix the code in solving the question problem. This research is focusing on developing an online programming semantic error feedback by using a dynamic template matching model.

II. Related Work

The teacher-student ratio can reach thousands to one by implementing the advancement of Massive Open Online Courses (MOOCs) [14]. This makes the feedback design more specific and personalized. Unfortunately, providing manual teacher feedback for programming assignments is determined as a traditional method and it is no longer suitable for MOOCs. Current automatic feedback methods have some weaknesses, such as the inability to extend to larger programs, manual teacher involvement, and lack of accuracy in determining errors.

There are two techniques to design the programming feedback; static and dynamic approaches. Static approaches identify and study the source code without running the computer program [9]. It is used to evaluate the syntax and semantic error and programming style. The dynamic approach

is based on the execution of the computer program [10]. It is used to evaluate run-time errors, programming design, and software metrics such as timing and resources utilization.

For beginners, static feedback is crucially needed in helping them to visualize the logic of the computer program in solving a question. As to master the programming skills, lots of exercises need to be completed by a student. With the advance of an e-learning platform, many platforms offer programming exercises to be done online. There are a lot of programming tools where users can learn and train their programming skills by solving the given problem with some programming code to find the best solution for that problem [11]. Most of these tools are developed as web applications. Some of these tools are CodingBat [12], betterprogrammer, Practice-It, and CodeWorkout [13]. By using these systems, users can get feedback about their submitted answers because these systems already provide a set of practical programming problems to be solved in the web browser and the results are evaluated by checking them against unit tests or test cases. Unfortunately, this dynamic feedback is difficult to be understood by the beginner who wants to start learning the logic or flow of the programming. The feedback is general in highlighting how the output should be generated. Writing hints and preparing the feedback in this way needs meta cognition and involves critical thinking which is not yet developed among the beginners.

The novice programmer tries to imitate the steps prepared by the teacher, and some errors that the novice programmer could not solve appeared during compilation [15]. One of the challenges in writing coding for novice programmers is insufficient feedback error messages. The only feedback that is available is the compiler-based error on the syntax errors [16]. Therefore, the best compiler errors are those that can deliver important messages that are desperately needed by programmers in response to all the errors they make. Decaf is a Java editor that serves as a medium for improving javac compiler error messages. An error message will be produced by the compiler if there are some errors contains in the student's source code. Then, the error codes and error messages are analyzed to produce enhanced error messages that provide more valuable information to students, in the hope that the error can be corrected more effectively as compared to the ordinary error messages alone.

With the existence of the standard error and enhanced errors, students can avoid making the same error in the future by referring to the both types of error Decaf is an enhanced compiler error message as shown in Fig. 1 that elaborate the common syntax error produced by Java.

However Decaf only provide feedback in clarifying the error related to the programming syntax. A logic error which is part of the semantic error is not presented in most of the compilers as it depends on the individual question requirements.

A semantic error feedback is meant to provide feedback based on specific question requirements using a solution template [17]. A template consists of a correct program instructions sequence (keywords, symbols, numbers). This computer programs need to be converted to certain features numbers before it can be processed as a matching template. [8] was using the instruction ratio (IGR) and the instruction count ratio (ICR) as the features to represent the computer program. IGR is the ratio of sequential instructions or symbol sequences in a program to instructions or symbol sequences of templates with some skippable instructions. Meanwhile, ICR is the average ratio of the amount of all unique instructions within the program that matches the amount of all unique instructions laid out in the template. Based on these features, the K-Means algorithm was used to assign similar computer programs to certain clusters. Based on each cluster, programs that have a similar Euclidean distance to the centroid in the cluster are grouped. These groups represent unique rules that will be associated with semantic feedback. Under this rule, an expert will add an assisted feedback to add comments of what further actions need to be done in solving the question.

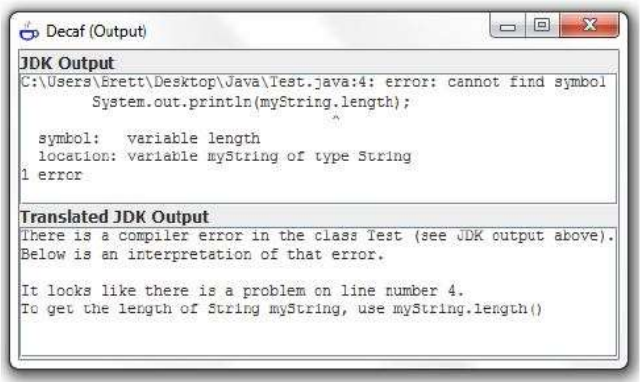


Fig. 1. Example of an enhanced Compiler Error Message Produced by Decaf.

However, the technique needs to design the enhanced feedback manually to make sure their student understands the way on how to fix the error in their code. It needs to be done on pre-defined templates. This method requires more resources from the teacher not only to prepare the template but also need to manually assign students' program clusters with feedback from time to time. This research further enhanced this technique by making the matching template more accurate by comparing forward and reverse sequences of the n-gram algorithm. It also provides automated feedback based on the missing instruction sequence based on the selected dynamic templates mining from the correct submission answers from other students.

III. Methodology

The N-gram model is improved by calculating the sequence N-gram in two different ways. The first approach is using front N-grams where the value is gained by calculating the matched sequences based on the forward parsing of the codes. The second approach is using rear N-grams where the value is gained by calculating matched sequences based on the reverse parsing of the codes.

The combination of front and rear N-gram values which are referred to as the FR-Grams model are then used to enhance the similarity finding technique between two different programs. Fig. 2 shows the framework of generating the semantic error feedback using FR-Grams.

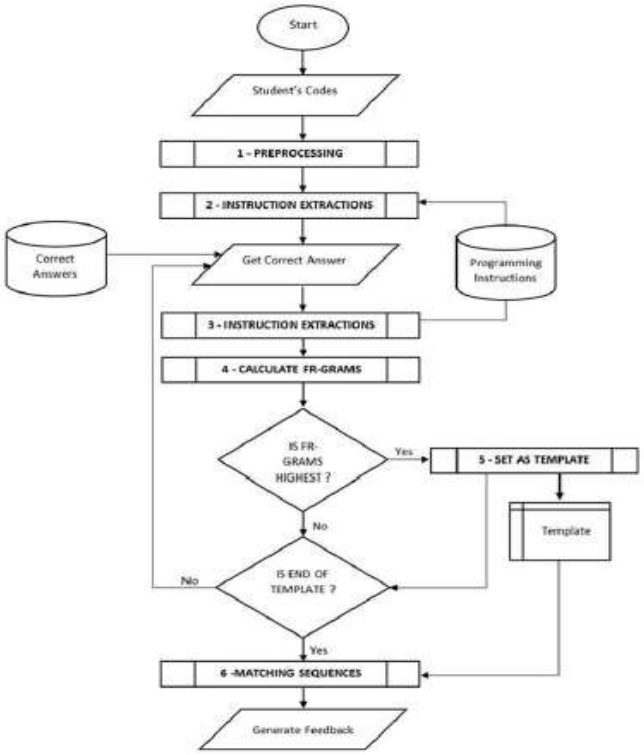


Fig. 2. Semantic Error Feedback Framework using FR-Grams Technique.

FR-Grams Model is then used to match a student's program with the dynamic program templates. The templates were auto-generated based on the list of correct answers of the submitted student's program. The answers were graded automatically by executing the answers and matched with the expected test cases. Based on the best-matched program templates between the student's program, the student will get feedback notifying any missing program instruction sequence that needs to be added to the program in order to get the correct answers according to the selected template.

A. Step 1: Pre-processing

The library of computer instructions keywords needs to be prepared first before the computer program instruction's extraction can be generated. For example, a list of Java instructions or keywords are "int", "if", "string", "nextInt()", "public" and others. These keywords are inserted into a file to be used as a key point to filter the submitted answers. In the pre-processing process, any essential code before the program body needs to be removed. In java, these codes are something like "import java.util.Scanner", "public static void main". This framework only considers the main program body.

B. Step 2: Instruction Extractions from Student's Answer

Student programs need to be converted as a sequence of instructions. The sequence of Java instructions from a student example answer shown in Fig. 3 will be produced as the following: -

"Scanner", "System", "String", "next()", "int"

```
1  import java.util.Scanner;
2  public class Q4 {
3    public static void main(String[] args) {
4      Scanner k = new Scanner(System.in);
5      String name = k.next();
6      int age = k.next();
7      int balance = 50-age;
8      System.out.println("Hi, " + name +"! Your age is "+ i +".");
9    }
10 }
```

Fig. 3. Sample of a Student's Program.

Then, the total instruction (N) is calculated based on the sequence of the program instructions.

C. Step 3: Instruction Extractions from Template

Templates are the collection of successful and accepted computer program submissions made by the previous student attempts to a question. These templates need to be converted into a sequence of instructions similar to the Step 1 to 2.

D. Step 4: Calculate FR-Grams

The number of FR-grams from student answers and correct answers is compared to calculate the sequence of programming. The algorithm to calculate N-grams is given in Fig. 4.

E. Step 5: Find the Best Template

All the correct students' answer stored in the database will be selected as the dynamic template matching. A student's attempt answer will be matched with these templates. The highest total FR-grams among them will be the considered as the most accurate template for further feedback generation.

F. Step 6: Feedback Generation

After the comparison, the total FR-grams value is produced based on the template selected in step 4. After finishing comparing the answers, the FR-Grams are calculated to get the total FR-grams for each template. The highest total will be processed as the feedback template.

```
1:   N = number of unique instructions (I') in the template
2:   NGRAM = 0
3:   for i=1 to N do
4:       SA = Student answer
5:       CA = Correct answer
6:       if SA == CA then
7:           NGRAM++
8:       else
9:           N++
10:      end if
11:      N++
12:  end for
13:  return NGRAM
```

Fig. 4. Algorithm to Calculate N-grams.

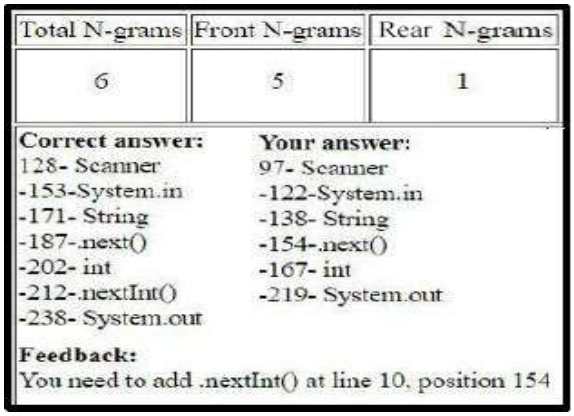


Fig. 5. Automated Feedback based on Template Matching.

Fig. 5 shows that the student's attempt consists of instructions sequence of "Scanner", "System.in", "String", ".next()", "int" and "System.out". While comparing to the template, a feedback will be generated by the system notifying that "You need to add .nextInt() at line 10, position 154". This is the missing instruction sequence that the student needs to add for the program to be tailored to the template. This feedback can provide some clues for the student on how to proceed and make the correction to the program.

IV. Result and Analysis

There were 60 student's Java programming answers were tested using all the dynamic templates. The answers were based on 3 different set of programming questions. The templates were randomly chosen for each student's attempt by the system. Table I is the sample feedback that was generated by the system along with the template chosen by the system.

An expert was assigned to manually match the student's answer with the three randomly chosen templates. The experts are chosen based on their experience in validating the source code and marking the student's programming answer. For each student's answer, the expert will be presented with the three template that have highest total FR-grams to be compared with the student's answer. The expert was provided with a rubric in order to choose which template should the student's attempt be referred most.

1) Check for similar variables: In a student's answer, many variables contain in the source code such as "string", "int", "char", "for" and others. If the student's answer contains a "string" variable, the experts will search this variable in the correct answers to get the most accurate template.

2) Check for quantity and type of variable: If the student's answer contains a "string" variable but in the template use "string[]" which is a string array type, the template will not be chosen.

3) Check for the simpler with the student's answer: If each template has passed the first and second rules which means that there is almost similarity between the templates, the experts will choose the simplest template according to the student's answer.

TABLE I. Sample Feedback by System

Student ID	Answer	Best Feedback	Result

01	import java.util.Scanner;
	public class Q2 {
	public static void main(String[] args) {
	Scanner k = new Scanner(System.in);
	String a="Apology";
	char [] b=new char[7];
	int i=a.length();
	System.out.print("*");
	System.out.print(a.charAt(1)+"*****");
	}
	}
		You need to
		add next()
		at line 5,
		position 43
			True

The result shows that there were 48 out of 60 or 80% similar decision made by the model and the expert. 9 from the 12 answers that were not matched with the expert's decision was due to the small difference of the total FR-Grams (only one missing sequence different) among the templates. On the other hand, these cases will also contribute difficulty for the expert to decide which template should be considered as the most matched. Based on the matched template, the student will be given feedback notifying the possible next programming instruction that can be included in the answer to get it correct as was achieved by the template. This will be like personal coaching to help students recover from any cluelessness on the programming command sequences to answer computational programming exercises.

However, there was a weakness for this sequence-based model as it does not recognize the template based on the data type usage. For example, answer that was using array data type should have only seek template that using the same data type. This will be the future research works need to be conducted in identifying template context in order to get the best template matching for a more accurate feedback.

V. Conclusion

In conclusion, the best semantic error feedback for the student should meet these criteria:

1) Can guide the student on what is missing.

2) Can highlight to the student what student needs to include in the source code to fix the error.

With this semantic error feedback, students can get a valuable idea or hint to solve the error in their source code. The critical thinking skills of students will increase based on computational logic skills practice tools. The system continuously collects feedback as a repository which eventually fully automated interactive assisted learning system can be achieved. Lastly, the student will keep interested and motivated to self-practice programming exercises towards problem-solving skill development.

Acknowledgment

Special thanks to the Ministry of Higher Education Malaysia and Universiti Sultan Zainal Abidin (UniSZA) for providing equipment and supporting this research project under the grant number of UNISZA/2018/GOT/03.

REFERENCES

[1] Tucker, A. (Ed.). (2006). A model curriculum for K-12 computer science: Final report of the ACM K-12 task force curriculum committee (2nd ed.). New York: Association for Computing Machinery (ACM).

[2] Queiros, Ricardo. (2014). Innovative teaching strategies and new learning paradigms in computer programming. 10.4018/978-1-4666-7304-5.

[3] Kwiatkowska, M., 2016. Measuring the Difficulty of Test Items in Computing Science Education. In: Proceedings of the 21st Western Canadian Conference on Computing Education, BC, Canada, 6 - 7 May 2016. ACM Press.

[4] Gross, P., & Powers, K. (2015). Evaluating assessments of novice programming environments. In Proceedings of the First International Workshop on Computing Education Research (pp. 99-110). New York: ACM. doi:10.1145/1089786.1089796.

[5] Anthony Robins, Janet Rountree and Nathan Rountree (2003). Learning and Teaching Programming: A Review and Discussion. Computer Science Education, Vol. 13, No. 2, pp. 137–172.

[6] S. Suhailan, M.K. Yusof, A.F.A. Abidin, S.A. Fadzli, M.S. Mat Deris and S. Abdul Samad (2018). Automated Ranking Assessment based on Completeness and Correctness of a Computer Program Solution. International Journal of Engineering & Technology, 7 (3.28) (2018) 278-283.

[7] S. Suhailan, S. Abdul Samad, M.A. Berhannuddin (2015). A perspective of Automated programming error feedback Approaches in problem solving exercises. Journal of Theoretical and Applied Information Technology. 70(1) 121-129.

[8] S. Suhailan, M.S. Mat Deris, S. Abdul Samad, M.A. Burhanuddin (2019). A Recommended Feedback Model of a Programming Exercise Using Clustering-Based Group Assistance. International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-7, Issue-5S4.

[9] Ala-Mutka, K. M. (2005). A survey of automated assessment approaches for programming assignments. Computer science education, 15(2), 83-102.

[10] Adidah Lajis, Shahidatul Arfah Baharudin, Diyana Ab Kadir, Nadilah Mohd Ralim, Haidawati Mohd Nasir and Normaziah Abdul Aziz (2018). A Review of Techniques in Automatic Programming Assessment for Practical Skill Test.

[11] Ashlesha Patil (2010). Automatic Grading of Programming Assignments.

[12] Priyanka Mohan (2015). Student Perceptions of Various Hint Features while Solving Coding Exercises.

[13] Kevin Buffardi and Stephen H. Edwards (2014). Adaptive and Social Mechanisms for Automated Improvement of eLearning Materials.

[14] Ke Wang, Benjamin Lin, Bjorn Rettig, Paul Pardi, and Rishabh Singh (2017). Data-Driven Feedback Generator for Online Programing Courses.

[15] Aniket Bhawkar, Rohit Belsare, Fenil Gandhi and Pratiksha Somani (2013). Analysis of Errors - A Support System for Teachers to Analyze the Error Occurring to a Novice Programmer.

[16] Brett A. Becker (2016). An Effective Approach to Enhancing Compiler Error Messages.

[17] S. Suhailan, S. Abdul Samad, M.A. Berhannuddin. Nazirah (2017). Program Statement Parser for Computational Programming Feedback. Journal of Engineering and Applied Sciences, 12(5S) 7057-7062.

Comparing MapReduce and Spark in Computing the PCC Matrix in Gene Co-expression Networks

Nagwan Abdel Samee[1], Nada Hassan Osman[2], Rania Ahmed Abdel Azeem Abul Seoud[3]
Information Technology Department, College of Computer & Information Sciences[1]
Princess Nourah bint Abdulrahman University, Riyadh, 11461 Saudi Arabia[1]
Computer Engineering Department, Misr University for Science and Technology, Giza, 12511 Egypt[1]
Department of Communication &Electronics Engineering, Fayoum University, Egypt[2, 3]

***Abstract*—Correlation between gene expression profiles across multiple samples and the identification of inter-gene interactions is a critical technique for Co-expression networking. Due to the highly intensive processing of calculating the Pearson's Correlation Coefficient, PCC, matrix, it often takes too much processing time to accomplish it. Therefore, in this work, Big Data techniques including MapReduce and Spark have been employed in a cloud environment to calculate the PCC matrix to find the dependencies between genes measured in high throughput microarray. A comparison between the running time of each phase in both of MapReduce and Spark approaches has been held. Both these techniques can dramatically speed up the computation allowing users to work with highly intensive processing. However, Spark has yielded a better performance than the MapReduce as it performs the processing in the main memory of the worker nodes and avoids the unnecessary I/O operations with the disks. Spark has yielded 80 times speed up for calculating the PCC of 22777 genes, however the MapReduce attained barely 8 times speed up.**

***Keywords*—Pearson's correlation; Hadoop; MapReduce; spark; gene co-expression networks; GCN; Affymetrix microarrays**

I. INTRODUCTION

Gene co-expression networks (GCN) [1] are gaining attention nowadays as useful representations of biologically interesting interactions among genes. Finding the interactions with significant genes [2] can help in understanding their molecular pathways. Constructing the similarity matrix between genes in GCN is the most complex part as the complexity rises quadratically. So, the most computationally demanding all pairwise combinations must be analyzed.

The analysis of correlated genes can help in finding other gene functions or relationships. The correlation between genes can be estimated based on their expression values and can be visualized via networks that reveal the interactions between co-expressed genes. Utilizing such gene expression values is currently effortless using the public accessible genomics data banks for RNA-seq, and high throughput microarrays. However, the genomics data is public and available, still the analysis of such data needs powerful platforms and algorithms for its processing. The parallel computing technology plays an essential role in processing and analyzing such huge amount of data. Even though, there are many paradigms and platforms from the parallel computing technology have been intensively reviewed and compared in previous studies[3],[4], [5],[6], there still an open question in utilizing the big data techniques in the processing of the gene expression profiles and finding their relationships.

High throughput technologies such as the Affymetrix microarrays have turned molecular biology into a data-intensive discipline that requires the usage of high-performance computing resources[7]. Flexible framework is needed to cover the resources which are required in highly intensive processing, and to help in data storage and processing. This requires a huge investment in both money and manpower. This problem can be overcome by cloud computing which has emerged as an additional technology offering virtualized environments[8], [9].

High throughput microarrays contain a huge number of genes. Determining the relationships between all these gene experiments proved to be very useful in biological analyses[10]. It has helped in understanding the molecular basis of complex disease traits as well as the prediction of treatment responses of individual subjects. Several methods existed to construct correlation or similarity matrix, i.e., a two-dimensional triangular matrix, where each value is the similarity coefficient of one gene pair. Some examples of those methods are Pearson's[11],Spearman[12],Theil-Sen[13] and Kendall[14] correlations.

Computation between gene expression profiles across multiple samples and the identification of inter-gene interactions is a critical technique for Co-expression networking, which usually relies on all-pairs correlation (or a similar measure). In this respect, Pearson's Correlation Coefficient (PCC) is one of the techniques that have been widely used for gene co-expression network construction. All pairs PCC computation has recently been widely used in Bioinformatics; yet, it is computationally demanding large numbers of gene expression profile. In the present work, it is important to calculate the PCC matrix to find the dependencies between all huge numbers of genes measured in our high throughput microarray. It requires an enormous amount of computation, resulting in slow data processing and takes more days to finish the calculation because it is considered very highly intensive processing. So, new approaches are needed to calculate and accelerate such a complicated process.

There are technologies that show great promise in bioinformatics, such as MapReduce[15], Hadoop[16] and Spark[17] which can contribute to solving the intensive computations. MapReduce and Spark are widely used high performance parallel frameworks that can solve the problem of

the Pearson's Correlation Coefficient matrix[18]. Apache Spark is an open source designed to enhance the computational speed in highly intensive processing. MapReduce works on the file system commonly known as Hadoop Distributed File System (HDFS), whereas Spark works in memory data processing engine. Whenever any operation is performed, Hadoop reads the data from the disk and uses the MapReduce to perform the task. While Spark keeps the data in memory and performs operation at a faster speed than Hadoop. However, the main drawback of MapReduce lies in its relatively high runtime for input datasets consisting of thousands of genes. This prevents the wide adoption of this method by the scientific community especially in intensive processing computation.

The present work focuses on holding a comparison between two Big Data techniques: MapReduce and Spark, which are considered parallel tools that accelerate the construction of intensive processing of pairwise correlation matrix between genes. Multithreading Programming Model [19] in both techniques are employed in this study to achieve efficient performance. The rest of the paper is organized as follows: Section II presents previous works related to the parallelization of the algorithm to calculate the Pearson's Correlation for intensive processing data to find the dependencies between all the huge numbers of genes. Section III describes the parallel implementation in MapReduce and Spark. Section IV provides the experimental evaluation regarding runtime efficiency. Section V presents the results and discussion. Finally, the conclusion is presented in Section VI.

II. Literature Review

This section shows recent work for determining the PCC matrix in Bioinformatics[20],[21] and Non-bioinformatics [22][23]applications. In[20], a parallel approach has been introduced to analyze the correlation between genes. In [24], a parallel implementation of transcription networks via GPUs, Graphical Processing Units, has been developed. In [25] and [26] a Message Passing Interface (MPI) implementation for the parallel construction of similarity matrices on multicore cluster processing have been provided. They have used MI (Mutual Information) instead of Pearson's Correlation for inferring the interactions between genes. Although the MI can detect non-linear connections better than Pearson's Correlation, some experiments have shown that it is not relevant in the case of gene co-expression networks. In [27], a parallel tool for the construction of GCN using GPUs was introduced. In [28], a distributed approach for computing the PCC matrix on Intel Xeon cluster has been proposed. A hybrid approach of MPI and OpenMP to compute the PCC matrix has been provided in [29]. A parallel approach on a multicore cluster using MPI, MPIGeneNet, has been introduced in [30] which uses GSL (GNU Scientific Library) and MKL (Math Kernel Library) libraries to perform mathematical functions. These libraries are in continuous evolution, so MPIGeneNet will benefit from future library updates without requiring any modification in its code. In [31], an algorithm for constructing the GCN using MapReduce in a cloud environment has been developed. In that algorithm, an approach from the information theory, ARACNE [21], has been employed. In [22] a parallel implementation of the Support Vector Machine, SVM, algorithm using the OpenMP for the Multicore platform has been developed. In [32], a parallel approach using GPU to compute the Pcc matrix in a Magnetic Resonance Imaging, MRI, images to estimate the functional interactions in human's brain. Another work for calculate the PCC matrix using a hybrid approach of the MPI, and the Compute Unified Device Architecture, CUDA has been developed in [33]. And in [23], a model for estimating the scalability of the parallel algorithms in the Cluster platform has been presented. Recently, in [34], a Parallel MapReduce (PMR) framework was proposed to compute bioinformatics applications and reduce the computation cost.

Each of the afore-mentioned frameworks provides a different paradigm of parallel programming and has their own strong and weak points. However, the performance of the Big Data techniques in the construction of similarity matrix in GCN still needs more examination. In this research, a comparison is held between the MapReduce, and Spark to compute the PCC matrix in real data of time series microarrays of Hepatocellular Carcinoma, HCC, containing a massive number of genes. The comparison has been done in a cloud environment which is more inexpensive, and flexible than the on-premises computing resources [31]. Cloud computing model has achieved an incredible performance for many applications in bioinformatics [35],[36].

III. Methods

The Pearson correlation coefficient is one of the most popular approaches in measuring the intensity of a linear association between two genes in a Gene Co-expression Network; hence, it has been applied here as a measure of dependencies between interacting genes. Let(X&Y) considered as a pair of gene expression profiles & (n) is the number of pairs in a gene expression data then the PCC can be calculated as shown in (1).

$$PCC = \frac{\sum_{i=1}^{n} XY - \frac{\sum_{i=1}^{n} X \sum_{i=1}^{n} Y}{n}}{\sqrt{(\sum_{i=1}^{n} X^2) - (\frac{(\sum_{i=1}^{n} X)^2}{n})} \sqrt{(\sum_{i=1}^{n} Y^2) - (\frac{(\sum_{i=1}^{n} Y)^2}{n})}} \quad (1)$$

The computation of the Pearson's Correlation Coefficient between genes expressed in a gene expression matrix has a quadratic complexity. Therefore, we are suggesting here a parallel algorithm that will break the entire calculation of PCC matrix into components. Each component represents an independent computation. Fig. 1 and Fig. 2 depict a flowchart for calculating the PCC matrix using the MapReduce and Spark.

Each technique receives an input matrix comprising the gene expression values for each gene in different conditions, samples. These expression data are saved in a numeric matrix, with n columns, the number of genes, and m rows, the number of samples.

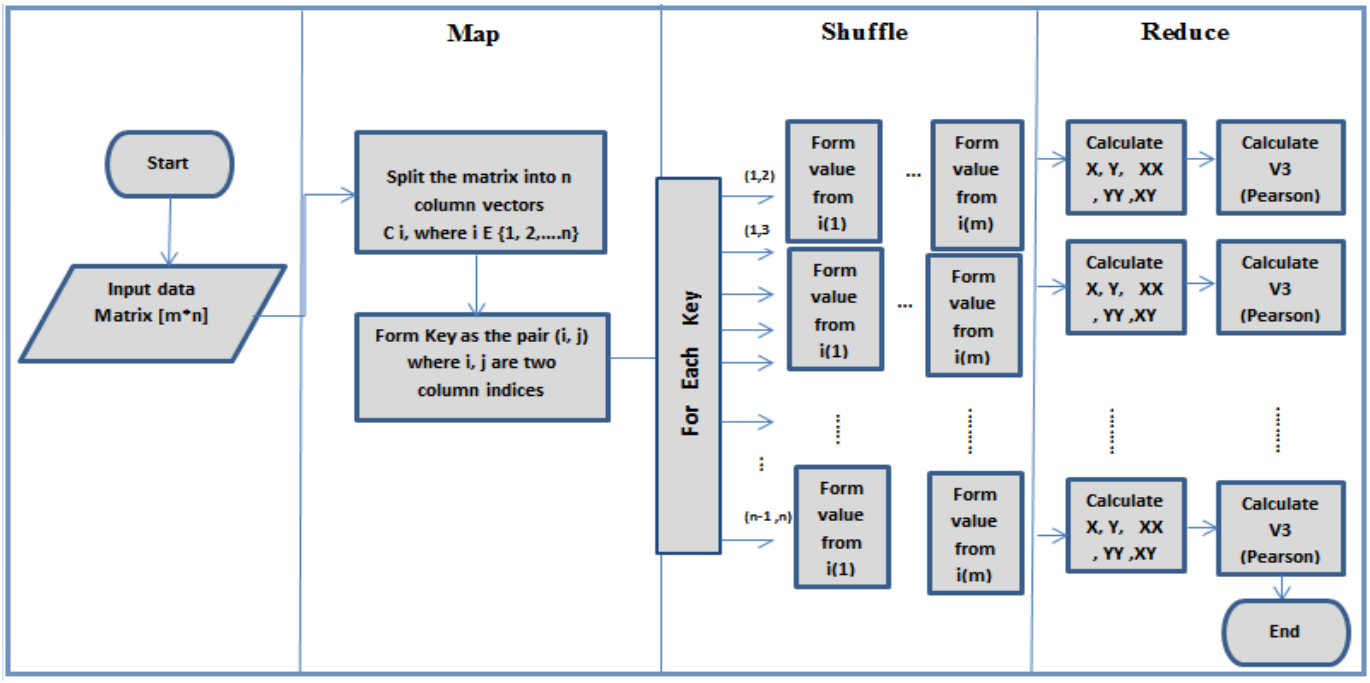


Fig. 1. Flowchart for Calculating the PCC Matrix between Genes Expressed in a Gene Expression Matrix using MapReduce.

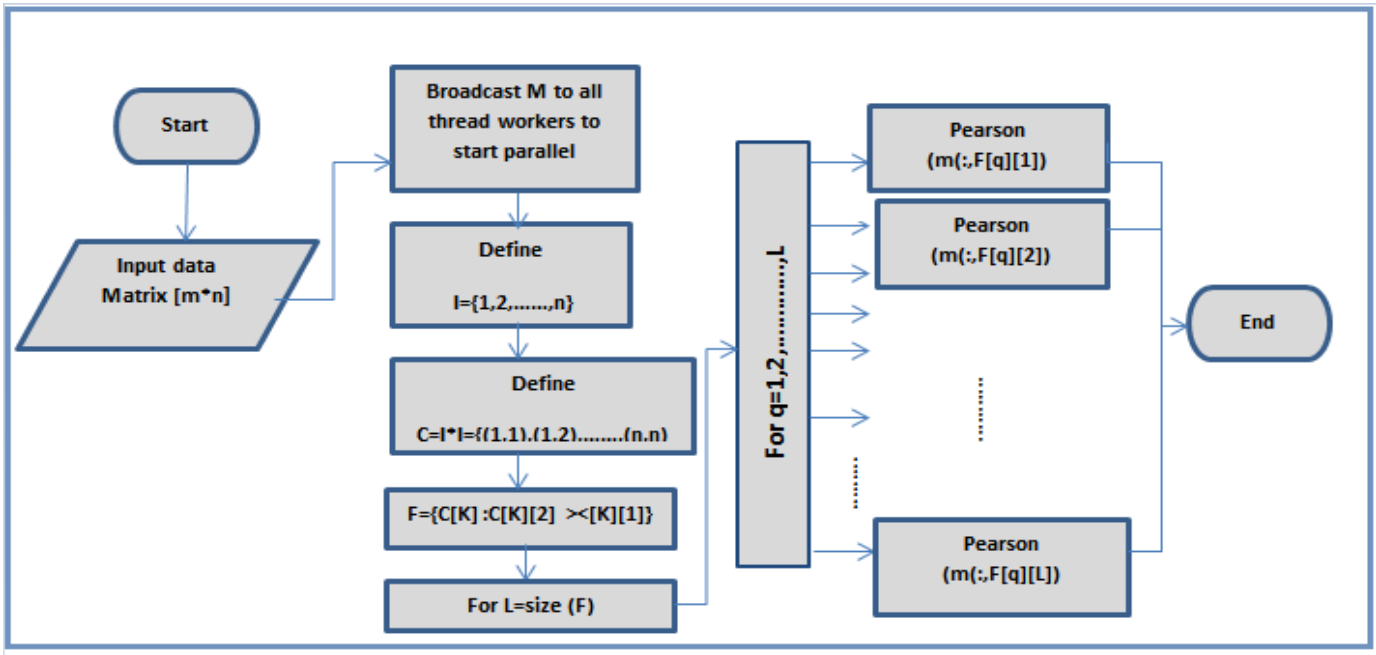


Fig. 2. Flowchart for Calculating the PCC Matrix between Genes Expressed in a Gene Expression Matrix using Spark.

MapReduce consists of mappers which perform a large portion of work and reducers which perform a relatively small amount of computation which achieves the best performance. In the implementation of the Multithreaded Mapper implementation, threads from a thread pool invoke a queue of key value pairs in parallel. Multiple threads running a map task can help to speed up the tasks, based on availability of cores in the system. The map and shuffle task only receive a key-value pair input <k1, V1> and get outputs with other key-value pair in parallel. However, the reduce task receives the input from shuffle<K2, List<V2>>, which is a key and a list of values associated with that key. It gets all this pairs of values associated with the i-th and j-th columns/variables and compute Pearson correlation. The output of the Reduce phase is (K3,V3) as the PCC between a pair of genes, the i-th and j-th genes.

Spark integrates the whole functionality in one program. This makes the tool easier to work with the users having only to launch the application once and avoid writing/reading from disks because it is based on memory which makes it is faster than MapReduce. Spark uses a hybrid approach that combines

MPI processes and threads. Each MPI process launches multiple threads to efficiently exploit the cores available on each node and to reduce the memory requirements. Regarding the workload distribution, as the PCC must be calculated for all gene pairs, the workload of this step can be represented with a 2D matrix, where both axes x and y include all genes. Further improvement in the performance of the proposed algorithm has been done by dividing the data into multiple partitions based on the number of threads and execute on available cores on multiple nodes and also only half of the matrix must be calculated Concretely sum (i, i=1 to N) = N * (N + 1) / 2 which called triangular numbers (Since cor(A,B)=cor(B,A)) which adding more time for saving. So, every pair of genes has been handled as a compute job, key them with a unique index, send to any compute unit available, and put them back into the result matrix using their key.

The detailed algorithm to construct the PCC matrix using the Map Reduce and Spark are shown below.

Algorithm: PCC using MapReduce

Require: Input data matrix A[m,n] where n length of A and split the matrix into n column ci where I {1,2,…..n} . (i,j) are two cloumn indicies. N as the number of values in array V -> N = length of array V.

```
Function Map(Array A)
for each thread
        for i = 0....A.length
                for j = i....A.length
                        Let k = pair(i, j)
                        Let v = pair(A[i], A[j])
                        Send the key (k) and the value pair (v)
Function Reduce(key k, value pairs V)
        Let X = 0
        Let Y = 0
        Let XX = 0
        Let YY = 0
        Let XY = 0
        for i=1 to N
                Let X = X + V[i][1]
                Let Y = Y + V[i][2]
                Let XX = XX + (V[i][1])^2
                Let YY = YY + (V[i][2])^2
                Let XY = XY + V[i][1] * V[i][2]
Let P = (XY - (X * Y / N)) / sqrt((XX - (X^2 / N)) * (YY - (Y^2 / N))
Compute P
end for
```

Algorithm: PCC using Spark

Require: Input data matrix A[m,n] where n length of A ,Number of threads=8, Obtain indices -> I[i], where i=1 to n.

```
Broadcast matrix to all thread workers
for each thread
Let I be an array of size n
for i = 1....I.size
        I[i] = i
Le let C be an array of pairs, where each element of C is holds an element of the cac cartesian product of the elments of I -> C[i], where i =1 to n^2
Let F be an empty array
Let j = 0
for i = 1....C.size
        if C[i][1] is more than or equal C[i][2]
                F[j] = C[i]
                Let j = j + 1
Let P be an array of C.size
For i = 1....F.size -> where (i=1 to n*(n + 1)/2)
        Compute the Pearson correlation M[:, P[i][1]] and M[:, P[i][2]]
end for
```

IV. Experimental Setup

A. Material

The data employed in this research is a non-benchmarking dataset for Liver cancer, Hepatocellular Carcinoma (HCC). It is a real data downloaded from GEO, Gene Expression Omnibus data bank [37] and contains thirty-five Microarray samples of HCC that have been downloaded from. These samples have been collected using the Affymetrix HG-U133A 2.0 platform. HCC is a complication of HCV (Hepatitis C virus) cirrhosis. The raw data has been preprocessed by the Affy package which is provided by the Bioconductor [38].

B. Platform

A cloud platform from IBM, IBM Analytics Engine, IAE[39] has been utilized in this research. The IAE offers a parallel infrastructure for MapReduce and Spark on the IBM cloud. It permits users to upload their data in a layer called the IBM Cloud Object Storage and provides clusters of computing nodes to work on the uploaded data. The separation of the computing and storage layers helps in having more scalability and flexibility in analyzing. Analytics libraries and open-source packages has been employed. More details about the hardware and software employed in each technique are listed in Table I.

TABLE I. CHARACTERISTICS OF THE CLOUD PLATFORM USED IN THE EXPERIMENTAL EVALUATION

Specification	Hadoop	Spark
Software package	AE 1.2 (Analytic Engine IBM)	
Version	Hadoop 3.1.1	Apache Spark 2.3.2
Nodes	2	
Cores	32	
Memory	128GB	
Driver Memory	50GB	
Executor Memory	20GB	
Spark driver max Result size	Should be more than or equal 25GB	
Thread	8	

V. RESULTS AND DISCUSSION

The performance of the employed techniques has been evaluated by calculating the whole running time on different chunks of data including 400, 1000, 4000, 8000, 12000, 16000, and 22777genes. Multithreading for each technique splits tasks into threads to execute them at the same time in parallel.The retrieved running time using the proposed MapReduce, and Spark algorithms is shown in Fig. 3, and Fig. 4 correspondingly. As illustrated in Fig. 3, the running time for finding the correlation between 4000 genes was an hour and twenty minutes and it increases exponentially until it reaches four hours and fifty minutes for the whole number of genes, 22777. The exponential increase in the running time reveals a high time complexity in using the MapReduce approach. That complexity is due to storing the whole dataset to HDFS after running each job. In addition, the dataset is replicated three times in HDFS by default. Thus, the time complexity of Map task in intrinsic operations (sorting, shuffling, sending data, etc.) is O(n^2) operations per line and O(n) in the Reduce task.

As noticed from Fig. 3, a long processing time has been retrieved using the MapReduce. Such long processing time can be clarified as follows:

- The size of the whole dataset employed in this work is small (3.7 MB) which is less than the minimum block size 64 MB of HDFS. This means that if a block size of HDFS is 64 MB with 3 replication and we have 1 MB file, we do not lose (63 MB * 3) 189 MB. Since physically just three 1 MB files are stored with the standard block size of the underlying file systems and Hadoop will typically try to spawn a mapper per block. So, if we have 40 blocks with10 KB files, then Hadoop may end up spawning 40 mappers eventually per block even if the size of file is small.
- The number of mappers and reducer tasks are directly proportional to the input splits, which depend on DFS block size (no. split= no. Map =input size/block size) which helps in working all tasks in parallel well. In our present work, we had one Map task, and one reduce. So, using HDFS in intensive parallel processing with small dataset was not helpful in MapReduce. This is a real wastage overhead time when using MapReduce.

On the other side, as illustrated in Fig. 4, the running time for the Spark algorithm is less than the time consumed using the MapReduce algorithm. The consumed time for retrieving the correlation between 4000 genes was just five minutes and fifty-four seconds and it increases exponentially until it reaches twenty-nine minutes for the whole number of genes, 22777.

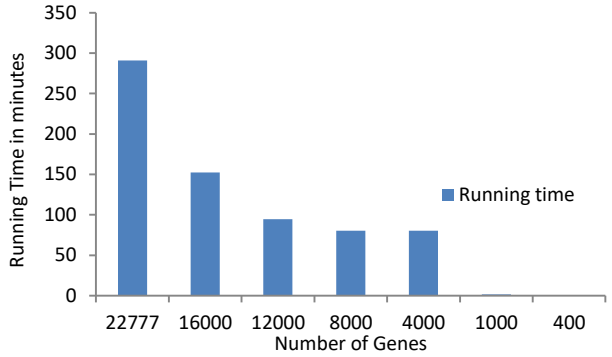


Fig. 3. The Running Time of MapReduce Algorithm on different Chunks of Data on IBM Analytic Engine.

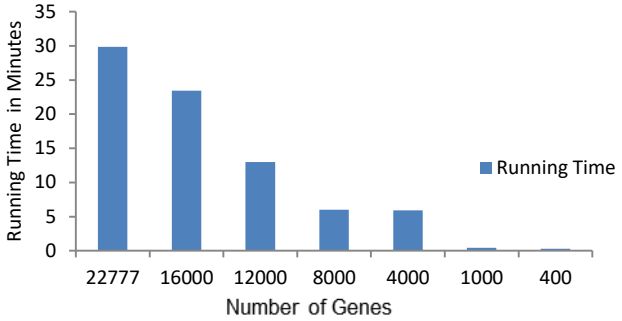


Fig. 4. The Running Time of Spark Algorithm on different Chunks of Data on IBM Analytic Engine.

In Spark, Resilient Distributed Datasets (RDD) is employed to manage the data through partitions. RDD helps to distribute the data processing with negligible network traffic for sending data between executors in parallel. The best way to decide the number of partitions into smaller chunks of data is to make the number of partitions equal to the number of cores in the cluster; so that all the partitions will process in parallel, and the resources will be utilized in an optimal way.

We used 32 partitions like the number of Virtual cores and 8 threads in the cluster to achieve the best performance as we can. Compared to MapReduce, Spark does not work the entire dataset in the HDFS until completing each task as done in MapReduce.

To sum up the comparison between the MapReduce and Spark in calculating the correlation coefficient between genes in a gene expression matrix, a bar chart for the retrieved running times using them is depicted in Fig. 5.

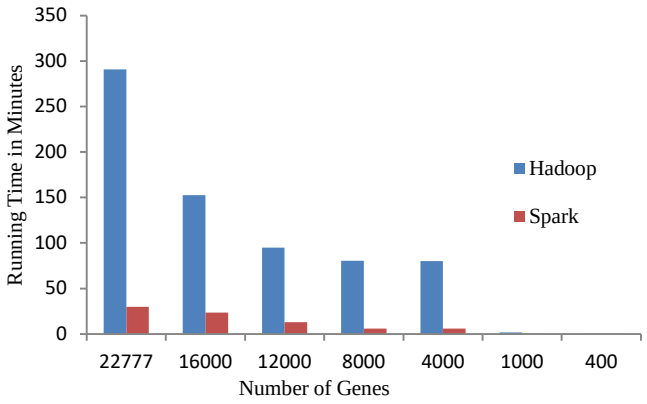


Fig. 5. A Comparison between the Retrieved Running Time in Calculating the PCC Matrix on different Chunks of Genes in using MapReduce and Spark.

A. *Comparing the Performance*

In this section, the performance between MapReduce and Spark algorithms is compared to a single CPU processing of the PCC. We have calculated the speedup for different data sizes as listed in Table II. In Table II, N, and M represents the number of genes and number of samples. It can be noticed that the maximum speed up of Spark, 304.1x, is considerably greater than the corresponding one on the MapReduce, 83.12x, for a data size of 1000 genes represented in 15 samples.

In addition, we have performed more investigation to the performance of the employed techniques, MapReduce, and Spark, by analyzing the total running time. The running time in the MapReduce technique is composed of three components including the Mapping, Shuffle & Merge, and Reduce. However, in the Spark, it is composed only of the Mapping, and Reduce phases. As depicted in Table III, long time is consumed in the Mapper phase compared to the Reduce phase.

Long time for the phases of the MapReduce technique can be clarified as follows. Reading the data from the disk and then running the mappers has consumed too much time, four hours. The Generation of a lot of keys has taken a lot of time to sort them. Storing the output of mappers back on disk also has consumed around 6 minutes. Then the reading/storing of the data from the disk in Shuffling phase has contributed to the increased processing time, 3 minutes. Finally, the Reduce phase has spent 36 minutes.

Six times the disk is accessed to complete one job in the MapReduce which slowdown its processing. Thus, MapReduce has a big drawback since it must operate with the entire set of data in the HDFS on the completion of each task, which in turn increases the time and the cost of processing data, so we found that Spark is faster than MapReduce with 10.2613%.

Although, Spark has two phases including transformation (Map) and reduce tasks as in the MapReduce approach. Spark has spent only 23 minutes and the other is for action (reduce) which spent only 2 minutes. An interpretation to the retrieved performance of Spark can be given as follows:

- In Spark, a Directed a cyclic graph (DAG)[40] is created when it starts executing a job.
- Looking at the DAG, remembering the steps in the DAG. The Spark's DAGs enable optimizations between steps compared to Hadoop which does not have any cyclical connection between MapReduce steps. That is, at that level, no performance tuning can occur which proven that Spark is much faster for applications.
- Having eight steps for transformations but does not really go to disk to perform the transformations.

TABLE II. COMPARING THE PERFORMANCE BETWEEN MAPREDUCE & SPARK.

Matrix Size (N*M)	**Serial PCC Time(s) / Speedup**	**MapReduce Time(s) / Speedup**	**Spark Time(s) / Speedup**
400*10	826.05/ 1x	25.98/ 31.79 x	17.652/ 46.77x
1000*15	7,773.29/ 1x	93.516/ 83.12 x	25.56198/ 304.1x
4000*20	22034.14/ 1x	4813.2/ 4.58 x	355.98/ 61.88x
8000*25	66402.9/ 1x	4830/ 13.75x	360/ 184.45x
22277*35	143,382.74/ 1x	17443.98/ 8.22x	1789.98/ 80.1x

TABLE III. A COMPARISON BETWEEN THE RUNNING TIME OF EACH PHASE IN BOTH OF MAPREDUCE AND SPARK APPROACHES IN CALCULATING THE CORRELATION MATRIX BETWEEN GENES IN A HIGH THROUGHPUT MICROARRAY

Technique	**Phase**	**Phase Time**
MapReduce	Mapping phase	4 hours + 6 minutes+53 seconds
	Shuffle & merge phase	3 minutes + 48 seconds
	Reduce Phase	36 minutes+ 46 seconds
Spark	Mapping phase	23 minutes
	Reduce phase	2 minutes

At this point, a Spark job goes to disk, performs the first transformation, keeps the result of transformation in memory, performs the second transformation, keeps the result in memory and so on until all the steps are completed; so only two accesses to disk to write the output of the job which makes Spark faster than MapReduce to access the disk. The higher efficiency on Spark can be explained as follows. In Spark, each MPI process launches multiple threads to efficiently exploit the available cores on each node and to reduce the memory requirement. The launched threads have a shared memory space on the cluster cores and allowing parallel execution with up to 32 working processes. This approach can be implemented easily with Spark which makes it faster than MapReduce.

VI. Conclusion

In the present paper, we have introduced a parallel implementation for an algorithm for computing the PCC matrix in GCN using Spark and MapReduce approaches. Spark has yielded a better performance than the MapReduce as it performs the processing in the main memory of the worker nodes and prevents the unnecessary I/O operations with the disks. So, the memory in the Spark cluster should be at least as large as the amount of data needed to process. However, Spark is more expensive when setting up the cluster because it requires more RAM compared to MapReduce. MapReduce is highly fault-tolerant because it was designed to replicate data across many nodes but in Spark, data is replicated across executor nodes, and can generally be corrupted if the node or phase between executors and drivers fails.

As a future work for this study, we are recommending the implementation of PCC matrix using GPUs and comparing its performance to the results retrieved in this research.

Acknowledgment and Funding

This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University through the Fast-track Research Funding Program.

References

[1] N. M. A. Samee, N. H. Solouma, and Y. M. Kadah, “K4. Gene network construction and pathways analysis for high throughput microarrays,” Natl. Radio Sci. Conf. NRSC, Proc., no. April, pp. 649–658, 2012, doi: 10.1109/NRSC.2012.6208578.

[2] N. M. Abdel Samee, N. H. Solouma, and Y. M. Kadah, “Detection of biomarkers for Hepatocellular Carcinoma using a hybrid univariate gene selection methods,” Theor. Biol. Med. Model., vol. 9, no. 1, p. 1, 2012, doi: 10.1186/1742-4682-9-34.

[3] A. O, K. AR, and A. A, “Parallel Algorithms for Inferring Gene Regulatory Networks: A Review,” Curr. Genomics, vol. 19, no. 7, pp. 603–614, Jun. 2018, doi: 10.2174/1389202919666180601081718.

[4] H. Shi, B. Schmidt, W. Liu, and W. Müller-Wittig, “Parallel mutual information estimation for inferring gene regulatory networks on GPUs,” BMC Res. Notes, vol. 4, p. 189, 2011, doi: 10.1186/1756-0500-4-189.

[5] F. F. Borelli, R. Y. de Camargo, D. C. Martins, Jr, and L. C. Rozante, “Gene regulatory networks inference using a multi-GPU exhaustive search algorithm,” BMC Bioinformatics, vol. 14, no. Suppl 18, p. S5, 2013, doi: 10.1186/1471-2105-14-S18-S5.

[6] W.-P. Lee, Y.-T. Hsiao, and W.-C. Hwang, “Designing a parallel evolutionary algorithm for inferring gene networks on thecloud computing environment,” BMC Syst. Biol., vol. 8, no. 1, p. 5, Jan. 2014, doi: 10.1186/1752-0509-8-5.

[7] O. Spjuth et al., “Experiences with workflows for automating data-intensive bioinformatics,” Biol. Direct, vol. 10, no. 1, pp. 1–12, 2015, doi: 10.1186/s13062-015-0071-8.

[8] M. C. Schatz, B. Langmead, and S. L. Salzberg, “Cloud computing and the DNA data race,” Nat. Biotechnol., vol. 28, no. 7, pp. 691–693, 2010, doi: 10.1038/nbt0710-691.

[9] L. D. Stein, “The case for cloud computing in genome informatics,” Genome Biol., vol. 11, no. 5, 2010, doi: 10.1186/gb-2010-11-5-207.

[10] P. Minguez and J. Dopazo, “Assessing the biological significance of gene expression signatures and co-expression modules by studying their network properties,” PLoS One, vol. 6, no. 3, 2011, doi: 10.1371/journal.pone.0017474.

[11] A. Gobbi and G. Jurman, “A null model for pearson coexpression networks,” PLoS One, vol. 10, no. 6, 2015, doi: 10.1371/journal.pone.0128115.

[12] J. Nie et al., “TF-Cluster: A pipeline for identifying functionally coordinated transcription factors via network decomposition of the shared coexpression connectivity matrix (SCCM),” BMC Syst. Biol., vol. 5, 2011, doi: 10.1186/1752-0509-5-53.

[13] H. Peng, S. Wang, and X. Wang, “Consistency and asymptotic distribution of the Theil-Sen estimator,” J. Stat. Plan. Inference, vol. 138, no. 6, pp. 1836–1850, 2008, doi: 10.1016/j.jspi.2007.06.036.

[14] F. Gómez-Vela, C. D. Barranco, and N. Díaz-Díaz, “Incorporating biological knowledge for construction of fuzzy networks of gene associations,” Appl. Soft Comput. J., vol. 42, pp. 144–155, 2016, doi: 10.1016/j.asoc.2016.01.014.

[15] J. Dean and S. Ghemawat, “MapReduce: Simplified data processing on large clusters,” Commun. ACM, vol. 51, no. 1, pp. 107–113, 2008, doi: 10.1145/1327452.1327492.

[16] Tom White, “Hadoop: The Definitive Guide [Book],” O’Reilly Media, Inc. https://www.oreilly.com/library/view/hadoop-the-definitive/9780596521974/ (accessed Jul. 20, 2020).

[17] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, “Spark: Cluster computing with working sets,” 2nd USENIX Work. Hot Top. Cloud Comput. HotCloud 2010, 2010.

[18] S. Wang et al., “Optimising parallel R correlation matrix calculations on gene expression data using MapReduce,” BMC Bioinformatics, vol. 15, no. 1, pp. 1–9, 2014, doi: 10.1186/s12859-014-0351-9.

[19] K. Kavi, “Multithreading Implementations The University of Texas at Arlington,” no. September 1998, 2013.

[20] M. M. Zhu and Q. Wu, “Transcription network construction for large-scale microarray datasets using a high-performance computing approach,” BMC Genomics, vol. 9, no. SUPPL. 1, pp. 1–10, 2008, doi: 10.1186/1471-2164-9-S1-S5.

[21] P. Zoppoli, S. Morganella, and M. Ceccarelli, “An information theoretic approach to reverse engineering of regulatory gene networks from time-course data,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6160 LNBI, pp. 97–111, 2010, doi: 10.1007/978-3-642-14571-1_8.

[22] H. Zhu, P. Li, P. Zhang, and Z. Luo, “A high performance parallel ranking SVM with OpenCL on multicore and many-core platforms,” Int. J. Grid High Perform. Comput., vol. 11, no. 1, pp. 17–28, 2019, doi: 10.4018/IJGHPC.2019010102.

[23] L. B. Sokolinsky, “BSF: A parallel computation model for scalability estimation of iterative numerical algorithms on cluster computing systems,” J. Parallel Distrib. Comput., vol. 149, no. May, pp. 193–206, 2021, doi: 10.1016/j.jpdc.2020.12.009.

[24] J. Ingram and M. Zhu, “GPU accelerated microarray data analysis using random matrix theory,” Proc.- 2011 IEEE Int. Conf. HPCC 2011 - 2011 IEEE Int. Work. FTDCS 2011 -Workshops 2011 Int. Conf. UIC 2011-Work. 2011 Int. Conf. ATC 2011, pp. 839–844, 2011, doi: 10.1109/HPCC.2011.119.

[25] J. Zola, M. Aluru, A. Sarje, and S. Aluru, “Parallel information-theory-based construction of genome-wide gene regulatory networks,” IEEE Trans. Parallel Distrib. Syst., vol. 21, no. 12, pp. 1721–1733, 2010, doi: 10.1109/TPDS.2010.59.

[26] S. Misra, K. Pamnany, and S. Aluru, “of Genome-Scale Networks on the Intel Ò Xeon Phi TM Coprocessor,” vol. 12, no. 5, pp. 1008–1020, 2015.

[27] M. Liang, F. Zhang, G. Jin, and J. Zhu, "FastGCN: A GPU accelerated tool for fast gene co-expression networks," PLoS One, vol. 10, no. 1, pp. 1–11, 2015, doi: 10.1371/journal.pone.0116776.

[28] Y. Liu, T. Pan, and S. Aluru, "Parallel Pairwise Correlation Computation on Intel Xeon Phi Clusters," Proc. - Symp. Comput. Archit. High Perform. Comput., pp. 141–149, 2016, doi: 10.1109/SBAC-PAD.2016.26.

[29] J. González-Domínguez and M. J. Martín, "Fast Parallel Construction of Correlation Similarity Matrices for Gene Co-Expression Networks on Multicore Clusters," Procedia Comput. Sci., vol. 108, pp. 485–494, 2017, doi: 10.1016/j.procs.2017.05.023.

[30] J. Gonzalez-Dominguez and M. J. Martin, "MPIGeneNet: Parallel calculation of gene co-expression networks on multicore clusters," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 15, no. 5, pp. 1732–1737, 2018, doi: 10.1109/TCBB.2017.2761340.

[31] Y. Abduallah, T. Turki, K. Byron, Z. Du, M. Cervantes-Cervantes, and J. T. L. Wang, "MapReduce Algorithms for Inferring Gene Regulatory Networks from Time-Series Microarray Data Using an Information-Theoretic Approach," Biomed Res. Int., vol. 2017, 2017, doi: 10.1155/2017/6261802.

[32] T. Eslami and F. Saeed, "Fast-GPU-PCC: A GPU-based technique to compute pairwise pearson's correlation coefficients for time series data—fMRI study," High-Throughput, vol. 7, no. 2, 2018, doi: 10.3390/ht7020011.

[33] E. Kijsipongse, S. U-Ruekolan, C. Ngamphiw, and S. Tongsima, "Efficient large Pearson correlation matrix computing using hybrid MPI/CUDA," Proc. 2011 8th Int. Jt. Conf. Comput. Sci. Softw. Eng. JCSSE 2011, no. May, pp. 237–241, 2011, doi: 10.1109/JCSSE.2011.5930127.

[34] A. A. Al-Absi, N. A. Al-Sammarraie, W. M. S. Yafooz, and D. K. Kang, "Parallel MapReduce: Maximizing cloud resource utilization and performance improvement using parallel execution strategies," Biomed Res. Int., vol. 2018, 2018, doi: 10.1155/2018/7501042.

[35] T. Nguyen, W. Shi, and D. Ruden, "CloudAligner: A fast and full-featured MapReduce based tool for sequence mapping," BMC Res. Notes, vol. 4, 2011, doi: 10.1186/1756-0500-4-171.

[36] S. Zhao et al., "Rainbow: A tool for large-scale whole-genome sequencing data analysis using cloud computing," BMC Genomics, vol. 14, no. 1, 2013, doi: 10.1186/1471-2164-14-425.

[37] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," Nucleic Acids Res., vol. 30, no. 1, pp. 207–210, 2002, doi: 10.1093/nar/30.1.207.

[38] R. C. Gentleman et al., "Bioconductor: open software development for computational biology and bioinformatics.," Genome Biol., vol. 5, no. 10, 2004, doi: 10.1186/gb-2004-5-10-r80.

[39] A. G. Bromley, "Babbage's Analytical Engine Plans 28 and 28a - The Programmer's Interface," IEEE Ann. Hist. Comput., vol. 22, no. 4, pp. 5–19, Oct. 2000, doi: 10.1109/85.887986.

[40] H. Wang, S. Li, X. Li, and H. Zhong, Microstructure and thermoelectric properties of doped p-type CoSb3 under TGZM effect, vol. 466. 2017.

Analysis of Different Attacks on Software Defined Network and Approaches to Mitigate using Intelligent Techniques

P. Karthika, Dr. A. Karmel
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai Campus, Chennai, India

***Abstract*—The detection of DDoS (Distributed Denial of Service) attacks is essential topic under network security. DDoS attacks cause network services to become unavailable by repeatedly flooding servers with unwanted traffic. The volume, magnitude, and complexity of these attacks increased dramatically as a result of low-cost Internet connections and easily available attack tools. Both Software Defined Networking (SDN) and Deep Learning (DL) have recently found a number of practical and fascinating applications in industry and academia. SDN enables centralized management, a global view of the overall network, and configurable control planes, allowing network devices to adapt to diverse applications. When applied to diverse categorization problems, DL-based approaches outperformed classic machine learning techniques, while SDN characteristics offer better network monitoring and security of the managed network when compared to traditional networks. By inheriting the non-linearity of neural networks, they increase feature extraction and reduction from a high-dimensional dataset in an unsupervised way. An overview of deep learning algorithms for sensing distributed denial of service attacks in software-defined networks with Deep learning is presented within this article. Furthermore, SDN environment is simulated in Mininet using RYU controller. In addition, each paper's mitigation method is examined in the survey.**

***Keywords*—Distributed Denial of Service (DDoS); Software Defined Networking (SDN); attack detection; Mininet; OpenFlow; mitigation; machine learning; deep learning**

I. Introduction

As a result of consecutive evolution of network infrastructure, unending extension of network professional requirements, the massive development of Internet economy in the Internet environment, network facilities containing critical business and industry information have permeated modern society's production and life. The introduction of DDoS assaults can result in irregularities in associated network services, resulting in significant economic losses and even disastrous effects. DDoS assaults are a severe danger to the Internet's network security. The accurate and rapid detection of DDoS assaults is a critical study area in the security industry. The network and control planes are separated in SDN, which is a novel network design. [1-2] enabling network programmability, centralized administration control and interface opening.

Controllers operate solely as packet forwarders in a new networking paradigm, isolating control logic from forwarding and switching aspects. The data plane is made up of network components such as switches that are controlled by the controller in the control plane (also known as Open Flow or simply referred to as OF switches). In large-scale and high-performance computer systems, decoupling the routing plane and forwarding plane is crucial for gaining higher performance. Additionally, it simplifies network management by centralizing configuration and management within the controller. This technique enables for more frequent modifications because the administrator does not have to configure and reconfigure all of the network devices to execute network updates and adjustments. They can utilize the controller to quickly and effectively implement policy and network configuration needs.

To manage data plane, the controller requires numerous core services. It enables the exchange of data with application layer services that perform network functions such as routing, load balancing and intrusion detection. The application layer's services the applications are mapped to entire network by an operating system of network installed on the controller and provides a high level of optimization, automation and network control. Java APIs for local communication and representational state transfer (REST) APIs for remote communication are used by the applications to interface with the controller.

However, a very factor that propels SDN networks to prominence and popularity too exposes them to slew of novel security threats. The distributed denial-of-service (DDoS) attack is a unique of these consumes the utmost devastating outcome on an SDN network. If the network is not adequately protected, DDoS attacks can overwhelm the controller. To defend the SDN network against DDoS attacks, there is a variety of documentation available. In networks, Intrusion Detection Systems (IDSs) sniff packets and alert the administrator if a Distributed Denial of Service (DDoS) assault is identified. One strategy that is attracting the attention of researchers is the use of machine learning to detect distributed denial of service assaults. Defending SDN against threats is continuing research area.

A. Motivation

In past 5 years, the DDoS attacks have strained more attention towards the cyberspace. In large networks, Intrusion Detection Systems (IDS) are widely used to safeguard the network from threats. However, IDS are not a practical option for real-time monitoring, leaving systems open to various

attacks. Attackers continue to develop new processes and strategies for deceiving protection systems, allowing them to illegally use accessible software and harm service providers. Several ways of dealing with DDoS attacks have been proposed in previous research. Various ML/DL techniques have been proposed in earlier studies to fight against DDoS attacks. The goal of this research is to aid the research field in developing and inventing new DDoS attack remedies.

The following are the main contributions of this survey work:

- In the context of SDN, an overview of several types of DDoS attacks are provided.
- Mininet was used to emulate the SDN environment.
- Based on machine learning and deep learning approaches, an in-depth assessment of the most important DDoS detection and mitigation solutions are provided.
- The research issues in SDN deployment and security that need to be investigated are highlighted.

The rest of this paper is structured as follows. Section II details a related work that includes an overview of DDoS attack types, mitigation approaches, and the creation of SDN in Mininet. Section III discusses the need for artificial intelligence in SDN and the various methodologies arrived at using Deep Learning discusses in section IV. The research issues in the deployment and security of SDN are outlined in section V, and the discussion is presented in section VI.

II. Related Work

A. Overview of DDoS Attack

This kind of attack results in the inability of legitimate users to access services and is thus denoted as DoS (Denial of Service) attacks [3]. Consider the following attack situation: A hacker can send several service inquiries to the enterprise to register with organization or obtain connection to some enterprises legitimate service instances. The organizational server will get overwhelmed with service requirements and cannot deliver services to other right customers/users. Another possible assault scenario is one in which numerous machines are used to perform a denial-of-service attack:

Organization's or enterprise's network connects a significant number of machines. Suppose an attacker obtains access to individual or more extra computers belonging to an organization or enterprise. This can abuse the opportunity plus perform DoS attacks against further systems in similar network subnet. This attack surface is extensive in this case; an attacker can take over many machines (Zombies) as well utilize them to execute DoS. Aforementioned type of DoS assaults sometimes referred to as a Distributed Denial of Service attack (DDoS). Fig.1, classifies DDoS attacks.

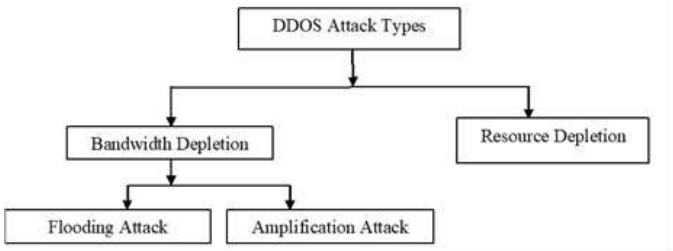


Fig. 1. DDoS Attack Classification.

In addition to Bandwidth deficiency and resource deficiency attacks, around two more classes of DoS attacks are available: Bandwidth Depletion plus Resource Depletion. Bandwidth Depletion is an attack that attempts into overwhelm network with network packets. Bandwidth Depletion attacks are classified as follows: Attackers who use flooding or amplification.

- Flooding attacks seek to overwhelm the network's resources by sending an excess quantity of ICMP or UDP packets.
- Amplification attacks attempt towards the advantage of the IP address broadcast features found on majority of routers. Aforementioned aspect enables a directing system to provide a broadcast internet protocol address instead of a specific address as the destination address. Smurf and Fragile assaults are examples of such attacks [4]. In Resource Depletion assaults, the attacker suffocates the target system's resources. This attack perhaps conducted by attacking a network protocol (for example, Neptune, mail bomb) or generating malformed packets (for example ping of death, Apche2, teardrop Back, land, etc.) and sending them over the network to the victim machine. A concise description of several of these attacks [5] is provided in Table. I.

1) DDoS attack detection: The primary approaches for detecting DDoS attacks are classified as detection of attack established on traffic features as well detection of attack created on traffic abnormality. The first collects numerous attack characteristics and produces a database of DDoS assault characteristics. We can determine whether DDoS, attacks a network by relating and examining the data statistics included in current network data packet as well nature of database. Expert systems, model reasoning, features matching and state transition are primary implementation methods. The latter is generally used to construct a traffic model including analyse aberrant flow variations to assess whether or not the traffic is abnormal and determine whether or not the server has been attacked. Fig. 2, depicts a flowchart of identifying DDoS assault in different stages.

TABLE I. DDOS – ATTACKS, DESCRIPTIONS AND FEATURES

DDoS Attack	Description	Attack features
Land attack	The attacker transmits a manipulated SYN packets with the similar source and destination address. It is helpful in several TCP/IP implementation	Consider the feature 'Land' to detect the attack. If 'Land' is 1, the source address and destination address are alike. Thus, trait is critical in detecting assault.
Smurf attack	Smurf attack is denial-of-service amplification attack whereby an attacker transmits many ICMP echo packets through fake address of the victim's computer to broadcasting internet protocol address. Each host on the broadcast network answers when the packet is received that the victim's system uselessly uses their resources.	This attack might be identified on the victim system by looking at an enormous amount of victim machine ICMP echo responses without transmitting packets from the victim machine to an ICMP echo request.
Teardrop attack	The attacker attempts to transmit the fragmented packets to the intended recipient. Attackers adjust the fragment offset such that the following packets overlap. If the receiving target operating system's IP fragmentation reassembly code contains a fault, the computer will crash owing to inappropriate processing of the overlapping packets.	The feature 'Wrong Fragment', is sum of the connection's faulty checksum packets, provides some insight into the erroneous IP packets. As a result, this attribute is critical in identifying the attack.
Ping of Death attack	An attacker sends an IP packet more significant than the 65,536-byte limit to elicit a "ping of death" denial of service (DOS). The maximum permissible IP packet size is 65.535-bytes, comprising 20-bytes long packet header. This crashes or freezes the machine.	By recording the scope of every ICMP packet and identifying which are larger than 65,535-bytes, and tried Ping of Death can be found.
Mail bomb attack	Mail bomb attacks occur when unauthorized users send a massive sum of e-mail messages through considerable additions to specific mail server, clogging up disc-space and denying other users email capabilities.	This type of attack can be spotted by the presence of thousands of e-mail information from a single person in a short-period.
SYN flood attack	TCP/IP implementation is used in SYN flood. The SYN request is sent to the victim system by an attacker. The victim responds with an ACK and waits for a response. Each half-open connection's information is added to the pending connection queue by the server. The victim server system's half-open links will soon plug the queue, and the system will turn out to be inadequate towards acquire further connections.	A flood assault via SYN may be separated from regular network traffic while searching for many simultaneous SYN packs intended for a specific machine that comes from unattainable host.

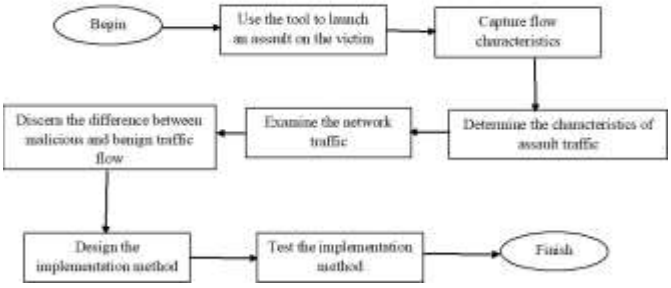


Fig. 2. Stages of Identifying DDoS Assault.

B. Software Defined Network

Deep packet analysis is possible via a complete network view in the revolutionary architecture environment of SDN [6]. It allows for quick response and changes to traffic policies and procedures. The SDN allows perceptual regulators of global visualization illustration to be flexible and timed. Quick deployment that is schedule-aware and intelligent scheduling that is service-aware.

Though assuring network facilities plus lowering implementation value, the software defined network improves user experience and enables more comprehensive network rollout promotion. Fig. 3, shows software defined network architecture. It is visibly clear that the architecture is divided into Applications, Controller and Data plane, which enables us to identify and mitigate attacks in SDN.

Lin and Wang [7] offered DDoS assault detection and defence technique based on SDN. Still, system required three Open flow management tools to accomplish anomaly detection using Flow standard, making implementation and operation complicated.

Yang et al. [8] described a strategy for combining flow statistics and IP entropy-specific information. Using a single flow as well as internet protocol entropy characteristic information, the flow and IP entropy distinctive information are detected, resulting in a more effective and precise detection impact. While information entropy is adaptable and appropriate, it must be used with other technologies to determine the threshold and multi-element weight distribution.

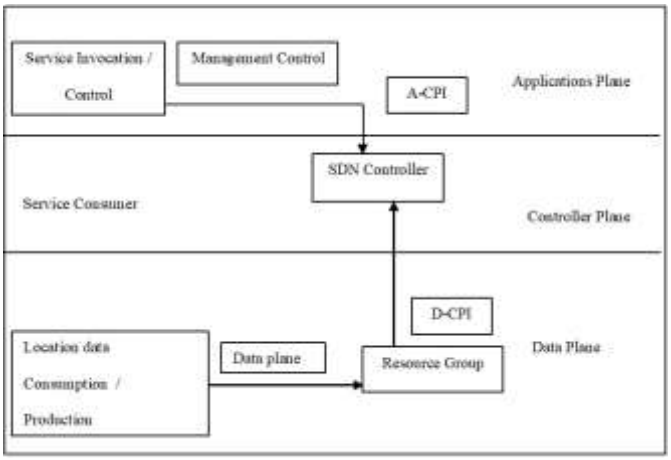


Fig. 3. Software Defined Network Architecture.

Author [9] suggested that to detect DDoS attacks, the approach must analyse the features of each ICMP/TCP/UDP protocol using the training ANN algorithm, which is difficult and ineffective.

In [10], the author presented a strategy for identifying and preventing DDoS assaults in a large network, however it is not suitable for simple implementation. [11] offers a logical source and destination IP address database-based DDoS attack detection system. When a DDoS attack occurs, it investigates the unusual properties of the source and destination IP addresses. It successfully verifies the DDoS attack using the non-parametric cumulative algorithm CUSUM, but the approach needs to change and set the threshold.

Data entropy and the usage of the data-mining method, in which the SOM methodology is most prominent, have been found to be the most important factors in DDoS detection in SDN networks. The SOM algorithm requires determining the number of neurons in advance because of the high false-positive information entropy rate.

1) Mininet and openflo: Mininet is a virtual network device emulator that simulates virtual network devices such as hosts, switches, controllers, and links. Mininet switches offer OpenFlow for highly flexible custom routing and Software-Defined Networking, and its hosts run conventional Linux network software. Mininet makes it easier to conduct research, development, learning, prototyping, testing, and debugging on a laptop or other PC.

Mininet :

- Low-cost and easy-to-use testbed for developing OpenFlow applications.
- Rapid software-defined network prototyping.
- Without the requirement to set up a physical network, complex topology testing may be performed.
- The same topology can be worked on by multiple developers at the same time.

OpenFlow :

- The interface between the OpenFlow controller and the OpenFlow switches is defined by the OpenFlow protocol.
- The OpenFlow protocol assists the OpenFlow controller in instructing the OpenFlow switches how to handle incoming packets.
- Using multiple packet header data, identify and classify packets from an ingress port.
- The packets are dropped or pushed to a specific egress port or to the OpenFlow Controller.

2) Creating SDN in mininet: First, use the following command to construct a topology with a single switch and five separate hosts.

sudo mn --topo single,5 --mac --controller remote --switch ovsk

We need to execute as a sudo instance since we need to access the kernel protocol stack as root. Fig.4 depicts the creation of SDN in Mininet.

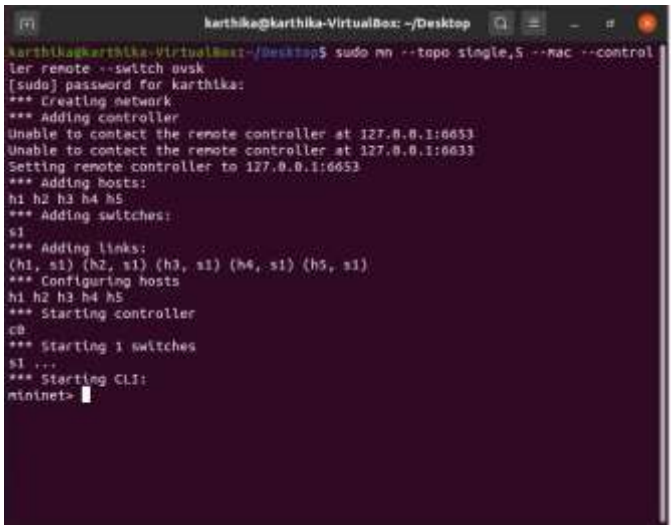


Fig. 4. SDN in Mininet.

It has added switches to three separate hosts, h1, h2, and h3, and that the links are h1 to s1, h2 to s1, and h3 to s1, forming a star topology. It was unable to reach the remote controller on the local PC every time it attempted to add the controller. The controller is generally connected to two ports: 6653 and 6633. It is looking for the controller, but no controller has been executed yet.

The next step is to run the controller in the RYU controller's mininet directory. The following command is used to start the controller,

PYTHONPATH=. ./bin/ryu-manager
ryu/app/simple_switch_13.py

The ryu-manager application is set to run in verbose mode, and it will configure the switch as well as install the forwarding rules. The default python script used inside the RYU controller as shown in Fig.5 and it performs similar to a forwarding manager. It assists in packet forwarding from one machine to another.

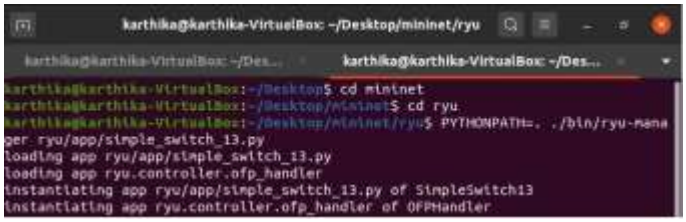


Fig. 5. Connection of RYU Controller to the Switch.

Now, the following command is used to ping the hosts, h1 ping h2

Fig.6 depicts how specific packets are delivered to the controller, which then configures the associated switch based on that packet.

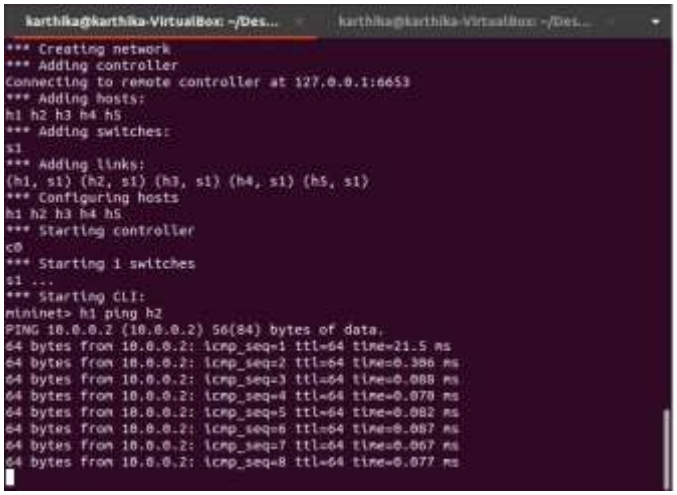


Fig. 6. Successful Packet Transfer.

When we examine the switch's response time, the initial packet sent took 21.5ms, while the remaining ping packets took 0.306ms and 0.088ms. Because the switch has no knowledge of how to forward the first packet when it arrives. As a result, the switch generates an OpenFlow event, which is forwarded to the controller. Fig.7 depicts the OpenFlow event that have been generated.

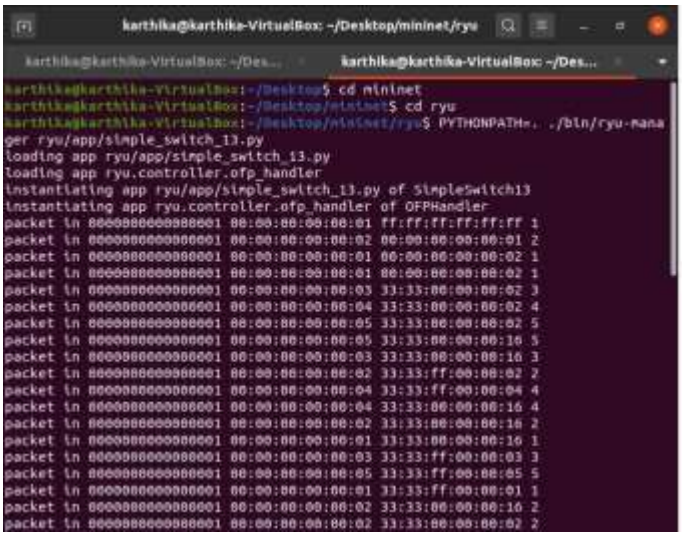


Fig. 7. OpenFlow Event.

The OpenFlow event will be generated and transmitted to the appropriate switch, which will then forward it to the appropriate RYU controller application. That specific switching application will build the rules, configure the switch with the rules, and then forward the packet. The packet will remain in the switch's buffer throughout this period. As a result, the initial packet has a higher delay, whereas the remaining packets have a shorter delay.

III. Artificial Intelligence

Artificial Intelligence (AI) is the process of teaching machines to require human intelligence, particularly the human brain and its reasoning abilities. AI systems develop the ability to reason and conduct actions that have the best likelihood of reaching a certain goal, similar to the human brain.

A. Need for Artificial Intelligence in SDN

The diverse network infrastructure adds complexity to networks and creates a slew of issues for organizing, controlling, and maximizing network resources effectively. Traditional network systems are designed to be dispersed, with every node, such as a remote device like switches and routers, seeing and reacting to just a minor portion of the system. Learning to offer control outside the local domain from nodes with only a partial perspective of the entire system is a challenging process. The training process has been made easier because to recent improvements in Software Defined Networking (SDN).

In SDN, both the control and data planes are decoupled. In an SDN architecture, the data plane contains real and virtual switches that serve as forwarding devices. Remote switches are software-based switches that work with a number of different operating systems. Using the Control Plane's structure, these data plane switches are responsible for forwarding, discarding, and manipulating packets (CP). The CP can use the Southbound Interfaces (SBIs) interface to regulate the data plane's converting and forwarding capabilities.

Control plane stands "brain" regarding SDN system, capable of programming network sources, dynamically updating forwarding guidelines as well enabling formative and agile network administration. The central controller, which is responsible for managing communication between forwarding devices and applications, is the most important part of CP. On the one hand, the controller takes network status data from the data plane and passes it along to the application plane. In other circumstances, the controller develops custom rules based on application requirements and assigns them to promotional items. Important network application capabilities including network topologies storage, state data notification, device structure, and shortest path routing are all provided by the controller.

The Networking Operating System (NOS) handles network resources with a logically centralized controller (NOS). The SDN controller has the ability to programme the network in real time. The centralised controller has a complete perspective of the network by observing and accumulating real-time network state and configuration data, as well as packet and flow graininess statistics. The following factors justify the usage of machine learning performances in SDN.

1) Recent advances in computing technology, such as the Graphics processing unit (GPU) and Tensor Processing Unit (TPU), give a perfect chance to apply credible machine learning approaches to the network area (e.g., Deep Neural Networks) [12], [13].

2) Accounting Data is vital factor to the algorithms for data-driven basic cognitive process. The central Controller has a comprehensive network interpretation and the ability to collect a large amount of network data, allowing machine learning approaches to be used.

3) By accessing data, upgrading networks, and automating network service delivery with legitimate and previous network data, machine learning algorithms can provide data to the SDN controller. Furthermore, SDN's programmability allows the network to implement the optimal network solutions (Example: Resource allocation & configuration) identified by machine learning algorithms in real-time.

ML is an area of particular study focuses on design methods that can acquire automatically from information and encounter hidden design not including explicitly programmed to do so [14]. Classification of ML algorithms depend on their learning approach and functional similarities [14]. Fig .8, summarizes ML methodologies according to their learning approach.

Machine learning approaches are considered efficient strategies in order to increase detection rates, decreasing false alarm rates, and decreasing the costs of computing and transmitting [15]. Machine learning approaches are classed as either supervised, unsupervised, or semi-supervised [16].

Because of their high classification power and computational efficiency, support vector machine (SVM) approaches are extensively used in NIDS research. They can be used with information that has a lot of dimensions. It is, nevertheless, critical to utilize the correct kernel function. A resource-intensive program places a high premium on computational processing units and memory [14]. While random forest method [17] is collective supervised learning approach for dealing along unequal data and vulnerable to over fitting.

Unsupervised learning methods derive the configuration and illustrations of data from enabled inputs. Unsupervised learning algorithms anticipate unidentified data by modelling entire system or delivery of the data [15]. Techniques for feature contraction, such as PCA, and clustering, such as self-organizing maps, are included in unsupervised learning methods (SOM).

PCA is an approach that significantly accelerates unsupervised feature learning [24]. Numerous scholars utilize PCA to pick features before performing classification. Clustering techniques like the K-means algorithm and other distance-based learning algorithms are used to find anomalies. The problem with using clustering algorithms to discover anomalies is that they are vulnerable to early conditions like the centroid, which can lead to a large number of false positives [18].

Semi-supervised learning is a type of supervised learning that uses unlabelled data for training and labelled data for testing. The training data set is made up of a small amount of tagged data and a big number of unlabelled data. It's beneficial in situations where significant amounts of tagged data aren't available, such as image archives with only a subset of the images labelled (for example, a person's image within a group photo). Simultaneously, the vast majority are not labelled [19]. MPCK-means, a semi-supervised clustering algorithm, was employed to improve the detection system's performance [20].

B. Distributed Denial-of-Service Attack Mitigation in Software Defined Network

Mitigation of distributed denial of service (DDoS) attacks is also crucial for protecting network resources under assault. Researchers used packet migration, intake bandwidth restriction, connection migration, modifying time outs, and a controller to manage protocols to resist DDoS attacks in networks based on Software-Defined networking architecture.

Shin et al. [21] developed a technique for mitigating saturation attacks by extending the Open Flow data plane's capabilities. They improved Avant-Guard by including two new modules: a network migration section and a trigger activation module. Before alerting the control plane, the connection migration module might move failed TCP sessions to it. The actuating trigger element collects network and packet payload data and uses it to trigger various flow rules depending on the situation. To demonstrate their solution, they employed the Net FPGA architecture.

Wang et al. [22] promoted protection for SDN networks using a lightweight, active and protocol-autonomous structure called Flood Guard. The proactive segment dynamically generates aggressive flow procedures based on SDN controller's run-time logic, preserving network strategy requirement. To avoid getting overwhelmed, the packet migration segment caches packets and transfers them to the controller via rate-limiting and round-robin forecast.

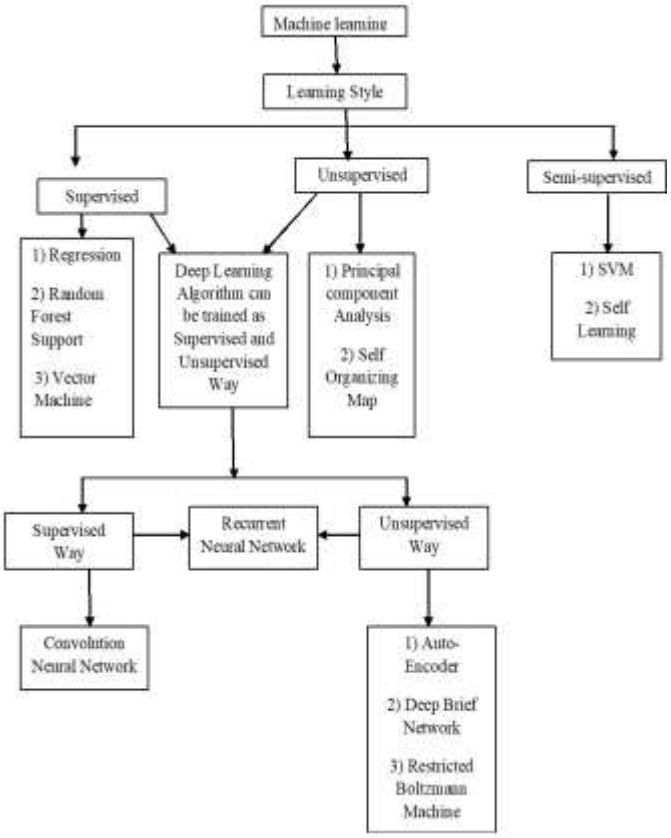


Fig. 8. ML Methodologies.

Piedrahita et al. [23] developed FlowFence, quick and lightweight DDoS attack mitigation method. The degree of use of router and SDN controller interfaces is monitored in this approach to determine the state of congestion. When a router identifies congestion on one or more interfaces, it alerts the controller, who orders the router to limit bandwidth on those interfaces.

Wang et al. [24] proposed an assured method for access control that requires entities to be authenticated. One such approach comprises three modules:

- Policy management, authentication and registration.
- Access mechanism and communication strategy.
- Trace back with audit strategy.

To communicate with another entity, it must first schedule with validation and registration segment which offers a passcode for subsequent message. They realized all components at the SDN architecture's application layer. By creating a POX controller, they validated their technique.

Yuan et al. [25] employed a peer support technique to minimize DDoS attacks on flow table overflows by pooling the available unused RAM throughout the entire SDN system. Their approach takes into account all switches on a peer-to-peer basis. When a switch is attacked, other switches will assist the targeted switch by donating their unused flow table space, thereby minimizing the DDoS attack. They approximated the vacant areas of switches that are not under attack using queuing theory.

Dridiet al. [26] proposed a unique SDN guard system for defending SDN networks versus DDoS outbreaks by dynamically rerouting malicious traffic as well managing flow-time outs. They built the solution by means of Mininet as well validated that it can reduce controller performance by up to 32%.

To avoid flooding attacks, Phan et al. [27] presented an effective approach based on support vector machines dubbed Idle-time Adjustment (IA). Before begin, the flow collector accumulates data from switches, which is subsequently extracted by the extractor. Following that, SVM-I processes the related features. Following that, whichever the flow is passed to the strategy implementation module or also the IA algorithm, depending on the outcome of SVM-I. The IA algorithm will handle the flow if the result is standard; if it isn't, it will be sent to strategy implementation, which will run a novel framework.

Sahay et al. [28] suggested a solution called ArOMA towards mitigating DDoS attacks by leveraging the SDN's centralized manageability and programmability highlights. At the ISP end, a controller receives the alarm and generates a switch policy to manage the DDoS attack. They utilized a RYU controller to validate the strategy.

Hameed et al [29] developed a combined way for defending SDN against DDoS attacks. They set the Controller-to-Controller protocol (C-to-C), enabling SDN controllers to impart and securely exchange threat information. They used Mininet to create the POX controller for authentication purposes.

Conti et al. [30] suggested a DDoS mitigation strategy in SDN that combined route spoofing and resource fatigue. Selective Blocking gathers internet protocol and MAC address data and sends it to the controller for further processing. Regular observation measures the entropy of destination address (Internet protocol) and port to establish the dataspace between them to detect probable aberrant behavior. On Mininet, they implemented a target scenario.

The northbound program was utilized by Karmakar et al. [31] to mitigate DDoS assaults in SDN. To combat DDoS attacks, this system took advantage of the specification and storing of security policies. They used the ONOS controller to validate their technique.

To secure the control plane from DDOS attacks, Wang et al. [32] proposed the Safe-Guard-Scheme (SGS). The BPNN approach is used by the anomaly detection module to find any irregularities in the given network flow. Using flow-blocking rules to remap a controller's flows stops the hosts from transmitting bogus traffic.

To counteract the Domain Name System amplification threat, Houda et al. [33] developed the wisdom SDN. To map DNS requests and responses one to one, the suggested method employs a proactive and stateful technique. The DDoS detection module collects flow characteristics to assess network traffic unpredictability before using a Bayes network-based filtering algorithm to categorise bogus DNS requests based on entropy. If the classified illegal traffic features' speed exceeds the band, the DNS mitigation (DM) mechanism systematically drops the illegitimate DNS request.

Adaptable modular frameworks, according to Daz et al. [34], can identify and mitigate LR-DDoS assaults utilizing SDN settings. The proposed work employed the CIC Dos dataset to analyze the performance of six machine learning methods for training the intrusion detection system: Random Tree, J48, RF, REP Tree, SVM, and MLP.

In order to improve the accuracy of detection with low-rate DDoS attacks, Zhijun et al. [35] developed a multi-feature DDoS attack detection approach based on FM principles and investigated the mechanism of attacks outside of the SDN data layer. This paper proposes a defense strategy based on the fundamentals of dynamic deletion in flow rules, and the results are studied to demonstrate the defense strategy's effectiveness. Some of the existing approaches challenges are listed in Table. II.

TABLE II. CHALLENGES OF EXISITNG APPROACHES

Author	Approaches	Challenges
Shin et al. [21]	Avant-Guard	Network scanning attacks and TCP SYN flood resilience may be increased by the connection migration components of Advent - Guard. As a result, network developers protecting against DoS attacks or using TCP and UDP may not find it useful. Normal network connections experience a slight but noticeable delay when connection migration is used.
Wang et al. [22]	Flood Guard	The proposed method faces two difficulties. The first is the deployment of a single data plane cache to serve all switches. Another difficulty is the usage of TCAM, which does not have the memory to carry out all proactive requirements.
Piedrahita et al. [23]	FlowFence	The proposed effort focuses on simple bandwidth to reduce DoS impact rather than wider topologies.
Wang et al. [24]	Software defined security networking mechanism (SDSNM)	It has limited influence on finding the attacker with the host in the botnet when access control is lax, and it has no impact in finding the genuine attacker.
Yuan et al. [25]	QoS-aware mitigation strategy	The proposed effort focuses on preventing switches from becoming overloaded, rather than preventing the attacker node from gaining access to the network
Dridiet al. [26]	SDN-Guard	Instead of discarding the flow, the proposed solution predicted it as harmful if it crossed the threshold value and routed it to its destination via least-used links with high time-out. As a result, the amount of bandwidth consumed by switches grows.
Phan et al. [27]	Idle-time Adjustment (IA)	The proposed study focuses on specific sorts of assaults, such as ICMP and TCP SYN flooding, rather than broader forms of attacks.
Sahay et al. [28]	ArOMA	The proposed method was tested using a simple network environment with only one controller and no real-time mitigation mechanism is provided.
Conti et al. [30]	Route Spoofing and Resource Exhaustion	The number of attacks detected using the proposed approach is higher, but the precision is a little weak.
Houda et al. [33]	wisdom SDN	Large values cause flow rules to stay in the OF table for a long period, exhausting the TCAM of OF switches, while tiny values cause legitimate DNS responses to be dropped.
Daz et al. [34]	6 ML algorithm (J48, Random Tree, REP Tree, Random Forest, MLP, SVM)	The administrator must manually intervene in order to reset the host's flow, drop probability.

IV. DEEP LEARNING

DDoS attacks are still the most common and lethal danger to current and next-generation network systems. DDoS attacks have evolved besides in frequency and severity but also in sophistication overtime. Transport layer DDoS attacks like TCP-SYN and UDP flooding, as well as network layer DDoS operations like ICMP flooding, were the most common threats to networks. As ML and DL's capacity to detect threats improves, more challenging and precise DDoS operations, known as application-layer attacks, emerge. DDoS application-layer assaults are more advanced and focused threats that exploit a server's resources. As a result, traditional attack detection techniques that rely on packet-level data are rendered ineffective.

To identify DDoS attacks, data from network traffic flow must be used to build a network-based Intrusion Detection System (IDS) that employs cutting-edge networking techniques like Software-Defined Networking (SDN). The control plane (CP) is detached from the network in SDN, which is a revolutionary networking prototype. The aforementioned technique differs from traditional network design in how it works. Users can use this technology to dynamically recreate routing operations in network systems like switches and routers. These capabilities enable in-line and network-based threat detection and mitigation measures to be implemented.

Deep Learning algorithms are new evolution of Artificial Neural Networks (ANN)which use plentiful, inexpensive computers. Deep learning enables an algorithm to discover representations for data that exhibit varying degrees of generalization. These algorithms have been used in various fields, including network intrusion, object detection, detection and visual object recognition [36]. A deep learning structure perhaps trained in either supervised or unsupervised fashion [15]. Supervised training of a deep learning algorithm, Convolution Neural Networks (CNNs) [37] remain usually taught in a supervised manner. CNN is presently de facto typical model for the applications of computer-vision.

A. Unsupervised Deep Learning Algorithm

The auto encoder [38] utilized to discover a description (encoding) for a collection of data to reduce its dimension. When trained unsupervised on collection of examples, a Deep Belief Network (DBN) [39] might train to rebuild its data. After that, the layers operate as feature detectors for the data. Following aforementioned learning stage, a DBN is trained further to do categorization in supervised manner. DBNs, also known as restricted Boltzmann machines RBM's or an auto-encoders, are helpful for feature learning, dimension reduction, topic modelling, regression and collaborative filtering.

B. Supervised or Unsupervised Algorithm

Recurrent Neural Network (RNN) algorithm [38] is a method for supervised or unsupervised learning. This network might process inputs in random order by utilizing internal memory. RNNs are frequently used in speech recognition [38]. These networks are effective at predicting characters in the text and recognizing patterns that have existed for a long time. Recent advances in deep learning algorithms for identifying and mitigating DDoS assaults in SDN are summarized in the Table. III.

TABLE III. TECHNIQUES ON RECENT DEEP LEARNING- DETECTING AND MITIGATING DDOS ATTACKS IN SDN

Publication	Deep Learning Techniques	Traffic collection	Tool used	Inference / Challenges	Accuracy
Nisha Ahuja et al, 2021 [40]	SAE-MLP	Mendeley Data	Ryu controller	On the basis of the dataset's features, network traffic is classified as normal or malicious.	99.75%
Auther Makuvaza et al, 2021 [41]	DNN	CICIDS 2017	CICFlowMeter	Improved F1 score, precision and recall	97.59%
Noe M Yungaceila-Naula et al, 2021 [42]	RF, SVM, KNN, MLP, CNN, GRU, LSTM	SDN Controller and CICDoS2017, CICDDoS2019	ONOS Controller and CICFlowMeter	The architecture's deployment simplifies its migration to production environments.	Above 99% using two public datasets
Arul and Punidha, 2021 [43]	Supervised Deep Learning Vector Quantization	MemCached server	Learning vector quantization (LVQ)	By analyzing the efficiency of cloud-mounted systems, the limitations of a static and interactive grouping of various DDoS-encrypted cross-site assault detection methodologies are overcome.	97.23%
Lu Wang, Ying Liu, 2020 [44]	CNN	POX SDN Controller and CICIDS2017	POX Controller	With a high recall and F-score, accurate and precise results are obtained. The time required to train a neural network can be reduced.	98.98%
Shahzeb Haider et al, 2020 [45]	CNN	CICIDS-2017	CICFlowMeter	Attack detection is precise, despite the computational complexity.	99.45%
Lotfi Mhamdi et al, 2020 [46]	SAE- 1SVM	CICIDS2017	CICFlowMeter	Works well with unbalanced and unlabeled datasets, resulting in more accurate and improved attack detection.	99.35%
Mahmoud Said Elsayed et al, 2020 [47]	RNN	CICDDoS2019	CICFlowMeter	Results that are significantly improved in respect of precision, F-score, and recall	99%
Beny Nugraha and Rathan Narasimha Murthy, 2020 [48]	CNN-LSTM	SDN Controller	ONOS Controller	When confronted with a huge dataset, deep learning prototype performs conventional prototype.	99.99%
Trung V. Phan et al, 2020 [49]	RL Technique	SDN Controller	ONOS Controller	Improved precision, recall, F-score and accuracy. But the proposed scheme was validated for selected network scenarios	Above 90%

V. REASERCH CHALLENGES

Though SDN enhances network speed and network monitoring management, intelligence centralization comes with its own set of security, scalability, and elasticity issues. SDN presents a number of security challenges, which are listed in this section.

A. OpenFlow Switches / Flow Table Pace

DDoS attacks against OpenFlow switches can be launched through a number of network devices to slow down or stop legal flow. The size of the OpenFlow table of the switches is one of the main vulnerabilities of SDN. Due to the growing demand for a fast and reliable data plane, flow tables are typically implemented using TCAM, which is highly expensive and limited in size [25]. By forwarding attack flow for route discovery, these compromised switches will overwhelm the controller. As a result, these compromised switches will become a major constraint for the entire network.

B. Traffic Flow

The majority of DDoS attacks are intended to generate traffic that appears to be legitimate (Low-rate DDoS attack) and is difficult to detect [23, 34,35]. The mitigation module will block the flow if the present flow exceeds the rate limit because it is unable to discriminate between regular and malicious flows. This degrades the network performance. As a result, a legitimate and robust security solution is required that can effectively differentiate between benign and anomalous network data flows.

C. Communication Links

Network performance could be harmed if communication links between switches and controllers fail. The attacker can utilize resources in both the data plane and the control plane by delivering a large number of table-miss messages. When a switch receives a new flow for which there are no matching flow rules in the flow table [22], the data plane will request actions from the control plane. As a result, scalability and security problems arise.

D. Single Point of Failure

The control plane and the data plane are decoupled in Software Defined Networking (SDN), which makes it easier to deploy new services. In the meantime, a controller faces a security threat. Because of SDN's centralized nature, the

controller can become a bottleneck, and attackers can use this flaw to perform distributed denial-of-service (DDoS) attacks against it through switches [30, 32, 44]. The attackers may be able to bring down the entire network if the centralized controller is compromised. The research community faces an open problem in developing a robust and reliable controller.

VI. Discussion

In this study, various proposal for detection and mitigation of DDoS attacks in SDN are discussed. However, the main goal of this study is to derive certain conclusions about ML/DL detection methods.

Many of the studies included in this paper employ a simulated dataset rather than a real one, which reduces the accuracy level. The learning phase of ML/DL techniques is used to learn from a specified dataset and build a training model to detect patterns. Although several studies have shown promising results in detecting assaults, it is usually recommended that the methodologies be tested in a large-scale network.

As attackers might devise new techniques to launch new attacks, various studies sought to mitigate specific types of attacks, leaving the approaches open to other types of DDoS attacks. Another note is that few studies used simulation tools to initiate an attack flow and normal flow, but real-world DDoS attackers employ a compromised host to launch a DDoS attack. This method should be used to validate the effectiveness and resilience of a defense system in a real-world setting.

VII. Conclusion

Software-defined networks are the way of the future. It enables abstraction with its programmable features. The rise of SDNs also poses security concerns due to the architecture's centralized intelligence. With the continued growth of extensive data and computing capacity, deep learning methods have exploded in popularity and are now widely used in various fields. Deep learning has the potential to extract more accurate representations from data to generate significantly more accurate models. This paper examines the use of ML/DL approaches in SDN systems to mitigate DDoS attacks. The Convolutional Neural Network-Long-Short Term Memory (CNN-LSTM) model is determined to be an effective and efficient way for identifying slow DDoS attacks in the software-defined network environment, according to the accuracy gained in the review paper. With the survey mentioned above on Deep learning techniques, we intend to continue working and touching on other areas in the future to fully exploit the significant potential of deep learning techniques for DDoS.

References

[1] H. Zhang, Z. Cai, Q. Liu, Q. Xiao, Y. Li, and C. F. Cheang, “A survey on security-aware network measurement in SDN,” Security and Communication Networks, Article ID 2459154,2018.

[2] J. Cao, M. Xu, Q. Li, K. Sun, Y. Yang, and J. Zheng, “Disrupting SDN via the data plane: a low-rate flow table overflow attack, “in Proceedings of the 13th EAI International Conference on Security and Privacy in Communication Networks, Niagara Falls, Canada, October 2017.

[3] G. Mantas, N. Stakhanova, H. Gonzalez, H. H. Jazi, and A. A.Ghorbani, “Application-layer denial of service attacks: taxonomy and survey,” International Journal of Information and Computer Security, vol. 7, no. 2-4, pp. 216–239, 2015.

[4] D. Kumar, “DDoS attacks and their types,” Network Security Attacks and Countermeasures, p. 197, 2016.

[5] MIT. (1999) Darpa intrusion detection attacks database. [Online]. Available: http://www.ll.mit.edu/ideval/docs/attackDB.html

[6] Y. Li, Z. Cai, and H. Xu, “LLMP: exploiting LLDP for latency measurement in software-defined data center networks,” Journal of Computer Science and Technology, vol. 33, no. 2, pp. 277–285, 2018.

[7] H. Lin and P.Wang, “Implementation of an SDN-based mechanism against DDOS attacks,” in Proceedings of the 2016 Joint International Conference on Economics and Management Engineering (ICEME 2016) and International Conference on Economics and Business Management (EBM 2016),Pennsylvania, Penn, USA, 2016.

[8] J. G. Yang, X. T. Wang, and L. Q. Liu, “Based on traffic and IP entropy characteristics of DDOS attack detection method,” Application Research of Computers, vol. 33, no. 4, pp. 1145–1149,2016.

[9] A. Saied, R. E. Overill, and T. Radzik, “Detection of known and unknown DDOS attacks using artificial neural networks,” Neurocomputing, vol. 172, pp. 385–393, 2016.

[10] N. Z. Bawany, J.A. Shamsi, and K. Salah, “DDOS attack detection and mitigation using SDN: methods, practices, and solutions,” Arabian Journal for Science and Engineering, vol. 42, no. 2, pp.425–441, 2017.

[11] X. Wang, M. Chen, C. Xing, and T. Zhang, “Defending DDOSattacks in software-defined networking based on and destination IP address database,” IEICE Transaction on Information and Systems, vol. E99D, no. 4, pp. 850–859, 2016.

[12] M. Wang, Y. Cui, X. Wang, S. Xiao, and J. Jiang, “Machine learning for networking: Workflow, advances and opportunities,” IEEE Network, vol. 32, no. 2, pp. 92–99, March 2018.

[13] M. Usama, J. Qadir, A. Raza, H. Arif, K.-L. A. Yau, Y. Elkhatib,A. Hussain, and A. Al-Fuqaha, “Unsupervised machine learning for networking: Techniques, applications and research challenges,” arXivpreprint arXiv:1709.06599, 2017.

[14] Atkinson RC, Bellekens XJ, Hodo E, Hamilton A, Tachtatzis C(2017) Shallow and deep networks intrusion detection system: ataxonomy and survey. CoRR, arXiv preprint arXiv:1701.02145.2017 Jan 9.

[15] Zamani M, Movahedi M (2015) Machine learning techniques for intrusion detection. CoRR, arXiv preprint arXiv:1312.2177. 2017Jan 9

[16] Aburomman AA, Reza MBI (2016) Survey of learning methods in intrusion detection systems. International conference on advances in electrical, electronic and system Engineering (ICAEES), Putrajaya, pp 362–365.

[17] Niyaz Q, Sun W, Javaid AY, Alam M (2016) A deep learning approach for network intrusion detection system. International conference wireless networks and mobile communications (WINCOM).

[18] Bennett KP, Demiriz A (2017) Semi-supervised support vector machines. NeuralComput &Applications 28(5):969–978.

[19] Haweliya J, Nigam B (2014) Network intrusion detection using semi supervised support vector machine. Int J ComputAppl 85, 9.

[20] LeCun Y, Bengio Y, Hinton G (2015) Deep learning review. Weekly journal of science in nature international. Nature 521.

[21] S. Shin, V. Yegneswaran, P. Porras, G. Gu, Avant-guard: Scalable and vigilant switch flow management in software-defined networks, in:Proceedings of the 2013 ACM SIGSAC Conference on Computer &Communications Security, 2013, pp. 413–424.

[22] H. Wang, L. Xu, G. Gu, Floodguard: A dos attack prevention extension in software-defined networks, in: 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, IEEE, 2015, pp.239–250.

[23] A.F.M. Piedrahita, S. Rueda, D.M. Mattos, O.C.M. Duarte, Flowfence: adenial of service defense system for software defined networking, in:2015 Global Information Infrastructure and Networking Symposium, GIIS,IEEE, 2015, pp. 1–6.

[24] X. Wang, M. Chen, C. Xing, SDSNM: a software-defined security networking mechanism to defend against DDOS attacks, in: 2015 Ninth International Conference on Frontier of Computer Science and Technology, IEEE,2015, pp. 115–121.

[25] B. Yuan, D. Zou, S. Yu, H. Jin, W. Qiang, J. Shen, Defending against flow table overloading attack in software-defined networks, IEEE Trans. Serv.Comput. 12 (2) (2016) 231–246.

[26] L. Dridi, M.F. Zhani, SDN-guard: DOS attacks mitigation in SDN networks, in: 2016 5th IEEE International Conference on Cloud Networking, Cloudnet, IEEE, 2016, pp. 212–217.

[27] T.V. Phan, T. Van Toan, D. Van Tuyen, T.T. Huong, N.H. Thanh, Open-FlowSIA: An optimized protection scheme for software-defined networks from flooding attacks, in: 2016 IEEE Sixth International Conference on Communications and Electronics, ICCE, IEEE, 2016, pp. 13–18.

[28] R. Sahay, G. Blanc, Z. Zhang, H. Debar, ArOMA: An SDN based autonomic DDOS mitigation framework, Computer. Security. 70 (2017) 482–499.

[29] S. Hameed, H. Ahmed Khan, SDN based collaborative scheme for mitigation of DDOS attacks, Future Internet 10 (3) (2018) 23.

[30] M. Conti, C. Lal, R. Mohammadi, U. Rawat, Lightweight solutions DDOS attacks in software defined networking, Wireless. Networks. 25(5) (2019) 2751–2768.

[31] K.K. Karmakar, V. Varadharajan, U. Tupakula, Mitigating attacks in software defined networks, Cluster Computing. 22 (4) (2019) 1143–1157.

[32] Y. Wang, T. Hu, G. Tang, J. Xie, J. Lu, SGS: Safe-guard scheme for protecting control plane against DDOS attacks in software-defined networking, IEEEAccess 7 (2019) 3469934710.

[33] Z. A. El Houda, L. Khoukhi and A. S. Hafid, "Bringing Intelligence to Software Defined Networks: Mitigating DDoS Attacks," in IEEE Transactions on Network and Service Management, doi: 10.1109/TNSM.2020.3014870.

[34] J. A. Pérez-Díaz, I. A. Valdovinos, K. -K. R. Choo and D. Zhu, "A Flexible SDN-Based Architecture for Identifying and Mitigating Low-Rate DDoS Attacks Using Machine Learning," IEEE Access, vol. 8, pp. 155859-155872, 2020, doi: 10.1109/ACCESS.2020.3019330.

[35] W. Zhijun, X. Qing, W. Jingjie, Y. Meng and L. Liang, "Low-Rate DDoS Attack Detection Based on Factorization Machine in Software Defined Network," in IEEE Access, vol. 8, pp. 17404-17418, 2020, doi: 10.1109/ACCESS.2020.2967478.

[36] Deng L, Yu D (2014) Deep learning methods and applications. Microsoft Research. Available https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.

[37] Alom MZ, Bontupalli VR, Taha TM (2015) Intrusion detection using deep belief networks. Aerospace and electronics conference, IEEE.

[38] Hughes T, Mierle K (2013) Recurrent neural networks for voice activity detection IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, pp 7378–7382.

[39] Eid HFA, Darwish A, Hassanien AE, Abraham A (2010) Principal components analysis and support vector machine based intrusion detection system. International conference intelligent systems designand applications.

[40] Nisha Ahuja, Gaurav Singal, Debajyoti Mukhopadhyay, 2021, DLSDN: Deep Learning for DDOS attack detection in Software Defined Networking, International Conference on Cloud Computing, Data Science & Engineering, 683 – 688.

[41] Auther Makuvaza, Dharm Singh Jat, Attlee M. Gamundani, 2021, Deep Neural Network (DNN) Solution for Realtime Detection of Distributed Denial of Service (DDOS) Attacks in Software Defined Networks (SDNs), SpringerNature Computer Science, Vol 2, Issue 1, pp 1 -10.

[42] Noe M. Yungaicela-Naula, Cesar Vargas-Rosakes, Jesus Arturo Perez-Diaz, 2021,SDN-Based Architecture for Transport and Application Layer DDOS Attack Detection by Using Machine and Deep Learning, IEEE Access, Vol 10, pp 1 – 18.

[43] E. Arul, A. Punidha, 2021, Supervised Deep Learning Vector Quantization to Detect MemCached DDOS Malware Attack on Cloud, Springer Nature Computer Science, Vol 2, Issue 1, pp 1 -12.

[44] Lu Wang, Ying Liu, 2020, A DDoS Attack Detection Method Based on Information Entropy and Deep Learning in SDN, IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference, pp 1084-1088.

[45] Shahzeb Haider, Adnan Akhunzada, Iqra Mustafa , Tanil Bharat Patel, Amanda Fernandez, Kim-Kwang Raymond Choo , Javed Iqba , 2020, A Deep CNN Ensemble Framework for DDOS Attack Detection inSoftware Defined Networks, IEEE Access, Vol. 20, pp 53972-53983.

[46] Loftfi Mhamdi, Desmond McLernon, Fadi El-moussa, Syed Ali Raza Zaidi, Mounir Ghogho, Tuan Tang, 2020, A Deep Learning Approach Combining Autoencoder with One-class SVM for DDOS Attack Detection in SDNs.

[47] Mahmoud Said Elsayed, Nhien-An Le-Khac, Soumyabrata Dev, Anca Delia Jurcut, 2020, DDOSNet: A Deep-Learning Model for Detecting Network Attacks, IEEE International Symposium on "A World of Wireless, Mobile and Multimedia Networks", pp 391 – 396.

[48] Beny Nugraha, Rathan Narasimha Murthy, 2020, Deep Learning-based Slow DDOS Attack Detection in SDN-based Networks, IEEE Conference on Network Function Virtualization and Software Defined Networks, pp 51 – 56.

[49] T. V. Phan, T. G. Nguyen, N. Dao, T. T. Huong, N. H. Thanh and T. Bauschert, "DeepGuard: Efficient Anomaly Detection in SDN With Fine-Grained Traffic Flow Monitoring," in IEEE Transactions on Network and Service Management, vol. 17, no. 3, pp. 1349-1362, Sept. 2020, doi: 10.1109/TNSM.2020.3004415.

Multi-objective Batch Scheduling in Collaborative Multi-product Flow Shop System by using Non-dominated Sorting Genetic Algorithm

Purba Daru Kusuma
Computer Engineering, Faculty of Electrical Engineering
Telkom University, Bandung, Indonesia

***Abstract*—Batch scheduling is a well-known topic that has been studied widely with various objectives, methods, and circumstances. Unfortunately, batch scheduling in a collaborative flow shop system is still unexplored. All studies about batch scheduling that are found were in a single flow shop system where all arriving jobs come from single door. In a collaborative flow shop system, every flow shop handles its own customers although joint production among flow shops to improve efficiency is possible. This work aims to develop a novel batch scheduling model for a collaborative multi-product flow shop system. Its objective is to minimize make-span and total production cost. This model is developed by using non-dominated sorting genetic algorithm (NSGA II) which is proven in many multi objective optimization models. This model is then compared with the non-collaborative models which use NSGA II and adjacent pairwise interchange algorithm. Due to the simulation result, the proposed model performs better than the existing models in minimizing the make-span and total production cost. The make-span of the proposed model is 10 to 17 percent lower than the existing non-collaborative models. The total production cost of the proposed model is 0.3 to 3.5 percent lower than the existing non-collaborative models.**

Keywords*—Batch scheduling; flow shop; NSGA II; collaborative system

I. INTRODUCTION

Batch scheduling is a well-known topic in supply chain management, especially in production process. In general, batching mechanism is grouping the jobs that must be processed based on their similarities so that each batch is processed together [1]. Although its concept is simple, there are many studies in batching process because the circumstances in the production system are complex and various, such as multi-site plant [2], multiple products [3,4], deteriorating jobs [5], limited batch size [6], parallel batching [7], serial batching [8,9], and so on. There is not any single model that is the best to solve all problems in batching process. Most studies in batching process focused on the batch scheduling [10,11] while others focused on determining the batch size [2] and the number of batches [2].

Batch scheduling studies were also studied in flow shop system. In a flow shop, production is divided into several stages [12]. These stages can be production, assembling, or inspection. The flow shop can be permutation or non-permutation. In the permutation flow shop, once these jobs are sequenced, this sequence will be fixed for all stages [13]. In the non-permutation flow shop, the jobs sequence among stages may be different [14]. In the batch scheduling in a flow shop system, the jobs are batched first before sequenced.

Like the flow shop scheduling problem, batch scheduling problem is also a combinatorial optimization problem. There are many methods or algorithms to solve the batch scheduling problem, such as genetic algorithm (GA) [8,13], mixed integer linear programming (MILP) [15], shortest processing time (SPT) [10], variable neighborhood descent (VND) [10], backward dynamic programming [5], memetic algorithm [16], back-off decomposition algorithm [17], longest due date (LDD) [9], adjacent pairwise interchange (API) [9], permutation method (PM) [9], and so on. There are several parameters that were used as objective in a batch scheduling studies. The most common parameter is make-span [14,15]. Other parameters are total actual flow time (TAFT) [9,10], total tardiness [18], total earliness [18], revenue [17], overstocking [16], delay [16], inefficiency cost [16], total completion time [5,12], total energy cost [19], number of late jobs [6], run time [20], and so on.

The problem in the existing models is that most of studies in batch scheduling in the flow shop system used single flow shop system. In the single flow shop system, the orders or jobs that arrive from the customers are pooled in a single shipping/receiving unit. These orders are then transferred into the production facility if there is only one production facility or distributed into several production facilities, such as in a distributed flow shop or parallel flow shop. Due to production capacity constraint, in some studies, outsourcing is possible. Meanwhile, there is another environment where there are certain number of single flow shops that besides run autonomously, they also collaborate among each other. Each flow shop receives its own orders, but joint production is possible. Reciprocally, this flow shop can also handle orders from other flow shops to achieve global optimization.

There are several questions following this problem. The first question is what model that can be used as a basis for developing batch scheduling model for collaborative flow shop system. The second question is what optimization technique that is suitable for this model and how about its performance.

Based on these problem and question, this work aims to develop batch scheduling model for collaborative flow shops system. Its objective is to minimize make-span and total

production cost. In this work, the flow shop produces multiple products, and each order also consists of multiple products. This model is developed by using non-dominated sorting algorithm (NSGA II) because this algorithm is widely proven in many multi objective optimization works, for example in workshop scheduling [21], automatic train operation [22], wireless sensor network [23], and so on.

The contributions of this work are as follows.

1) This work proposes a novel batch scheduling for collaborative flow shop system which is different from the common single flow shop system.

2) This work proposes a novel bilateral interchange between flow shops in the joint production to ensure take-and-give mechanism which is different from the centralized production system.

3) This model proposes a more profitable mechanism by implementing additional outsourcing charge so that both parties get benefit from the interchange mechanism.

This paper is organized as follows. The background, research purpose, contribution, and paper organization are described in section one. The latest literatures or works related to the batch scheduling is reviewed in section two. The proposed batch scheduling model is explained in section three. The simulation and result are explained in section four. The deeper analysis and findings are discussed and elaborated in section five. In the end, the work is concluded in section six.

II. Related Work

In this section, some latest works that focused on the batch scheduling problem, especially in flow shop system, were explored. This exploration focuses on the objective, method, and circumstance of each work. After the explanation, this section is closed by the summarization of these literatures and stating clear reasoning of why this work is needed. This exploration is as follows.

Wu et al. [3] proposed batch scheduling with storage constraints. The objective is minimizing make-span. The mixed integer linear programming (MILP) is used to formulate the model. In it, the manufacturing system produces multiple products. The process consists of two stages: production units and storage tank. Quality checking is conducted in the second stage. The scheduling is conducted in both stages.

Feng and Hu [8] proposed order batching and sequencing in a vegetable industry. The system was modeled as parallel machine with serial batching. There are two main processes: order picking and order packing. Both processes are conducted manually. In the beginning, similar orders are batched and scheduled. The objective was minimizing total completion time. They used genetic algorithm to solve the problem.

Ackerman, Fumero, and Montagna [2] developed batch scheduling in a multisite manufacturing environment. The system consists of multiple plants. Each plant produces goods in multiple stages. There are parallel non-identical processing units in every stage. In it, orders are sent by the customers with specific release date and due date. Meanwhile, the system must deliver complete orders before their due date. The objective was to minimize make-span over all installations. The decision parameters included number of batches, batch size, batch sequencing, and batch distribution. They used MILP to formulate the problem.

Chiu, Wu, Yeh, and Wang [20] developed batching model for production system that allows outsourcing mechanism. The outsourcing is allowed due to the limited capacity of the internal production facility. The system is a hybrid fabrication plant. Meanwhile, the outsourcing mechanism offers higher production cost. During the production process, there may be some non-conforming products. These non-conforming products then will be reworked. Its objective is to find optimum run time.

Hertrich, Weib, Ackerman, Heydrich, and Krumke [6] developed scheduling for the batching machines in the flow shop system. Its objective was to minimize make-span, total completion time, weighted total completion time, maximum lateness, total tardiness, and number of late jobs. In this system, there are certain number of installed machines. As a flow shop, the production system consists of several stages. Each step is handled by a single dedicated machine. The processing time are job-independent, or it is also known as proportionate flow shop. Each machine can handle multiple jobs at the same time in batch. This system adopts parallel batching machine so that the processing time of a batch remains fix. It is different from the serial batching machine where the processing time of a batch is the sum of the processing time of all jobs in a batch. In this work, the maximum batch size becomes constraint.

Maulidya, Suprayogi, Wangsaputra, and Halim [10] proposed batch scheduling model that is implemented for hybrid flow shop system. Its objective was to minimize total actual flow time. In this work, they used heuristic algorithm which consisted of two sub algorithms: the shortest processing time and variable neighborhood descent. In this flow shop system, the production process consists of three stages. The first stage consists of unrelated parallel machines for processing common and unique components. Then, these components are assembled in the second stage. Finally, these assembled products are then differentiated into various product types in the third stage.

This previous work was then improved by Suryandhini, Sukoyo, Suprayogi, and Halim [11]. Its objective was to minimize total actual flow time. They also used the heuristic algorithm. The improvement was including the sampling inspection in the three-stage flow shop system. In this flow shop system, the production process occurred in the first and second stages. Meanwhile, the inspection process was conducted in the third stage. The batch size and sequence are same in the production stages. In there, each batch consisted of one product type. Meanwhile, the third stage consisted of common inspection machine. The inspection mechanism was sampling and it followed Dodge-Romig sampling rule. During the production, there was non-conforming products, and they could not be reworked.

Ferretti and Zavanella [19] proposed batch scheduling model for general flow shop system. Its objective is to minimize total energy cost in the system. The circumstances were as follows. The flow shop consisted of two machines.

There was not any intermediate storage for interstage buffer. Batches size could be various with minimum and maximum sizes constraint. Processing time were constant, and it was independent on the batch size. The flow shop implemented no-wait scenario. The production capacity was limited. On the other side, demand had to be satisfied during the total time. The system was designed to produce identical parts. Processing time of the second stage was longer than the first stage.

Miao, Xia, Zhang, and Zou [5] proposed batch scheduling model with proportional deteriorating jobs. Its objective was to minimize total completion time. They used backward dynamic programming to solve the problem. The system adopted parallel batch machine so that the production facility could process several jobs simultaneously. The system also did not allow interruption. All jobs were ready at time zero. Batch could be full or not.

Mostafaei and Hanjunkoski [4] proposed formulation for batch scheduling. Its objective was to minimize the completion time. They used continuous time integer linear programming to formulate the model. The system allowed multiple intermediate due date. The production units were shared (not dedicated) and could produce multiple products.

Valdez-Navarro and Ricardez-Sandoval [17] integrated dynamic optimization and scheduling in the batch plant. The system was batch plant that consisted of multiple production units that could produces multiple products. Its objective was to maximize process revenue under fixed make-span. They used back-off decomposition algorithm to solve the problem. The demand was stochastic.

Ogun and Alabas-Uslu [18] proposed mathematical model to solve the batch scheduling problem in the system that produced multiple products. On the other side, each order consisted of multiple products too. Its objective was to minimize total tardiness and earliness. This system implemented parallel batch model. All parts of the same products that come from different orders could be processed in the same batch. The batch size was limited. This mathematical model was developed by using non-linear integer programming and linear integer programming.

Wu et al. [15] proposed a scheduling model that combined production and maintenance. Its objective was to minimize make-span. They used mixed integer linear programming to formulate the problem and model. This model was implemented in a batch plant that produced multiple products. In it, product degradation might occur. The system produced multiple grade products. It was a flow shop system. Each stage had several parallel units although the processing time was serial due to the quantity of products in a batch.

Yusriski, Astuti, Biksono, and Wardani [9] proposed integer batch scheduling model with single machine scenario. Its objective was to minimize total actual flow time. They compared three optimization models: longest due date (LDD), adjacent pairwise interchange (API), and permutation (PM) . In it, the jobs arrived with different due date. Every job consisted of one or more parts. A machine processed a job into number of batches. The decision parameters were jobs sequence, number of batches, batches sequence, and batches size. It followed serial batch model where the processing time of a batch was the sum of processing time of all parts in it.

Based on this exploration, these works can be summarized as follows. All of studies conducted single flow shop system with single or parallel production lines. Some works deployed parallel batching while others deployed serial batching. Most of them focused on minimizing make-span. The other objectives were to minimize TAFT and cost. Based on it, batch scheduling model that occurs in a collaborative flow shop system with multiple flow shops in the environment is still unexplored. So, developing a batch scheduling model in a collaborative flow shops environment is challenging.

III. Proposed Model

Before the proposed model is explained, first, the comparison between the environment in the single flow shop system and the collaborative flow shop system will be explained. The illustration of a single flow shop system is shown in Fig. 1.

The explanation of the Fig. 1 is as follows. There is a flow shop system with single door for shipping and receiving activities. The role of this door is to receive order from customers and deliver products to customers based on their order. This flow shop system has one plant or multiple parallel plants. The shipping/receiving unit then distributes orders to the plants. Then, after the plants complete the production, these products are then transferred back to the shipping/receiving unit to be delivered to the customers.

This condition is different from the collaborative flow shop system. In a collaborative system, there are multiple flow shops. Each flow shop interacts with its own customers. It receives orders from its customers and delivers the requested goods to them. Each flow shop has its own production facility. The production cost, production capacity, and processing time may be different among flow shop. In some circumstances, it is better to outsource some jobs or parts of jobs to other flow shops in the system, for example, due to the limited production capacity [20]. Reciprocally, a flow shop may get jobs or parts of jobs from other flow shops. This collaboration is illustrated in Fig. 2. In Fig. 2, there are two flow shops in the environment. Each flow shop has its own customers, shipping/receiving unit, and plant. Each flow shop only serves its own customers. Meanwhile, production outsourcing may occur between them.

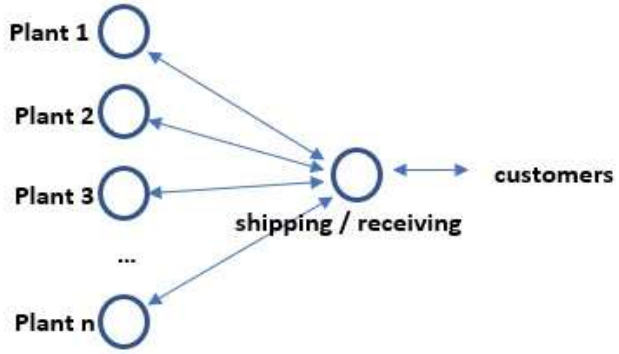


Fig. 1. Single Flow Shop System.

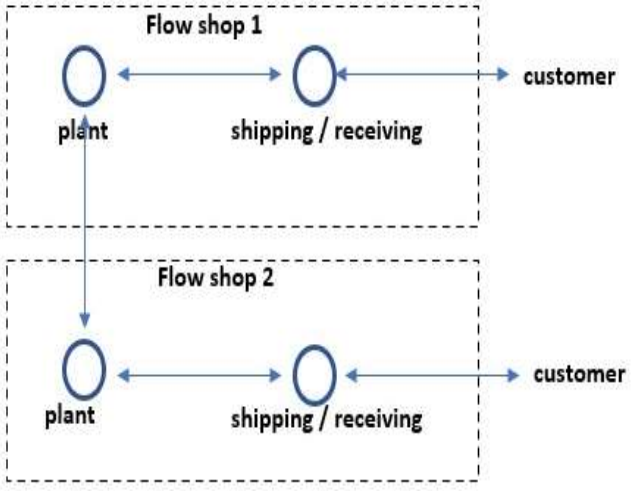


Fig. 2. Collaborative Flow Shop System.

In this work, the environment is a set of flow shops that interact to each other. Each flow shop has its own customers and production facility. The processing time, processing cost, and batch size may be different among flow shops. The objectives are to minimize make-span and total production cost. The assumptions in this model are as follows.

1) The flow shop is permutation so that once a sequence is generated, this sequence is fixed for all stages [13].
2) All flow shops have same production stages [24].
3) All products must pass all stages [24].
4) All products can be produced in any flow shop [25].
5) All orders are ready at time zero [26].
6) All production facilities are ready at time zero [26].
7) There is no due date.
8) Intermediate storage is unlimited.
9) Interruption and pre-emption are not allowed [26].
10) Parallel batching is implemented so that the batch processing time is independent to the number of jobs [7].
11) The batch size is determined based on the number of product units in a batch.

The process is described as follows. In the beginning, each flow shop receives orders from its customers. An order represents a job. A job consists of one or more products. A job is then split into packets based on the product. It means the number of packets in a job is equal to the number of products in a job. After all jobs in all flow shops are mapped into packets, then packet interchange among flow shop can be done. In this work, packet interchange is preferred to outsourcing because when a flow shop transfers a packet to be produced by other flow shop, it also receives packet from its counterpart. After packets interchange session ends, the next session is batching. Batching is conducted based on the product type and batch type. Each batch contains one or several packets with the same product.

Before further explanation of the model, here are notations that are used in this work.

a	action
b	batch
c	production cost
f	flow shop
f_{pr}	flow shop who processes the packet (packet processor)
f_{ow}	flow shop who owns the packet (packet owner)
f_{ms}	make-span fitness function
f_{tpc}	total production cost fitness function
g	product
o	order /job
p	packet
p_{sel}	selected packet
pos	packet position in a packet sequence
r	outsourcing rate
s	stage
q_b	batch's maximum quantity (batch size)
q_{cb}	batch's current quantity
q_p	packet's quantity
t_{end}	batch processing end time
t_p	processing time
t_{start}	batch processing start time
tr_{cr}	interchange threshold
tr_m	mutation threshold
B	set of batches
F	set of flow shops
G	set of products
O	set of orders /jobs
P	set of packets
S	set of stages

The first step is packets arrangement in every flow shop. The packets arrangement in the system can be modeled as $\{P(f_1), P(f_2), P(f_3), \ldots, P(f_{n(F)}),\}$. $P(f)$ is a set of packets that is owned by a flow shop and it can be modeled as $\{p(1,f), p(2,f), p(3,f), \ldots, p(n(P(f)),f)\}$. In other word, it is a two-dimensional list. The first dimension represents the flow shop. The second dimension represents the packets in a flow shop. The number of packets in every flow shop is formalized by using (1). The initial packet arrangement process is shown in algorithm 1.

$$n\big(P(f)\big) = \sum_{\forall O(f)} n\big(G(o,f)\big) \quad (1)$$

algorithm 1: initial packets arrangement

```
1  for i = 1 to n(F) do
2  begin
3   j = 0
4   for k = 1 to n(O(f_i)) do
5   begin
6   for l = 1 to n(P(o_k,f_i)) do
7   begin
8   initialize(p_{i,i})
9   q_p(j,i) = q_g(k,i,l)
10  j++
11  end for
12  end for
13 end for
```

Algorithm 1 represents the packets arrangement based on the packet owner. Packet owner is a flow shop who handles the related order where the packet belongs to. In this collaborative work, packets are then arranged based on the packet processor. In the interchange process, a packet may be processed or produced by another flow shop. It means, the packet processor is a flow shop who executes or produces the packet. If the packet is not interchanged, then the packet owner is same as the packet processor. Otherwise, the packet owner is different from the packet processor.

The explanation of the parameters and variables in algorithm 1 is as follows. The initial packets arrangement is conducted in all flow shops which is represented by the looping from 1 to $n(F)$. Then, the second loop is conducted for all orders in the flow shop which is represented by $n(O(f_i))$. The third loop is conducted to arrange all packets in every order, which is represented by $n(P(o_k,f_i))$.

After all packets in all flow shops have been arranged, the next step is batch scheduling by using NSGA II. This model is the improvement of the previous multi-objective evolutionary algorithm (MOEA) by solving the difficulties of the MOEA: (1) computational complexities, (2) non-elitism, and (3) specified sharing parameters [27]. In general, the NSGA-II consists of six steps [28]: (1) population initialization, (2) non-dominated sorting, (3) crowd distance, (4) selection, (5) genetic operators, and (6) recombination. As a population-based method, in this work, a population or a solution represents a two-dimensional packet list in a system. Based on this general process [28], the algorithm of the collaborative batch scheduling in this work is as follows.

1) In the beginning, certain number of populations are set.

2) New off-springs are generated from the current population. The off-springs population is equal to the population size.

3) Cross over and mutation occurs. In this work, cross-over means interchanging the packets between two flow shops. Cross over process is formalized by using (2) to (8). Mutation means interchanging the position between two packets in the packet arrangement of a flow shop. This internal flow shop mutation occurs based on certain probability. This mutation is formalized by using (9) to (12).

4) Batching process occurs in all flow shops for all population.

5) Production process occurs based on the batch sequence.

6) Fitness calculation is conducted.

7) Non-dominated sorting is conducted based on the fitness functions (make-span and total production cost).

8) Crowd distance is conducted to sort solution in every group or front.

9) The first half of population is selected as the next parents and the rest are delisted from the population.

*10)*Step 2 and 9 is repeated until the maximum iteration is reached.

The initial packet interchange occurs in the initial population set up. Its process occurs based on stochastic process. This process is formalized by using (2) to (8). Equation (1) shows that interchange occurs only when the generated uniform random number is below the interchange threshold. Equation (3) shows that the interchange occurs by switching two selected packets in two selected flow shops. These flow shops are selected by using uniform random among flow shops in the system as it is shown in (4) and (5). Then, the packet position candidates are determined by using (6) and (7). Finally, the final packet position is the lowest position between these candidates to avoid misposition, due to the different number of packets in both flow shops, as it is shown in (8).

$$a = \begin{cases} interchange, U(0,1) < tr_{cr} \\ not\ interchange, else \end{cases}, \forall p\ in\ f \quad (2)$$

$$a = switch\big(p(f_{sel1}, pos_{sel}), p(f_{sel1}, pos_{sel})\big) \quad (3)$$

$$f_{sel1} = U\big(f_1, f_{n(F)}\big) \quad (4)$$

$$f_{sel2} = U\big(f_1, f_{n(F)}\big) \quad (5)$$

$$p_{sel1} = U\big(p_1, p_{n(P)}\big), \forall p\ in\ f_1 \quad (6)$$

$$p_{sel2} = U\big(p_1, p_{n(P)}\big), \forall p\ in\ f_2 \quad (7)$$

$$p_{sel} = min(p_{sel1}, p_{sel2}) \quad (8)$$

Crossover process is like the initial interchange process. Different from the initial interchange process, the crossover occurs in every iteration. The number of crossovers in every iteration is equal to the number of total flow shops in a system. The selection of the flow shops and packets follows (4) to (7).

Mutation occurs in all flow shops in every iteration. It occurs stochastically. This mechanism is formalized by using (9) to (12). Equation (9) shows that the mutation occurs when the generated uniform random number is less than the mutation threshold. The mutation occurs by switching the packets position of the two selected packets in a flow shop as it is shown in (10). Both packets are selected among packets in a flow shop as it is shown in (11) and (12).

$$a = \begin{cases} mutate, U(0,1) < tr_m \\ not\ mutate, else \end{cases}, \forall p\ in\ f \quad (9)$$

$$a = switch\big(p(f, pos_{sel1}), p(f, pos_{sel2})\big) \quad (10)$$

$$p_{sel1} = U\big(p_1, p_{n(P)}\big), \forall p\ in\ f \quad (11)$$

$$p_{sel2} = U\big(p_1, p_{n(P)}\big), \forall p\ in\ f \quad (12)$$

Hereafter, the batching sequencing is conducted by grouping packets based on the product. The product in a batch is homogeneous. Batching is also limited to the batch size. The process occurs sequentially from the first packet to the last packet in a flow shop. New batch will be created when there is not any available batch in a flow shop. Else, the packet is added into the available batch. This process is formalized by using (13). After allocating packet to batch, then the batch current quantity is accumulated.

$$a = \begin{cases} add, \exists b \wedge g_b = g_b \wedge \big(q_p + q_{cb}\big) \le q_b \\ create, else \end{cases}, \forall p\ in\ f \quad (13)$$

The next step is production process. The batch scheduling occurs here. It follows permutation rule where batch

overlapping is not permitted. The batch scheduling is formalized by using (14) to (16). Equation (14) and (15) is used to determine the batch start time. Meanwhile, the batch end time is determined by using (16).

$$t_{start}(b,s) = \begin{cases} 1, b = 1 \\ t_{end}(b-1,s) + 1, b > 1 \end{cases}, s = 1 \quad (14)$$

$$t_{start}(b,m) = \begin{cases} t_{end}(b,s-1) + 1, b = 1 \\ max\left(t_{end}(b-1,s), t_{end}(b,s-1)\right) + 1, b > 1 \end{cases}, s > 1 \quad (15)$$

$$t_{end}(b,s) = t_{start}(b,s) + t_p(b,f) - 1 \quad (16)$$

After the production process ends, the next step is calculating the fitness function. Due to its multi objective model, there are two fitness functions: make-span and total production cost. These fitness functions calculation is formalized by using (17) to (19). Equation (17) shows that the make-span function is the maximum end time of the last stage of the last batch for all flow shops. Equation (18) shows that the total production cost is the accumulation of production cost of all packets of all flow shops. The packet production cost is calculated by using (19). In it, the packet production cost is obtained by multiplying the quantity of the packet and the unit production cost of the selected product, and its unit production cost is based on the packet processor. If the packet processor is different from the packet owner, then the outsourcing rate is included.

$$f_{ms} = max\left(t_{end}(f, n(B), n(S))\right), \forall f \quad (17)$$

$$f_{tpc} = \sum_{\forall f} \sum_{\forall p} c(p) \quad (18)$$

$$c(p) = \begin{cases} q(p).c_u(f_{pr}, g).(1+r), f_{pr} \neq f_{ow} \\ q(p).c_u(f_{pr}, g), f_{pr} = f_{ow} \end{cases} \quad (19)$$

The next step is non-dominated sorting. In this work, the minimization approach is chosen. Based on the general rule in non-dominated sorting [27], in this work, solution A dominates solution B if these conditions meet.

1) Make-span of solution A is less than or equal to B.
2) Total production cost A is less than or equal to B.
3) Make-span of solution A is less than B or total production cost A is less than B.

Based on this requirement, every solution will be compared with all other solutions reciprocally. Solutions are then grouped based on the number of solutions they dominate. These groups are called as front. The groups are descending sorted. The last work is calculating the crowd distance. This process is used to rank the solutions inside the fronts. Because both fitness functions use different metric, the fitness of every solution is normalized first before calculated. In this work, the min-max normalization is used [29].

IV. Simulation

This proposed model is then implemented into simulation to observe and evaluate the performance of the model. In the simulation, this proposed model is compared with two non-collaborative models. The first non-collaborative model uses NSGA-II algorithm [30] and the second one uses adjacent pairwise interchange algorithm [9]. In this simulation, C-NSGA II represents the proposed model which is collaborative NSGA II; NC-NSGA II represents the non-collaborative NSGA II; and NC-API represents the non-collaborative adjacent pairwise interchange. The reasons of this comparison are as follows. The proposed model, as a collaborative model is compared with the non-collaborative model to observe its improvement as it is stated as the research purpose. The reason of comparing the NSGA II with the API is to compare the performance of the multi objective approach (NSGA II) and the single objective approach (API).

In this work, the observed parameters are make-span and total production cost. Meanwhile, the adjusted parameters are number of flow shops and number of customers. Besides these parameters, the value of other variables is shown in Table 1. In the beginning, some variables are generated randomly. The number of products, average batch size, average batch processing time, average number of products per order, and average product quantity are generated randomly and follow normal distribution.

The first simulation is simulating model with various number of flow shops. In this simulation, the number of flow shops ranges from 2 units to 10 units. In this simulation, the number of customers is set 50 units. Its objective is to observe the relationship between the number of flow shops and the observed parameters. The result is shown in Fig. 3 and Fig. 4.

TABLE I. Default Variables

Variables	Default Value
Number of stages	3 stages
Number of products	3 products
Outsourcing rate	10 percent
Average batch size	200 units
Average batch processing time	10 time-unit
Average number of products per order	3 products
Average per-product order quantity	30 units

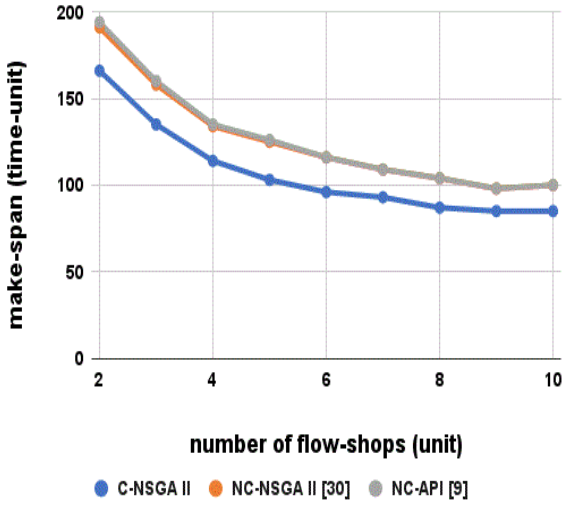


Fig. 3. Relation between the Number of Flow Shops and Make-Span.

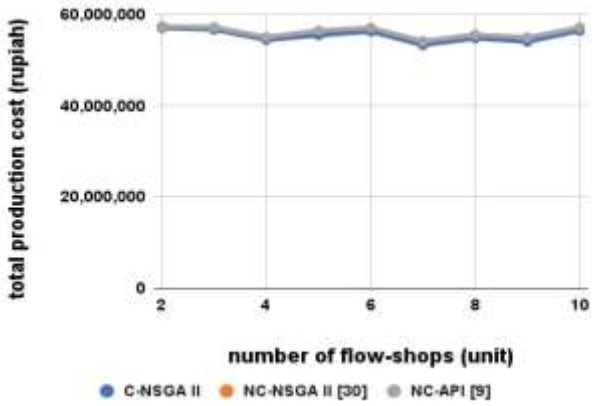


Fig. 4. Relation between the Number of Flow Shops and Total Production Cost.

Fig. 3 shows the relationship between the number of flow shops and make-span. It is shown that the make-span decreases due to the increasing of the number of flow shops. This decreasing of the make-span is the rational consequence of the increasing of the production units (flow shops) with the fixed number of customers. The make-span of the proposed model is the lowest one, compared with the existing NC-NSGA II [30] and NC-API [9]. It is 13 percent to 17 percent lower, with average 15 percent, than the existing models. Meanwhile, the make-span of the NC-NSGA II is almost equal to the NC-API with NC-NSGA II is slightly better.

Fig. 4 shows the relationship between the number of flow shops and total production cost. It is shown that the increasing of the number of flow shops does not affect the total production cost. The total production cost tends to fluctuate. By comparing among models, the total production cost is almost equal. The total production cost of the non-collaborative models is equal because the production is handled internally. Meanwhile, although almost equal, the total production cost of the proposed model is slightly lower than the existing non-collaborative models [9,30]. The total production cost of the proposed model is 0.6 percent to 1.8 percent lower, with average 1.3 percent, of the existing non-collaborative models.

The second simulation is simulating model with various number of customers. In this simulation, the number of customers ranges from 25 to 75 units. In this simulation, the number of flow shops is set 5 units. Its objective is to observe the relationship between the number of customers and the observed parameters. The result is shown in Fig. 5 and Fig. 6.

Fig. 5 shows that the make-span increases linearly due to the increasing of the number of customers. It is a rational consequence of the increasing jobs due to the fixed number of production units. The make-span of the existing non-collaborative models is almost equal with the NC-NSGA II [30] is slightly lower than the NC-API [9]. Meanwhile, the make-span of the proposed model is the lowest among models. The make-span of the proposed model is 10 to 17 percent lower, with average 14 percent, than the existing models.

Fig. 6 shows that the total production cost increases linearly due to the increasing of the number of customers. It occurs in all models. It is a rational consequence of the increasing jobs while on the other side, the number of production units is fixed. The total production cost is almost equal among all models. The total production cost of the NC-NSGA II [30] is equal to the NC-API because all jobs are handled internally. On the other side, the total production cost of the proposed model is slightly lower than the existing models. The total production cost of the proposed model is 0.3 to 3.5 percent lower, with average 1.6 percent, than the existing non-collaborative models. This gap is narrower when the number of customers is high.

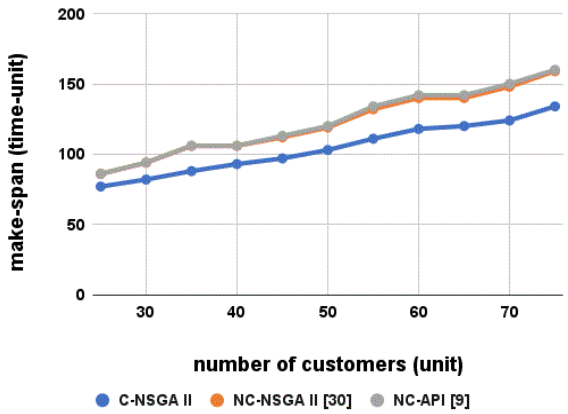


Fig. 5. Relation between the Number of Customers and Make-Span.

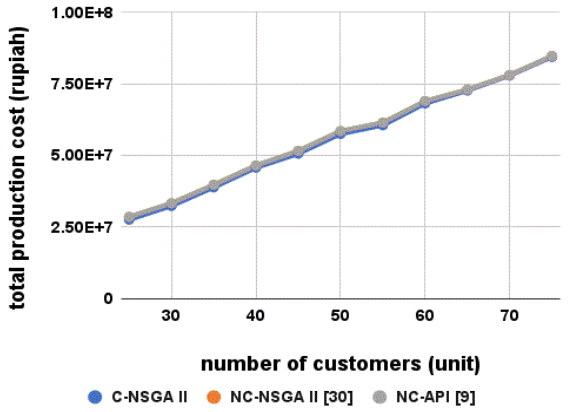


Fig. 6. Relation between the Number of Customers and Total Production Cost.

V. Discussion

Based on the simulation result, in general, the collaborative approach is better than the non-collaborative one. it is shown that the collaborative model outperforms the non-collaborative models in make-span and is slightly better in total production cost. The make-span of the proposed model is lower than the existing models, both the NC-NSGA II (30) and the NC-API [9], and its gap is wide. On the other side, although the gap is very narrow, the total production cost of the proposed model is still lower than the existing models.

In the make-span parameter, the make-span can be reduced in two ways. The first way is allocating jobs to the production

unit which has more competitive processing time. The second way is by sequencing jobs into batches so that the number of batches can be minimized. By using collaborative approach, the proposed model tries to allocate jobs to the production unit (flow shop) which has lower batch processing time. This process is conducted during the iteration of the NSGA II. On the other side, in the existing non-collaborative models [9,30], optimization or make-span reduction is conducted by arranging jobs that must be processed internally so that the number of batches can be reduced. Meanwhile, this process also occurs in the non-collaborative model. It means, the collaborative model outperforms the non-collaborative models in minimizing the make-span because it runs two optimization processes. On the other side, the non-collaborative models [9,30] run only one optimization process. The result also shows that the non-collaborative NSGA II performs better than the non-collaborative API in the make-span aspect. It means that the NSGA II [30] can arranges jobs into batches better than the API [9] although the gap is very narrow.

In the total production cost parameter, the production cost can only be reduced by transferring packets to the production unit (flow shop) which has lower production cost. This process occurs in the collaborative proposed model only. Meanwhile, this process cannot be conducted in the non-collaborative models [9,30] which execute orders internally. Unfortunately, this advantage is reduced by the outsourcing cost which is charged to packets which are produced externally. This circumstance makes the total production cost gap between the collaborative model and the non-collaborative ones is not significant.

VI. Conclusion

This work shows that the proposed collaborative batch scheduling model meets the research objective in minimizing the make-span and the total production cost. The collaborative NSGA II model outperforms the non-collaborative models (NSGA II and adjacent pairwise interchange). Due to the simulation result, the make-span of the proposed model is lower than the non-collaborative models. The make-span of the proposed model is 10 to 17 percent lower than the existing non-collaborative models. The total production cost of the proposed model is 0.3 to 3.5 percent lower than the existing non-collaborative model.

There are several future research potentials that can be conducted based on this proposed collaborative model. This model can be expanded into collaborative parallel flow shops where every flow shop has several production lines. This model is also can be improved by adding several constraints, such as due date, no-wait scenario, penalty, etc.

References

[1] K. R. Baker and D. Trietsch, Principles of Sequencing and Scheduling, Wiley, Canada, 2009.

[2] S. Ackerman, Y. Fumero, and J. Montagna, “Incorporating batching decisions and operational constraints into the scheduling problem of multisite manufacturing environments”, International Journal of Industrial Engineering Computations, vol. 12, pp. 345-364, 2021.

[3] O. Wu, G. D. Ave, I. Harjunkoski, A. Bouaswaig, S. M. Schneider, M. Roth, and L. Imsland, “Short-term multiproduct batch scheduling considering storage features”, IFAC PapersOnLine, vol. 53, no. 2, pp. 11794-11799, 2020.

[4] H. Mostafaei and I. Harjunkoski, “Single reference grid continuous-time formulation for batch scheduling”, IFAC PapersOnLine, vol. 52, no. 1, pp. 832-837, 2019.

[5] C. Miao, Y. Xia, Y. Zhang, and J. Zou, “Batch scheduling with deteriorating jobs to minimize the total completion time”, Journal of Operations Research Society of China, vol. 1, pp. 377-383, 2013.

[6] C. Hertrich, C. Weib, H. Ackermann, S. Heydrich, and S. O. Krumke, “Scheduling a proportionate flow shop of batching machines”, Journal of Scheduling, vol. 23, pp. 575-593, 2020.

[7] J. W. Fowler and L. Monch, “A survey of scheduling with parallel batch (p-batch) processing”, European Journal of Operational Research, 2021, in press.

[8] X. Feng and X. Hu, “A heuristic solution approach to order batching and sequencing for manual picking and packing lines considering fatiguing effect”, Scientific Programming, article ID: 8863391, pp. 1-17, 2021.

[9] R. Yusriski, B. Astuti, D. Biksono, and T. A. Wardani, “A single machine multi-job integer batch scheduling problem with multi due date to minimize total actual flow time”, Decision Science Letters, vol. 10, pp. 231-240, 2021.

[10] R. Maulidya, Suprayogi, R. Wangsaputra, and A. H. Halim, “A batch scheduling model for a three-stage hybrid flowshop producing products with hierarchical assembly structures”, International Journal of Technology, vol. 11, no. 3, pp. 608-618, 2020.

[11] P. P. Suryandhini, Sukoyo, Suprayogi, and A. H. Halim, “A batch scheduling model for a three-stage flow shop with job and batch processors considering a sampling inspection to minimize expected total actual flow time”, Journal of Industrial Engineering and Management, vol. 14, no. 3, pp. 520-537, 2021.

[12] G. M. Komaki, S. Sheikh, and B. Malakooti, “Flow shop scheduling problems with assembly operations: a review and new trends”, International Journal of Production Research, pp. 1-30, 2018.

[13] K. Peng, L. Wen, R. Li, L. Gao, and X. Li, “An effective hybrid algorithm for permutation flow shop scheduling problem with setup time”, Procedia CIRP, vol. 72, pp. 1288-1292, 2018.

[14] D. A. Rossit, F. Tohme, M. Frutos, M. Safe, and O. C. Vasquez, “Critical paths of non-permutation and permutation flow shop scheduling problems”, International Journal of Industrial Engineering Computations, vol. 11, pp. 281-298, 2020.

[15] O. Wu, G. D. Ave, I. Harjunkoski, A. Bouaswaig, S. M. Schneider, M. Roth, and L. Imsland, “Optimal production and maintenance scheduling for a multiproduct batch plant considering degradation”, Computers and Chemical Engineering, vol. 135, article ID: 106734, pp. 1-14, 2020.

[16] F. Miquel, M. Frutos, F. Tohme, and D. A. Rossit, “A memetic algorithm for the integral OBP/OPP problem in a logistics distribution center”, Uncertain Supply Chain Management, vol. 7, pp. 203-214, 2019.

[17] Y. I. Valdez-Navarro and L. A. Ricardez-Sandoval, “Integration between dynamic optimization and scheduling of batch processes under uncertainty: a back-off approach”, IFAC PapersOnLine, vol. 52, no. 1, pp. 655-660, 2019.

[18] B. Ogun and L. Alabas-Uslu, “Mathematical models for a batch scheduling problem to minimize earliness and tardiness”, Journal of Industrial Engineering and Management, vol. 11, no. 3, pp. 390-405, 2018.

[19] I. Ferretti and L. E. Zavanella, “Batch energy scheduling problem with no-wait/blocking constraints for general flow-shop problem”, Procedia Manufacturing, vol. 42, pp. 273-280, 2020.

[20] S. W. Chiu, H. Y. Wu, T. M. Yeh, and Y. Wang, “Solving a hybrid batch production problem with unreliable equipment and quality reassurrance”, International Journal of Industrial Engineering Computations, vol. 12, pp. 235-248, 2021.

[21] W. Yahui, S. Ling, Z. Cai, F. Liuqiang, and J. Xiangjie, “NSGA-II algorithm and application for multi-objective flexible workshop scheduling”, Journal of Algorithms & Computational Technology, vol. 14, 2020.

[22] P. Chu, Y. Yu, D. Dong, H. Lin, and J. Yuan, “NSGA-II-based parameter tuning method and GM(1,1)-based development of fuzzy immune PID controller for automatic train operation system”,

Mathematical Problems in Engineering, article ID: 3731749, pp. 1-20, 2020.

[23] V. Kanwar and A. Kumar, "Multiobjective optimization-based DV-hop localization using NSGA-II algorithm for wireless sensor networks", International Journal of Communication Systems, vol. 33, no. 11, 2020.

[24] Y. Sun and X. Qi, "A DE-LS metaheuristic algorithm for hybrid flow-shop scheduling problem considering multiple requirements of customers", Scientific Programming, article ID: 8811391, pp. 1-14, 2020.

[25] B. Naderi and R. Ruiz, "The distributed permutation flowshop scheduling problem", Computers & Operations Research, vol. 37, no. 4, pp. 754-768, 2010.

[26] F. C. Cetinkaya, P. Yeloglu, and H. A. Catkamas, "Customer order scheduling with job-based processing on a single-machine to minimize the total completion time", International Journal of Industrial Engineering Computations, vol. 12, pp. 273-292, 2021.

[27] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II", IEEE Transactions on Evolutionary Computation, vol. 6, no. 2, pp. 182-197, 2002.

[28] Y. Yussoff, M. S. Ngadiman, and A. M. Zain, "Overview of NSGA-II for optimizing machining process parameters", Procedia Engineering, vol. 15, pp. 3978-3983, 2011.

[29] H. Benhar, A. Idri, and J. L. Fernandez-Aleman, "Data preprocessing for heart disease classification: a systematic literature review", Computer Methods and Programs in Biomedicine, vol. 195, article ID: 105635, 2020.

[30] N. Farmand, H. Zarei, and M. Rasti-Barzoki, "Two meta-heuristic algorithms for optimizing a multi-objective supply chain scheduling problem in an identical parallel machines environment", International Journal of Industrial Engineering Computations, vol. 12, pp. 249-272, 2021.

Power Loss Minimization using Optimal Power Flow based on Firefly Algorithm

Chia Shu Jun[1], Syahirah Abd Halim[2]
Nor Azwan Mohamed Kamari[4]
Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering & Built Environment
Universiti Kebangsaan Malaysia, Selangor, Malaysia

Hazwani Mohd Rosli[3]
School of Engineering
Asia Pacific University of Technology and Innovation
Kuala Lumpur, Malaysia

Abstract*—Conventional methods are commonly used to solve optimal power flow problems in power system networks. However, conventional methods are not suitable for solving large and non-linear optimal power flow problems as they are influenced by initialization values and more likely be trapped in local optimum. Hence, heuristic optimization methods such as Firefly Algorithm have been widely implemented to overcome the limitations of the conventional methods. These methods often use random strategy that can provide better solutions to avoid being trapped in the local optimum while achieving global optimum. In this study, the load flow analysis was performed using the conventional method of Newton-Raphson technique to calculate the real power loss. Next, Firefly Algorithm was implemented to optimize the control variables for minimizing the real power loss of the transmission system. Generator bus voltage magnitudes, transformer tap settings and generator output active power were taken as the control variables to be optimized. The effectiveness of the proposed Firefly Algorithm was then tested on the IEEE 14-bus and 30-bus system using MATLAB software. The simulated results were then analyzed and compared with Particle Swarm Optimization's results based on the consistency and execution time. Implementation of the Firefly Algorithm has successfully produced minimum real power loss with faster computational time as compared to Particle Swarm Optimization. For the IEEE 14-bus system, the active power loss for the Firefly Algorithm is 6.6222 MW and the calculation time is 18.2372 seconds. Therefore, the application of optimal power flow based on Firefly Algorithm is a reliable technique, in which the optimal settings with respect to power transmission loss can be determined effectively.

Keywords*—Optimal power flow; firefly algorithm; real power loss; control variables

I. INTRODUCTION

Optimal power flow (OPF) is a non-linear and complex optimization technique in the operation of electrical power systems. OPF has been classified into conventional and intelligence as defined in [1]. Conventional methods are based on gradients and influenced by initial guess values such as voltage magnitude. This method requires a solution for a new linear system in each iteration [2].

In recent years, intelligence methods have been developed to overcome the limitations and short comings of the conventional methods. Intelligence methods such as Genetic Algorithm [1, 3], Ant-Lion Optimization Techniques [4, 5], Hybrid Firefly and Particle Swarm Optimization [6], Particle Swarm Optimization [7] and American Buffalo Algorithm [8] have been proposed to reduce active power loss in transmission system. In addition, the Firefly Mating Algorithm as in [9] has also been used to reduce active and reactive power loss in the transmission system. Firefly algorithm also has been suggested to tackle the economic dispatch problem [10]. Besides, the modified Firefly Algorithm has also been used to maintain various system constraints within operating limits while lowering the total cost of system generation [11].

Firefly Algorithm (FA) which acts as a nature-inspired algorithm has been used previously to solve various nonlinear design problems. It is based on the behavior of herds such as fish, bird schooling and insects in the environment [12]. The Firefly Algorithm has three ideal rules [8]. First is that all fireflies are of different genders, where one firefly will be attracted to brighter ones. The second rule is that the brightness of fireflies will decrease if the distance between them increases. This is caused by the absorption of light when it passes through the medium. However, if there is no lighter firefly nearby, the fireflies will move randomly. Thirdly, the brightness of a firefly is determined by the objective function. Compared with some other heuristic algorithm, FA is more suitable in solving optimization problems of various objectives and be able to find a better global optimal solution [13, 14]. The next major advantage of this method is the ability to learn, fast computation time and suitable to solve the problem of non-linear and convex optimal power flow [6, 15].

In this study, Firefly Algorithm was integrated with the Newton-Raphson load flow formulation to determine the optimal settings of the control variables such as generator bus voltage magnitudes, transformer tap settings and generator output active power. To demonstrate the effectiveness of the proposed method, the IEEE 14-bus and 30-bus systems were modelled in MATLAB software and utilized as the test systems. Further validations on the effectiveness of the method were conducted by comparing its consistency and computational time with Particle Swarm Optimization method. The implementation of the Firefly Algorithm-based power flow is expected to produce a minimum amount of active power loss while improving the voltage profile, subjected to the system operational constraints.

Ministry of Higher Education Malaysia (MOHE) and Universiti Kebangsaan Malaysia (UKM).

II. METHODOLOGY

A. Firefly Algorithm

Fireflies produce natural light to attract others. Firefly Algorithm is mathematically modeled based on hunting behavior. There are two important factors in the formulation of Firefly Algorithm, which are light intensity and attractiveness. The intensity of light varies according to the inverse square law as in (1).

$$I(r) = \frac{I_s}{r^2} \quad (1)$$

where $I(r)$ is the intensity of light at the distance of r and I_S is the intensity at its source. The attraction of a firefly is proportional to the light intensity detected by the neighbouring fireflies, therefore the attractiveness function can be defined as stated in (2).

$$\beta_j(r) = \beta_o e^{-\gamma r^m}, \mathrm{m} \geq 1 \quad (2)$$

where γ is the light absorption coefficient, β_o is the attractiveness for $r = 0$, r is the Cartesian distance between two fireflies which defined in (3).

$$r_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2} \quad (3)$$

where x_i is the location of firefly i, $x_{i,k}$ is the k[th] component of the spatial coordinate x_i of i-th fireflies. The movement of firefly i attracted to the brighter firefly j is defined by (4).

$$x_i^{k+1} = x_i^k + \left[\beta_o e^{-\gamma r^2}\right]\left(x_j^k - x_i^k\right) \quad (4)$$

$$+\alpha * [rand - 0.5]$$

where the first term on the left side is the initial position of firefly i, second term is the attraction of j^{th} fireflies while the third term introduces random movement, and the rand is a random number generator uniformly distributed between zero and one.

B. Formulation of the Optimal Power Flow

The purpose of solving the optimal power flow problem is to optimize an objective function by adjusting power system control variables with respect to the operating limits of the system. The non-linear constrained optimization problem can be formulated mathematically as follows [16]:

Minimize:

$$f(x,u), Objective\ function \quad (5)$$

Subject to:

$$g(x,u) = 0, Equality\ constraints \quad (6)$$

$$h(x,u) \leq 0, Inequality\ constraints \quad (7)$$

The vector of dependent variables, x consists of slack bus power P_{Gref}, load bus voltage V_{Li}, generator reactive power Q_{Gi} and transmission line loading S_{li}, which can be written as:

$$x^T = \left[P_{Gref}, V_{L1} \dots V_{L_{NL}}, Q_{G1} \dots Q_{G_{NG}}, S_l \dots S_{l\ nl}\right] \quad (8)$$

While the vector of control variables u consists of generator voltage V_{Gi}, real power output except at slack bus P_{Gi}, transformer tap setting T_i which can be stated as:

$$u^T = \left[V_{G1} \dots V_{G_{NG}}, P_{G1} \dots P_{G_{NG}}, T_1 \dots T_{NT}\right] \quad (9)$$

where *NG*, *NL*, *NT*, *nl* are the number of generators, number of load buses, number of regulating transformers and number of transmission lines respectively.

C. Objective Function

In this paper, the objective function is to optimize the total real power loss in transmission lines, which is expressed as follows [17]:

$$P_{loss} = \sum_{i=1}^{NL}\sum_{j=1}^{NL} g_{i,j}\ \{V_i^2 + V_j^2 - 2 * V_i V_j \cos(\delta_i - \delta_j)\} \quad (10)$$

The above objective function is optimized while satisfying the equality and inequality constraints, where,

P_{loss} is real power loss in system,

Vi is the voltage magnitude at the bus *i*,

NL is the total number of transmission lines,

δi is the voltage angle at the bus *i*.

$g_{i,j}$ is the conductance of line *i-j*.

D. Equality Constraints

The equality constraints are derived from the active power balance equations, which are described by power flow equation shown below [18].

$$P_i = P_{Gi} - P_{Di} = \sum_{j=1}^{NB} |V_i V_j Y_{ij}| \cos(\theta_{ij} - \delta_i + \delta_j) \quad (11)$$

where,

i is 1,2,3,……N_{bus-1}

P_{Gi} is the real power generation at bus *i*,

P_{Di} is the real power demand at bus *i*,

N_B is the total number of buses,

θ_{ij} is the angle of bus admittance element *i*, *j*.

E. Inequality Constraints

Inequality constraints include line power flow and limit of the control variables.

Continuous voltage control variables:

$$|V_{Gi}|min \leq |V_{Gi}| \leq |V_{Gi}|max, i = 1, \dots\dots, NG \quad (12)$$

Continuous active power control variables:

$$|P_{Gi}|min \leq |P_{Gi}| \leq |P_{Gi}|max, i = 1, \dots\dots, NG \quad (13)$$

Transformer tap-setting constraints:

$$|T_i|min \leq |T_i| \leq |T_i|max, i = 1, \dots\dots, NT \quad (14)$$

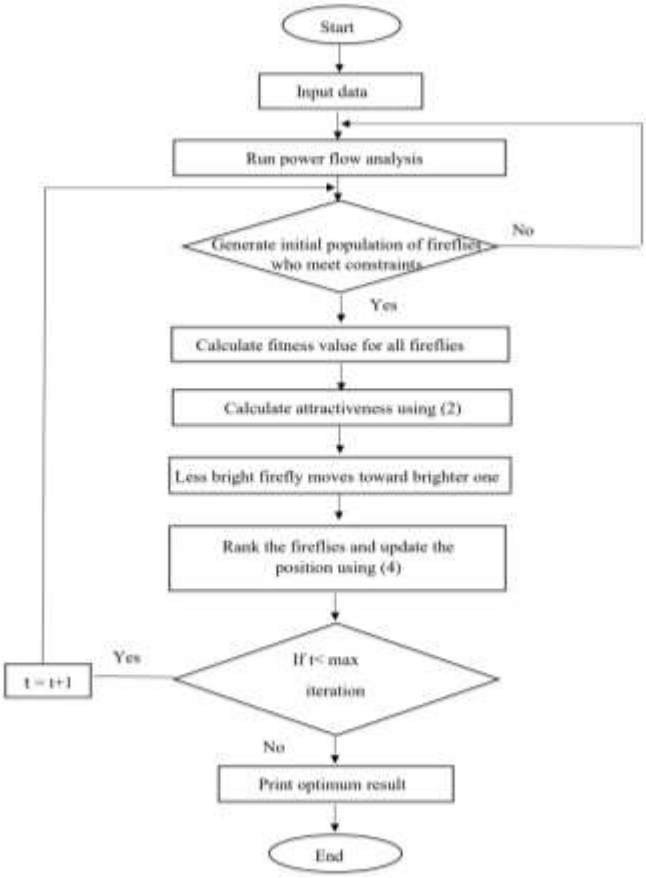


Fig. 1. Flowchart for Firefly Algorithm (FA).

The solution procedures as in Fig. 1 for the Firefly Algorithm are summarized as follows:

1) Input data such as bus data and line data.

2) Performs load flow analysis using Newton-Raphson method to calculate the amount of power loss.

3) The parameters βo, γ *and* α are set. Maximum iteration is set as $t = 100$.

4) Generate the original population of the fireflies as $\{xi\},(i=1,2,\ldots\ n)$ randomly representing a solution to an optimization problem with the objective function $f(x)$.

5) If the initial fireflies are within the limits, proceed to Step 6. Otherwise, go back to Step 2.

6) Calculate the fitness function (brightness) for each firefly.

7) The attraction between the fireflies is calculated by Equation (2).

8) The firefly *i* move towards firefly *j*, the distance is calculated by Equation (3).

9) Update the ranking of fireflies by Equation (4).

*10)*Find the current best global for fireflies.

*11)*If the t /< maximum iteration, Step 4 to 10 are repeated. Otherwise, continue to Step 12.

*12)*Print the output of active power loss.

Once the control variables are generated, the proposed algorithm uses the control variable as an initialization to run the load flow program based on the Newton-Raphson method. Objective functions are then evaluated for each population to ensure constraints are met. Then, the control variables are updated and the objective function for the updated firefly position is revalued.

The position of updated fireflies is organized and combined with the previous population and is designated as the best firefly. The best value of fireflies with minimum objective functions is noted as the best firefly position for each iteration. The iteration process will come to an end when the control variable has been set to its ideal value based on the fitness function's minimal value.

III. Result and Discussion

The proposed FA based method is analyzed based on the IEEE 14-bus and 30-bus systems. The Firefly Algorithm was developed and simulated by using MATLAB software. The control variables were adjusted within their lower and upper bounds. The ranges of control variables are shown in Table I. In this study, three types of control variables were used in the Firefly Algorithm optimization method to minimize the amount of active power loss. Next, three different cases were analyzed based on combinations of different control variable settings.

The first case study involved variable settings of active power output, while the second case study involved variable settings of active power output and voltage generator. The third case study involved all settings of the control variables of active power output, generator voltage and transformer tap.

A. Minimization of Real Power Loss

The reduction of real power loss in transmission line is the desired objective function. The proposed algorithm was simulated, and the optimal value of the real power loss was then calculated. In this study, case 3 which consists of three control variables output active power generated, generator voltage and transformer tap settings were within the operating limits. Case 3 was analyzed to obtain the optimal values of the control variables and total active power loss.

Table II shows the optimal values of the control variables for the IEEE 14-bus system. The implementation of Firefly Algorithm demonstrated a reduction in the amount of active power loss from the original case of 8.5429 MW to 6.6222 MW. The percentage of active power loss reduction for FA is 22.48%. The results showed that FA performed better compared to PSO, which only demonstrated a reduction of 15.43% from the original amount.

TABLE I. Parameters of FA

Parameters	Value
No. of fireflies (N)	40
Maximum iterations (t)	100
Absorption of light (γ)	0.01
Attractiveness (β_o)	1.0
Randomization (α)	0.95
No. of run	20

TABLE II. OPTIMAL CONTROL VARIABLES OF IEEE 14-BUS SYSTEM

Control Variables		Limit		Initial	FA	PSO
		Minimum	Maximum			
Output generator active power (MW)	P_{g1}	50	200	232	141.304	142.559
	P_{g2}	20	80	40	80	80
	P_{g3}	15	50	0	50	50
Generator voltage (p.u.)	V_1	0.95	1.05	1.06	1.05	1.0491
	V_2	0.95	1.05	1.045	1.0447	1.0237
	V_3	0.95	1.05	1.01	1.0241	1.0427
	V_6	0.95	1.05	1.07	1.0372	1.0125
	V_8	0.95	1.05	1.09	1.0486	0.9692
Transformer tap settings (p.u.)	T_{4-7}	0.95	1.05	0.9780	1.0252	0.9935
	T_{4-9}	0.95	1.05	0.9690	0.9530	0.9816
	T_{5-6}	0.95	1.05	0.9320	1.0134	1.0261
Time taken (s)		-	-	17.3945	18.2372	24.3924
Total active power loss (MW)		-	-	8.5429	6.6222	7.2247
Percentage of power loss reduction (%)		-	-	-	22.48%	15.43%

TABLE III. OPTIMAL CONTROL VARIABLES OF IEEE 30-BUS SYSTEM

Control Variables		Limit		Initial	FA	PSO
		Minimum	Maximum			
Output generator active power (MW)	P_{g1}	50	200	99.248	51.7606	51.822
	P_{g2}	20	80	80	80	80
	P_{g5}	15	50	50	50	50
	P_{g8}	10	35	20	34.849	35
	P_{g11}	10	30	20	30	30
	P_{g13}	12	40	20	40	40
Generator voltage (p.u.)	V_1	0.95	1.05	1.05	1.05	1.0065
	V_2	0.95	1.05	1.04	1.05	1.0389
	V_5	0.95	1.05	1.01	1.0359	1.0013
	V_8	0.95	1.05	1.01	1.0416	0.9912
	V_{11}	0.95	1.05	1.05	1.05	1.0300
	V_{13}	0.95	1.05	1.05	1.0479	0.9764
Transformer tap settings (p.u.)	T_{6-9}	0.95	1.05	1.078	0.9701	0.9782
	T_{6-10}	0.95	1.05	1.069	0.9653	1.0336
	T_{4-12}	0.95	1.05	1.032	1.0028	1.0318
	T_{28-27}	0.95	1.05	1.068	0.9786	0.9740
Time taken (s)		-	-	19.3942	22.5271	31.5624
Total active power loss (MW)		-	-	5.8632	3.3420	3.6658
Percentage of power loss reduction (%)		-	-	-	43.00%	37.48%

Table III shows the optimal control variables of the IEEE 30-bus system. The total active power loss for the original case for the 30-bus IEEE system is 5.8632 MW. The implementation of FA shows a lower total active power loss of 3.3420 MW, equivalent to a 43.00% reduction from the original value, while the implementation of PSO recorded a reduction of only 37.48%.

B. Bus Voltage Profile

The bus voltage profile indicates the stability of a system during operation. To validate the performance of the Firefly Algorithm in improving the voltage profile, the magnitudes of bus voltages for the IEEE 14-bus and 30-bus systems were analyzed and shown in Fig. 2 and Fig. 3 respectively. All three case studies and the base case were considered in this analysis. Fig. 2 and Fig. 3 illustrate that case 3 for both IEEE test systems resulted in better voltage profiles compared to the base case, case 1 and case 2.

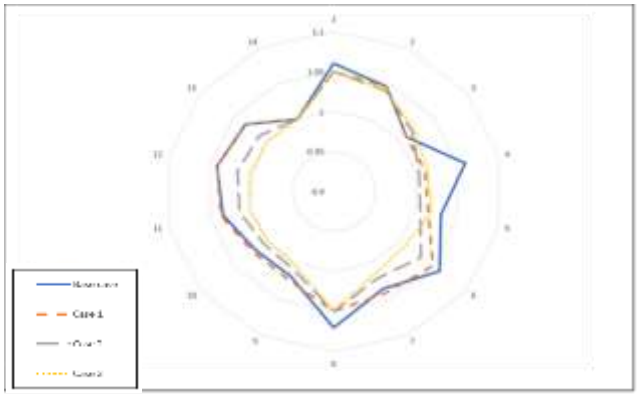


Fig. 2. Voltage Profile for IEEE-14 Bus System.

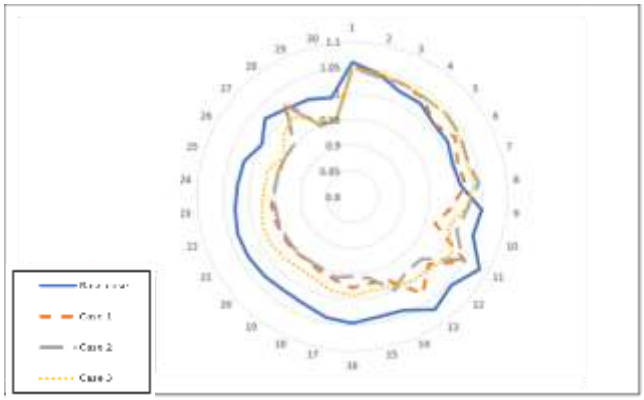


Fig. 3. Voltage Profile for IEEE-30 Bus System.

C. Total Active Power Loss

Statistical analysis was also performed to further investigate the effectiveness of the proposed FA and PSO method based on the IEEE 14-bus and 30-bus systems. The statistical result of the average, minimum, maximum objective values and standard deviations obtained from the 20 independent trials are illustrated in table. From Table IV and Table V, it is clearly shown that the statistical results of the average, minimum, maximum and standard deviation values of the active power losses obtained using FA are better than PSO.

For the IEEE 14-bus system, the standard deviation for case 1, case 2 and case 3 are 0.000, 0.004 and 0.008 respectively. The average active power loss for case 3 are 6.627 MW and 7.040 MW for FA and PSO method. For the IEEE 30-bus system, the standard deviation for case 1, case 2 and case 3 are 0.000, 0.040 and 0.045 respectively. The average active power loss of case 3 for FA and PSO are 3.395 MW and 3.675 MW respectively. The standard deviation values shown in Table 4 and Table 5 indicate that the results distribution of FA method was more concentrated in a smaller range than the PSO.

D. Consistency

Case 3 was analyzed for validating the consistency of the both the FA and PSO methods in solving the optimal power flow problem. For the IEEE 14-bus system, FA shows consistent value of the active power loss of 6.62 MW. Fig. 4 and Fig. 5 show that the result uniformity of the proposed FA is better compared with PSO method.

TABLE IV. ACTIVE POWER LOSS OF IEEE 14-BUS SYSTEM

Case	Method	Active power transmission loss (MW)			
		Lowest	**Highest**	**Average**	**Deviation**
Case 1	FA	6.928	6.928	6.928	0.000
	PSO	6.928	6.928	6.928	0.000
Case 2	FA	6.836	6.846	6.840	0.004
	PSO	6.948	7.331	7.090	0.113
Case 3	FA	6.617	6.645	6.627	0.008
	PSO	6.713	7.259	7.040	0.137

TABLE V. ACTIVE POWER LOSS OF IEEE 30-BUS SYSTEM

Case	Method	Active power transmission loss (MW)			
		Lowest	**Highest**	**Average**	**Deviation**
Case 1	FA	3.828	3.828	3.828	0.000
	PSO	3.828	3.828	3.828	0.000
Case 2	FA	3.534	3.703	3.573	0.040
	PSO	3.663	4.720	3.864	0.214
Case 3	FA	3.333	3.513	3.395	0.045
	PSO	3.529	3.840	3.675	0.083

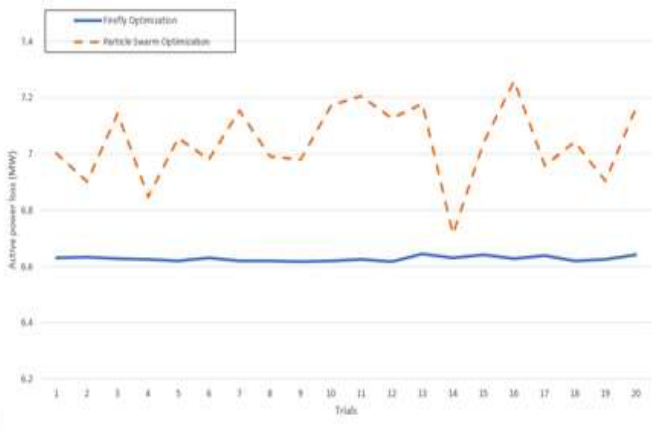


Fig. 4. Consistency Feature Case 3 for IEEE 14-Bus System.

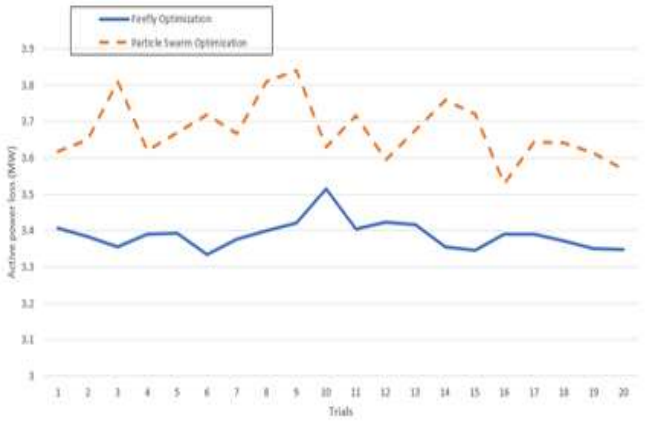


Fig. 5. Consistency Feature Case 3 for IEEE 30-Bus System.

E. Computation Time

Comparison of the computational times taken by each method to find the optimal solution for the power flow problem prove that the FA consume the least time. For IEEE 14-bus system, for case 3, Table VI shows that FA spent an average of 18.085 seconds while PSO spent 25.515 seconds to find the best solution. For the IEEE 30-bus system, Table VII shows that FA and PSO spent about 21.944 seconds and 31.562 seconds respectively. Therefore, FA method has been proven to perform better in reaching the best solutions.

TABLE VI. TIME TAKEN OF IEEE 14-BUS SYSTEM

Case	Method	Time Taken (Seconds)			
		Lowest	**Highest**	**Average**	**Deviation**
Case 1	FA	17.257	18.725	17.658	0.342
	PSO	22.348	24.756	23.531	0.634
Case 2	FA	17.113	18.954	18.051	0.662
	PSO	23.183	25.921	24.675	0.719
Case 3	FA	17.321	19.169	18.085	0.560
	PSO	23.975	27.955	25.515	1.240

TABLE VII. TIME TAKEN OF IEEE 30-BUS SYSTEM

Case	Method	Time Taken (Seconds)			
		Lowest	**Highest**	**Average**	**Deviation**
Case 1	FA	19.502	21.308	19.981	0.450
	PSO	25.439	32.634	28.587	1.635
Case 2	FA	20.954	24.950	22.329	0.792
	PSO	28.511	33.303	30.881	1.191
Case 3	FA	20.869	23.164	21.944	0.650
	PSO	30.391	36.241	31.562	1.339

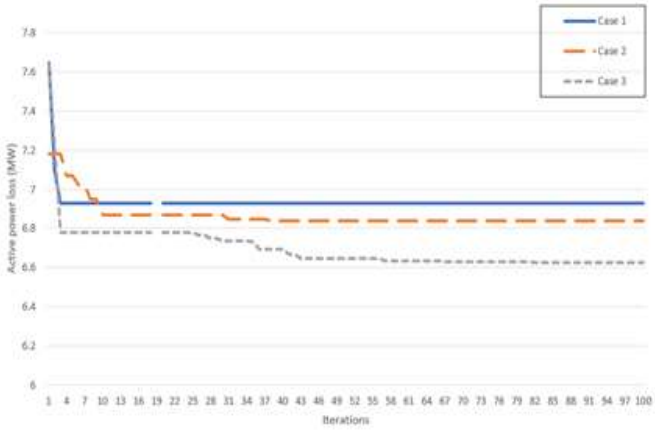


Fig. 6. Convergence Curve of FA for IEEE 14-Bus System.

F. Convergence Curve

If $n \geq m$, while m is the number of local optima in an optimization problem, the algorithm can be converged for any large number of fireflies. The starting position of fireflies is evenly dispersed around the whole searching space. When the method iterates, the fireflies converge towards all the local optimal. The global optima are found by correlating the best answers from all these optima. Therefore, FA can locate both the global and local optima in an efficient manner. Fig. 6 and Fig. 7 show the convergence curves for case 1, case 2 and case 3 of FA for the IEEE 14-bus and 30-bus systems respectively.

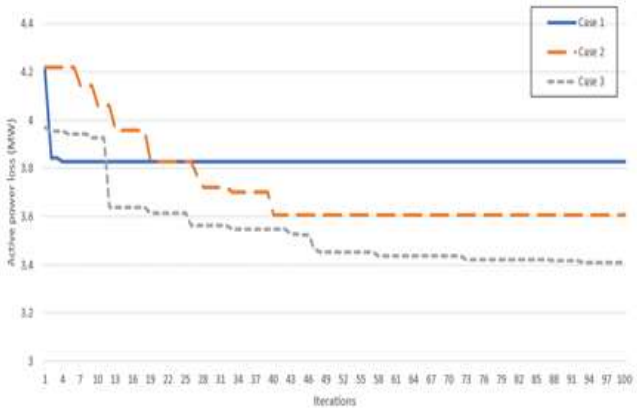


Fig. 7. Convergence Curve of FA for IEEE 30-Bus System.

IV. CONCLUSION

This study has successfully achieved the stated objective. The objective function of real power loss based on operational constraints of the IEEE 14-bus and 30-bus transmission systems has been successfully formulated. The simulated results based on the test systems for optimal power flow control indicate that Firefly Algorithm can find a well-distributed optimal solution. In addition, the implementation of the Firefly Algorithm optimization has successfully produced a minimum amount of active power loss with faster computational time. The active power loss obtained for the IEEE 14-bus and 30-bus system is 6.6222 MW and 3.3420 MW respectively. The results have shown good performance based on the consistency and computation time compared with the Particle Swarm Optimization. In conclusion, this study has proven that Firefly Algorithm is an ideal optimization method to reduce real power loss, subjected to the system operating constraints and control variable settings. Future research can be improved by simulating the OPF control using multi-objectives optimization, rather than a single-objective optimization as discussed in this paper.

ACKNOWLEDGMENT

The authors thank the Ministry of Higher Education Malaysia (MOHE) and Universiti Kebangsaan Malaysia (UKM) for supporting this work through FRGS research grant (FRGS/1/2018/TK04/UKM/02/12).

REFERENCES

[1] Balachennaiah, P., Suryakalavathi, M., & Nagendra, P., "Firefly algorithm based solution to minimize the real power loss in a power system," Ain Shams Engineering Journal, vol.9, no. 1, pp. 89-100, 2018.

[2] Vijayvargia, A., Jain, S., Meena, S., Gupta, V., & Lalwanib, M., "Comparison between different load flow methodologies by analyzing various bus systems," International Journal of Electrical Engineering, vol. 9, no. 2, pp. 127-138, 2016.

[3] Lamichhane, B., & Luintel, M. C., "Use of Genetic Algorithm on optimal power flow: An illustration of transmission line loss minimization," Kathford Journal of Engineering and Management, vol. 1, no. 1, pp. 1-4, 2018.

[4] Trivedi, I. N., Jangir, P., Jangir, N., Parmar, S. A., Bhoye, M., & Kumar, A. (2016, March). "Voltage stability enhancement and voltage deviation minimization using multi-verse optimizer algorithm". In 2016 International conference on circuit, power and computing technologies (ICCPCT), pp. 1-5. IEEE.

[5] Trivedi, I. N., Jangir, P., & Parmar, S. A., "Optimal power flow with enhancement of voltage stability and reduction of power loss using ant-lion optimizer," Cogent Engineering, vol. 3, no. 1, 1208942, 2016.

[6] Khan, A., Hizam, H., Abdul Wahab, N. I., & Lutfi Othman, M., "Optimal power flow using hybrid firefly and particle swarm optimization algorithm," Plos One, vol. 15, no. 8, e0235668, 2020.

[7] Abugri, J. B., & Karam, M., "Particle swarm optimization for the minimization of power losses in distribution networks". In 2015 12th International Conference on Information Technology-New Generations, pp. 73-78. IEEE.

[8] Lenin, K., "Active power loass reduction by Firefly Algorithm," *International Journal of Research-GRANTHAALAYAH*, vol. 6, no. 3, pp. 155-165, 2018.

[9] D Patel, T., & Acharya, A. G., "Minimize power loss using Particle Swarm Optimization technique," International Journal of Electrical Engineering and Technology, vol. 10, no. 2, 2019.

[10] Moustafa, F. S., El-Rafei, A., Badra, N. M., & Abdelaziz, A. Y., "Application and performance comparison of variants of the firefly algorithm to the economic load dispatch problem". In 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), pp. 147-151. IEEE.

[11] Liaquat, S., Fakhar, M. S., Kashif, S. A. R., Rasool, A., Saleem, O., Zia, M. F., & Padmanaban, S., "Application of dynamically search space squeezed modified firefly algorithm to a novel short term economic dispatch of multi-generation systems," IEEE Access, vol. 9, pp. 1918-1939, 2020.

[12] Sendra, S., Parra, L., Lloret, J., & Khan, S., "Systems and algorithms for wireless sensor networks based on animal and natural behavior," International Journal of Distributed Sensor Networks, vol. 11, no. 3, 625972, 2015.

[13] Govindaraj, T., & Tamildurai, V., "Firefly algorithm for optimal power flow considering control variables," International Journal of Innovative Research In Electrical, Electronics, Instrumentation And Control Engineering, vol. 2, no. 2, pp. 1116-1121.

[14] Deenadhayalan, H., "Real power loss minimization using firefly algorithm," International Journal of Advanced Information and Communication Technology, vol. 1, no. 8, pp. 677-682, 2014.

[15] Zhang, L., Liu, L., Yang, X. S., & Dai, Y., "A novel hybrid firefly algorithm for global optimization," PloS one, vol. 11, no. 9, e0163230, 2016.

[16] A Abbasi, M., Abbasi, E., & Mohammadi-Ivatloo, B., "Single and multi-objective optimal power flow using a new differential-based harmony search algorithm," Journal of Ambient Intelligence and Humanized Computing, vol. 12, pp. 851-871, 2021.

[17] Li, S., Gong, W., Wang, L., Yan, X., & Hu, C., "Optimal power flow by means of improved adaptive differential evolution," Energy, vol. 198, 117314, 2020.

[18] Warid, W., "Optimal power flow using the AMTPG-Jaya algorithm," Applied Soft Computing, vol. 91, 106252, 2020.

Distance Education during the COVID-19 Pandemic: The Impact of Online Gaming Addiction on University Students' Performance

Mahmoud Abou Naaj, Mirna Nachouki
Department of Information Technology
Ajman University, Ajman
UAE

***Abstract*—The COVID-19 pandemic has forced most universities worldwide to convert to distance education to ensure the educational process remains uninterrupted. The COVID-19 pandemic-related confinement orders have led students to be more engaged with online games. However, for a minority of students, excessive playing can become problematic and addictive. Few studies investigated the long-term effect of COVID-19 on game addiction among university students. The present study investigates the changes in online game addiction rates between May 2021 and May 2020 and aims at determining the impact of playing online games on students' academic performance. It also examines the demographic factors associated with video game addiction. A sample (n= 418) of students from one private university in UAE was randomly selected, and data were analyzed. The study has determined a reduction in online game addiction levels in the second year of pandemic compared with the first year. Gender and academic level were considered the most predominant features expressively related to online games addiction. It has also been found that digital game addiction is positively associated with academic performance.**

Keywords*—Distance education; COVID-19 pandemic; game addiction; students' performance

I. Introduction

In modern society, most of the information is collected from the internet, thus using electronic devices has become an essential means in all living areas. In particular, United Arab Emirates (UAE) emphasized the importance of building high-speed information infrastructures and informatization education at the national level. The statistic demonstrates that the UAE has the highest internet diffusion rate as of January 2021. It was found that 99 percent of the UAE population has access to the internet [1]. Playing games is one of the most common leisure activities for young people. It has been continuously increasing with the development of the internet and smart devices.

The emergence of the deadly COVID-19 pandemic at the end of 2019, which gradually swept all over the globe, brought instant and dramatic changes. People, who were required to stay at home, increased their consumption of all kinds of digital entertainment, especially online games [2, 3]. For instance, 75% increase had been recorded in online gaming activity since COVID-19 pandemic stay-at-home instructions in USA [4], with more than 20 million concurrent active users [5]. In addition, online gaming has been perceived as amplified due to public health efforts to encourage social distancing [6].

To control and decrease virus transmission, the governments in many countries worldwide, including the United Arab Emirates, implemented mandatory closure of educational institutions and all their face-to-face teaching and learning activities. This shift from traditional face-to-face learning to distance learning instruction greatly impacted everyone involved, particularly students. They registered for face-to-face education but suddenly switched from the conventional learning style they preferred and were accustomed to.

The lockdown caused by the COVID-19 pandemic may have further increased their use of online video games and, therefore, the risk of video gaming addiction. While moderate use of online gaming could lead to wide-ranging benefits [7], extreme usage of this activity experimented with various social problems and addictions [8]. The pandemic has contributed to increasing video game addiction in general and among students [9].

However, we do not currently have a thorough knowledge of the effect of the COVID-19 pandemic on students. Therefore, the current study will analyze video gaming addiction among university students during the lockdown, starting in March 2020.

Furthermore, few studies have been conducted to examine the relationship between playing video games and academic performance, especially in the United Arab Emirates. Most video game studies focus on the behavioral effects of video games. This issue makes it essential to determine whether the increased time spent playing online games during pandemics influenced students' academic performance. The study also identifies the main demographic factors (age, gender, academic level) related to online video game addiction.

Educational institutions, instructors can benefit from the findings of this research to shift their pedagogical approach from lecture-centered to student-centered and integrate the use of games as innovative learning technologies into their curricula.

II. Literature Review

The Covid-19 pandemic has affected teaching and learning at almost all higher education institutions worldwide. Most universities have faced sudden pressure to change from face-to-face courses to digitally enhanced teaching for distance learning. Two-thirds had reported replacing face-to-face classroom teaching with distance teaching and learning [10]. The pandemic also increased the risk of video game addiction among students [9].

Most of the students play video games, and it is known that video games are a common form of entertainment [11]. In addition, games are often seen as an important part of new technologies, and they have become an essential part of many students' lives [12]. However, students' online video game usage has generated significant concern due to its possible adverse effects on their health, socialization, and academic performance. Regarding this last aspect, some studies point out that online video games negatively affect academic performance, while others emphasize their positive effects.

Games can enhance logical thinking, analytical skills, social skills, visual abilities, collaboration, movement, and computing [13]. Psychomotor processes are influenced by computer and video games, and tension levels are reduced. In addition, playing games improve critical reasoning, analytical capabilities, movement, cognitive skills, perceptual ability, teamwork, and programming [14]. A study by [15] showed a positive relationship between the number of hours students play video games and their GPA, which means that students who take time playing video games can have better academic performance. Also, the author in [13] states that students' awareness and consciousness can be improved by playing video games. Online video games play an important role in increasing students' intelligence quotients (IQ). When students play games in the classroom, they form teams and devise their tactics for winning. They pay attention to and analyze the key points, aiming to minimize future challenges by looking more seriously [16]. The study of [14] found that 72% of participants believed that homework tasks involving computer or video games outside of the classroom could be a benefit to student education. In addition, about 77% accepted that video games should be used to incorporate and teach science, technology, and mathematics principles.

As stated by [13], video games can enhance logical thinking, analytical skills, social skills, visual abilities, collaboration, movement, and computing. Therefore, there can be a positive impact on student's GPA/academic achievements caused by video games. However, some studies, such as [12], states that video games cause negatively on students' academic performance due to addiction and distraction. Therefore, there is mixed information about the relation between video games and student's academic performance.

Students who did play video games had meaningfully lower GPAs than those who stated that they did not play video games [17]. This conclusion is similar to the study [18], which noticed a decrease in GPA and SAT scores for video game students. Therefore, it is more reliable to use GPA as it represents a continuous assessment of academic performance. This conclusion also matches the findings of [19], who experimented with decreases in academic performance for students who played video games.

Although there is no specific definition for the term video game addiction, it was found that students addicted to playing games had lower academic performance. Moreover, time spent playing online games was found as a negative indicator of academic performance for undergraduate students [20].

A study performed by [21] had used open-ended questions to encourage students to report their moods while playing video games. As a result, various negative repercussions indirectly related to academic performance associated with addiction to video gaming were reported, such as skipping classes, missing homework deadlines, etc. Researchers also found that these repercussions directly involved gender, an important demographic factor of video game addiction, as male students tend to play more often and lean towards losing time while playing. This statement was also confirmed by [18], who found that the academic performance of male students was more affected as they spent more time playing video games.

An experiment proved that academic performance improved after the students completely decreased their usage time of all technology, including video games, with a maximum of 30 minutes allowed per day. On the other hand, [22] found that playing video games for more than five hours per session has a negative impact on academic performance. They also found that these excessive hours spent on gaming promoted violence among students and prevented them from having normal social interaction and extra-curricular activities.

III. Purpose of the Study

The purpose of the study is:

- Check whether a change in students' online video game addiction is noticeable between May 2020 and May 2021,
- Observe the effects of some demographic factors (age, gender, academic level) on online video game addiction,
- Explore the impact of online video gaming on student's academic performance.

IV. Methodology

A. Participants

A total of 418 participants from one private university in the UAE were randomly selected to participate in this cross-sectional study. A survey questionnaire was applied to collect data from students in May 2020 and May 2021. In all, 101(24%) students had completed the survey form in May 2021and 317 (76%) students completed the survey in May 2020. Fig. 1 represents the demographic data of the students who participated in the survey. Of those who completed the survey form, 187 (45%) were males, whereas 231 (55%) were females. A total of 138 (33%) of the participant were aged less than or equal 20, 263 (63%) aged between 21 and 25, 14 (3%) aged between 26 and 29, and 3 (1%) aged 30 or above. A total 290 (69%) of the participant has completed less than 90 credit hours, and 128 (31%) has completed 90 credit or above. 5 (1%)

participants have CGPA less than 2, 80 (19%) have a CGPA between 2 and 2.49, 105 (25%) have a CGPA between 2.5 and 2.99, 129 (31%), or have a CGPA between 3 and 3.59 while 98 (24%) has a CGPA of 3.6 or above.

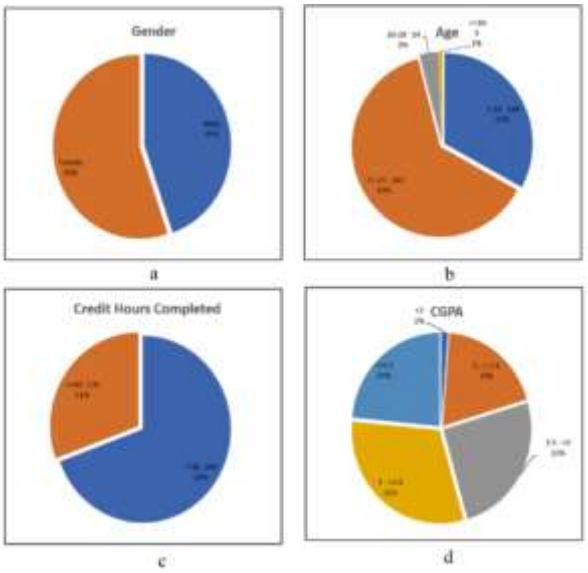


Fig. 1. Demographic Data of the Students who participated in the Survey.

B. Instrument

The research instrument consisted of a self-report questionnaire with a three-part structure designed to collect student's responses. The first section collected demographical/personal data based on five items related to participants' profiles, including age, gender, CGPA, major, and the number of credit hours completed. The second part consists of four questions. The first question was, "On average, how many hours do you spend playing video games daily?". The second question was, "Does gaming overall give you any real fulfillment in life?". The third question was, "If it is needed, could you quit playing video games easily?". While the fourth question was, "Did the average daily time you spend playing video games increase during the COVID-19 Pandemic?". Part 3 consisted of 20 items. The statement preceded them: "How often in the last six months…? All items were scored on the 5-point Likert scale, ranging from 1(never), 2(rarely), 3 (sometimes), 4 (often), and 5 (very often). These items were mainly based on a review of the literature.

C. Reliability

The researchers performed a reliability analysis using Cronbach's alpha to determine the internal questionnaire consistency once the data collection phase was completed. The satisfaction scale's alpha reliability coefficient was 0.956, which indicated that the instrument was highly reliable.

D. Research Hypotheses

The research hypotheses are listed below:

H1: There is no significant change in students' online game addiction between May 2020 and May 2021.

H2: There is no significant relationship between gender and online game addiction.

H3: There is no significant relationship between academic level and online game addiction.

H4: There is no significant relationship between age and online game addiction.

H5: Online game addiction has no significant impact on academic performance.

E. Data Analysis

Data analysis was conducted using the IBM SPSS Statistics version 27. An independent t-test was used to examine the relationships between video game addition, gender, academic level. In addition, between-participants one-way ANOVA was used to assess the effects of variables (age) on video game addiction and time spent gaming. Also, the one-way ANOVAs test was used to determine if video game addiction and time spent gaming affect academic performance. All statistical analysis was two-tailed with a .05 level of significance. Finally, linear regression was used to find out whether there is a correlation between CGPA and Addiction Hours.

V. Result and Discussion

Based on Fig. 2a, the question "On average, how many hours do you spend playing video games daily?". 15% of the participants do not play video games, followed by 50% of the participants play on average 2 hours daily. 27% of the participants play on average 5 hours daily, while 6% play on average 8 hours daily, and finally 2% of the participants play on average 10 hours every day.

Based on Fig. 2b, the question "Does gaming overall give you any real fulfillment in life?". 48% of the participants said yes, 14% were neutral, while 38 said no. Then, based on Fig. 2c, the question "If it was needed, could you quit playing video games easily?". 58% of the participants said yes, 8% were neutral, while 34% said no. Finally, based on Fig. 2d., the question "Did the average daily time you spend playing video games increase during COVID-19 Pandemic?". 146 (35%) of the participants said yes, 66 (16%) were neutral, while 206 (49%) said no.

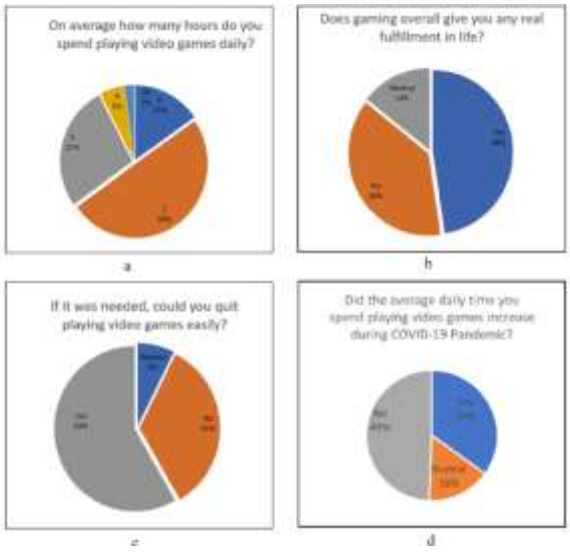


Fig. 2. Students responses with regards to their Gaming Behavior.

A. Comparison between May 2020 and May 2021 Survey Analysis

H1: There is no significant change in students' online game addiction between May 2020 and May 2021.

An independent t-test was used to test hypothesis H1. The results of the testing are reported in Table 2. The test aimed to determine if students surveyed in May 2020 have the same online games addiction level as those surveyed in May 2021. The result shows that there is a statistically important difference students surveyed in May 2020 online games addiction level (Mean= 2.45, Standard Deviation = 1.215) and on May 2021 online games addiction level (Mean= 2.044, Standard Deviation= 1.188; t(416) = 3.286, p= 0.010). Consequently, hypothesis H1 was rejected.

In conclusion, this result reveals an important difference between May 2020 students' online games addiction and May 2021 students' online games addiction. Specifically, in May 2021, students are less addicted to online games.

The average addiction hours for students surveyed in May 2020 was 3 hours, while 2.84 hours for those surveyed in May 2021.

B. Gender

H2: There is no significant relationship between gender and online games addiction.

An independent t-test was used to test hypothesis H2. The results of the testing are reported in Table 2. The test aimed to determine if female and male students have the same online games addiction level. The result shows that there is a statistically important difference in male students' video game addiction (Mean= 2.58, Standard Deviation = 1.159) and female students' video game addiction (Mean= 2.198, Standard Deviation= 1.198; t(317) = 3.436, p= 0.012). Consequently, hypothesis H2 was rejected.

In conclusion, this result reveals an important difference in students' online game addiction between male and female students. Specifically, female students are less addicted to online games than their male counterparts.

The researchers also conducted another independent t-test shown in Table 1. The test aimed to determine if female and male students spend the same time playing video games. The result shows that there is a statistically important difference in male students' video game addiction (Mean= 3.55, Standard Deviation = 2.253) and female students' video game addiction (Mean= 2.48, Standard Deviation= 2.234; t(416) = 4.139, p= 0.000).

In conclusion, male students spent more time playing online games than their female counterparts.

C. Academic Level

H3: There is no significant relationship between academic level and online games addiction.

An independent t-test was used to test hypothesis H3. The results of the testing are reported in Table 2. The test aimed to determine if senior students (completed 90 credit hours or more) and junior students (completed less than 90 credits hours) have the same video game addiction level. The result shows that there is a statistically important difference between senior online games add (Mean= 2.08, Standard Deviation = 1.130) and junior students' video game addiction (Mean= 2.50, Standard Deviation= 18130; t(416) = 4.435, p=.001). Consequently, hypothesis H3 was rejected.

In conclusion, this result means a significant difference in students' video game addiction levels between senior and junior students.

TABLE I. INDEPENDENT SAMPLE T-TEST ANALYSIS RELATED TO STUDENTS' TIME SPENT ON VIDEO GAMES WITH REGARDS TO THEIR GENDER

Category		**Gender**	
St. Type		**Male**	**Female**
Nº of Participants		187	231
Mean		3.55	2.48
Standard Deviation		2.253	2.34
t		4.852	
df		416	
p		0.000	
95% Confidence Interval of the Difference	**Lower**	0.636670	
	Upper	1.503895	

TABLE II. INDEPENDENT SAMPLE TEST ANALYSIS RELATED TO STUDENTS' ONLINE GAME ADDICTION

								95% Confidence Interval of the Difference	
Category	**Students' type**	**Number of Participants**	**Mean**	**SD**	**t**	**df**	**p**	**Lower**	**Upper**
Gender	Male	187	2.58	1.159	3.436	416	0.0012	0.162	0.602
	Female	231	2.198	1.11916					
Students	May 2020	317	2.45	1.215	3.286	416	0.010	0.172	0.684
	May 2021	101	2.044	1.1188					
Level	Junior	291	2.50	1.143	3.435	416	0.001	0.179	0.655
	Senior	127	2.08	1.130					

D. Age

Students were divided into four groups. Groups 1 for students who are less than or equal to 20 years old. Group 2 for students between 21 and 25 years old, group 3 for students between 26 and 29 years old, and group 4 for students who are 30 years or older. One-way ANOVA test analysis determined whether video game addiction and time spent playing video games differ significantly based on age. Table 3 indicates that university students' level of video game addiction does not differ according to age (p = .196). In addition, table 4 indicates that the time spent playing video games does not differ according to age (p = .871). Consequently, hypothesis H4 was accepted.

E. Academic Performance

In order to check whether video game addiction affects academic performance, the participants were divided into five groups. Groups 1 includes students with CGPA less than 2, group 2 for students with CGPA between 2 and 2.49, group 3 for students with CGPA between 2.5 and 2.99, group 4 for students with CGPA between 3.0 and 3.59, and finally group 5 for student CGPA between 3.6 and 4.0. One-way ANOVA test analysis was used to determine whether video game addiction and time spent playing video games are associated with academic performance. Table 5 indicates that university students' level of video game addiction is significantly different between CGPA groups (p = .000). Fig. 3 shows that the mean for game addiction decreases as CGPA increase. Table 6 indicates that the time spent playing video games is significantly different between CGPA groups (p = .000). Fig. 4 shows that the mean time spent playing video games decreases as CGPA increase.

Hypothesis H5 was rejected in both tests. This confirms that video game addiction has a significant impact on academic performance. Furthermore, students addicted to video games have worse performance than those who are less addicted.

TABLE III. ONE-WAY ANOVA RESULTS FOR THE LEVEL OF VIDEO GAME ADDICTION OF UNIVERSITY STUDENTS AGE

ANOVA					
AddicHRs					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	26.724	3	8.908	1.572	.196
Within Groups	1813.227	320	5.666		
Total	1839.951	323			

TABLE IV. ONE-WAY ANOVA RESULTS FOR TIME SPENT IN VIDEO GAME ADDICTION OF UNIVERSITY STUDENTS

ANOVA					
MeanofResp					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.534	3	.178	.237	.871
Within Groups	240.763	320	.752		
Total	241.297	323			

TABLE V. ONE-WAY ANOVA RESULTS FOR THE LEVEL OF VIDEO GAME ADDICTION OF UNIVERSITY STUDENTS CGPA

ANOVA					
MeanofResp					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	56.936	4	14.234	23.954	.000
Within Groups	244.821	412	.594		
Total	301.757	416			

TABLE VI. ONE-WAY ANOVA RESULTS FOR TIME SPENT IN VIDEO GAME ADDICTION OF UNIVERSITY STUDENTS WITH REGARDS TO CGPA

ANOVA					
AddicHRs					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	98.762	4	24.691	4.876	.001
Within Groups	2086.077	412	5.063		
Total	2184.839	416			

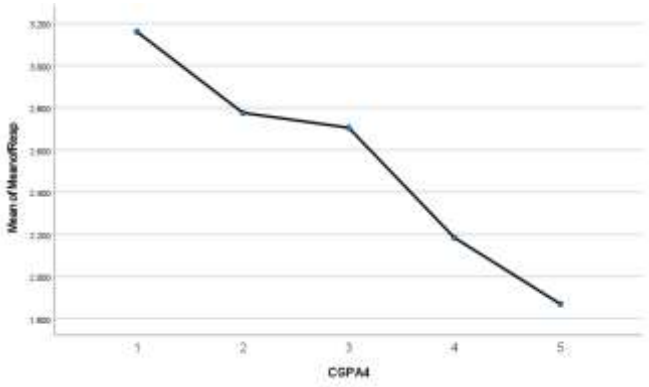


Fig. 3. Mean for Game Addiction with regards to CGPA.

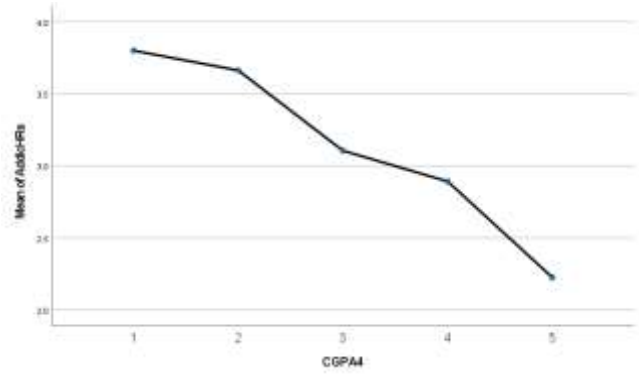


Fig. 4. Mean for Time Spent on Game Addiction with regards to CGPA.

The researchers have calculated the Pearson correlation and Significance One-Tailed Test of the CGPA and the mean of students' responses to the 20 survey questions. It was found that r (Pearson Correlation) is -.416. These results show a negative correlation between the CGPA and the mean of students' responses to the 20 survey questions at p-value= .000 < 0.05 (Table 7 and Table 8).

TABLE VII. PEARSON CORRELATION OF THE CGPA AND THE MEAN OF STUDENTS' RESPONSES

Correlations			
		AddicHRs	CGPA4
Pearson Correlation	AddicHRs	1.000	-.186
	CGPA4	-.186	1.000
Sig. (1-tailed)	AddicHRs	.	.000
	CGPA4	.000	.
N	AddicHRs	417	417
	CGPA4	417	417

TABLE VIII. SIGNIFICANCE ONE-TAILED TEST OF THE CGPA AND THE MEAN OF STUDENTS' RESPONSES

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4.391	.382		11.481	.000	3.639	5.143
	CGPA4	-.395	.103	-.186	-3.848	.000	-.597	-.193
[a] *Dependent Variable: AddicHRs*								

TABLE IX. PEARSON CORRELATION OF THE CGPA AND THE NUMBER OF ADDICTION HOURS

Correlations			
		VAR00027	CGPA4
Pearson Correlation	VAR00027	1.000	-.416
	CGPA4	-.416	1.000
Sig. (1-tailed)	VAR00027	.	.000
	CGPA4	.000	.
N	VAR00027	417	417
	CGPA4	417	417

TABLE X. SIGNIFICANCE ONE-TAILED TEST OF THE CGPA AND THE NUMBER OF ADDICTION HOURS

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	3.550	.130		27.224	.000	3.294	3.807
	CGPA4	-.326	.035	-.416	-9.323	.000	-.395	-.258
[a] *Dependent Variable: VAR00027*								

The researchers have calculated the Pearson correlation and Significance One-Tailed Test of the CGPA and the number of addiction hours. It was found that r (Pearson Correlation) is -.186. These results show a negative correlation between CGPA and the number of addiction hours at p-value= .000 < 0.05 (Table 9 and Table 10).

VI. CONCLUSION

Covid-19 pandemic changed the lifestyles of many students. At the same time, video games are thought to have a negative impact on students; it is believed that playing digital games has become a global relaxation activity among students and people in general. However, published research suggests that an important number of university students tend to play games excessively, which could be damaging their physical and mental well-being [1, 23]. Therefore, the World Health Organization [24] has recently included gaming disorder as a diagnosable mental disorder in the International Classification of Diseases (ICD) revision.

One of the objectives of the current study is to explore the deviations in online gaming behavior among university students in the context of COVID-19 between May 2020 and May 2021. This study revealed an important difference between May 2020 students' online games addiction and May 2021 students' online games addiction. Specifically, in May 2021, students are less addicted to online games. This is quite normal as students had increased their gaming behavior during the lockdown period following COVID-19 pandemic. The average addiction hours for students who were surveyed in May 2020 decreased from 3 hours in May 2020 to 2.84 in May 2021.

The present study had also found that video game addiction levels differ significantly according to gender. Female students had lower levels of video game addiction than males. In addition, female students play online games less than male students. The result of the t-test had also shown gender differences in time spent playing video games. Female students spend less time playing video games. The findings of this study are consistent with that of [17, 25, 26, 27, 28, 29].

The result of the one-way ANOVAs test had shown no significant relationship between age and video game addiction. This is logical because the participant and university student's age, in general, falls within a close age group of 18-25. However, the t-test had also shown an important correlation between the academic level and video game addiction. First to third-year students are more addicted to video games than fourth- and fifth-year students. This is due to the higher maturity of the fourth- and fifth-year students.

The result of the study showed that video game addiction has a significant impact on academic performance. This finding is consistent with that of [17, 30, 31], who saw a lower academic performance in students who engaged in playing video games. In addition, a negative correlation between students' CGPA and the number of hours they spend daily playing online games was found. As the CGPA increases, the number of addiction hours decreases, supporting the theory that academic performance can be negatively affected if the student is addicted to online video games.

Higher education institutions should take necessary actions to ensure that online video games cannot have a negative impact on educational achievement by:

1) Organize regular workshops and seminars to increase the population's awareness about the risk of video games' addiction and how to manage its use.

2) Integrating video games that promote critical thinking in the learning process so that this gaming addiction will turn up to be a catalyst for learning and improve students' academic performance, instead of decreasing academic CGPA.

Higher education institutions are constantly evolving by shifting their pedagogical approach from lecture-centered to student-centered and integrating the use of games as innovative learning technologies into their curricula.

VII. Limitation and Future Work

A. Limitation

The current study could serve as a driving force for future research focusing on online game addictions. However, one should interpret the data carefully as the sample was limited to one university in the United Arab Emirates.

B. Future Work

1) Include students from other universities would provide better opportunities to understand the impact of game addiction on academic performance among university students in general.

2) Track students' behavior during their whole academic journey would give a more accurate account of the CGPA and video gaming and produce more precise data to assess students' gaming habits and CGPA scores for four or five years.

3) The research could also be extended to include online gaming disorder, depressive, and anxiety symptoms among university students.

4) Finally, the research could be extended to cover the impact of online gaming on time management and the social life of university students.

References

[1] Johnson J., (2021), Leading online markets based on penetration rate 2021, https://www.statista.com/statistics/227082/countries-with-the-highest-internet-penetration-rate/.

[2] Javed, J. (2020). eSports and gaming industry thriving as video games provide escape from reality during coronavirus pandemic.

[3] Perez, M. (2020). Video games are being played at record levels as the coronavirus keeps people indoors.

[4] Pantling, A. (2020). Gaming usage up 75 percent amid coronavirus outbreak, Verizon reports. Retrieved from https://www.hollywoodreporter.com/news/gaming-usage-up-75-percentcoronavirus-outbreak-verizon-reports-1285140.

[5] Stephen, B. (2020). This is Twitch's moment [internet]. Retrieved from https://www.theverge.com/2020/3/18/21185114/twitchyoutube-

[6] Abel, T., & McQueen, D. (2020). The COVID-19 pandemic calls for spatial distancing and social closeness: not for social distancing!.

[7] Mishra, J., Anguera, J. A., & Gazzaley, A. (2016). Video games for neuro-cognitive optimization. *Neuron*, *90*(2), 214-218.

[8] Carbonell, X., Chamarro, A., Oberst, U., Rodrigo, B., & Prades, M. (2018). Problematic use of the internet and smartphones in university students: 2006–2017. *International journal of environmental research and public health*, *15*(3), 475.

[9] De Pasquale, C., Chiappedi, M., Sciacca, F., Martinelli, V., & Hichy, Z. (2021). Online videogames use and anxiety in children during the COVID-19 pandemic. *Children*, *8*(3), 205.

[10] Marinoni, G., van't Land, H., & Jensen, T. (2020). The impact of Covid-19 on higher education around the world. IAU Global Survey Report. International Association of Universities (IAU). https://www. youtube.com/channel/UCT5nt5FGVklxrtUHinF_LFA.

[11] Rüth, M., & Kaspar, K. (2020). Commercial video games in school teaching: two mixed methods case studies on students' reflection processes. *Frontiers in psychology*, *11*, 3802.

[12] Almalki, A. A., & Aldajani, H. M. (2021). Impact Of Playing Video Games On The Social Behavior And Academic Performance Of Medical Student In Taif City. *International Journal of Progressive Sciences and Technologies*, *24*(1), 572-585.

[13] Miller, C. T. (Ed.). (2008). *Games: Purpose and potential in education*. Springer Science & Business Media.

[14] Clark, A., & Ernst, J. (2009). Gaming research for technology education. *Journal of STEM Education*, *10*(1).

[15] Alqurashi, M., Almoslamani, Y., & Alqahtani, A. (2016). Middle School Students' Digital Game Eperiences in the City of Makkah in Saudi Arabia. *IJAEDU-International E-Journal of Advances in Education*, *2*(4), 167-175.

[16] Ashton, D. (2011). Playstations and workstations: identifying and negotiating digital games work. *Information Technology & People*.

[17] Wright, J. (2011). The effects of video game play on academic performance. *Modern psychological studies*, *17*(1), 6.

[18] Anand, V. (2007). A study of time management: The correlation between video game usage and academic performance markers. *CyberPsychology & Behavior*, *10*(4), 552-559.

[19] Wack, E., & Tantleff-Dunn, S. (2009). Relationships between electronic game play, obesity, and psychosocial functioning in young men. *CyberPsychology and Behavior*, *12(2)*, 241-244.

[20] Jackson, L., Zhao, Y., Kolenic III, A., Fitzgerald, H., Harold, R., & Von Eye, A. (2008). Race, gender, and information technology use: The new digital divide. *CyberPsychology and Behavior*, *11(4)*, 437-442.

[21] Wood, R., Griffiths, M., & Parke, A. (2007). Experiences of time loss among videogame players: An empirical study. *CyberPsychology and Behavior, 10(1),* 38-44.

[22] Jaruratanasirikul, S., Wongwaitaweewong, K., & Sangsupawanich, P. (2009). Electronic game play and school performance of adolescents in southern Thailand. *CyberPsychology and Behavior, 12(5),* 509-512.

[23] Stevanović, D., Đoric, A., Balhara, Y., Ćirović, N., Arya, S., Ransing, R., ... & Knez, R. (2020). Assessing the symptoms of Internet Gaming Disorder among college/university students: An international validation study of a self-report. *Psihologija*, *53*(1), 43-63.

[24] World Health Organization. (2020). #HealthyAtHome – Mental Health. Retrieved from https://www.who.int/news-room/campaigns/connecting-the-world-to-combat-coronavirus/healthyathome/healthyathome—mental-health.

[25] Gómez-Galán, J., Lázaro-Pérez, C., & Martínez-López, J. ˙(2021). Exploratory Study on Video Game Addiction of College Students in a Pandemic Scenario. *Journal of New Approaches in Educational Research, 10(2),* 330-346. doi: 10.7821/naer.2021.7.750.

[26] Miezah D, Batchelor J, Megreya AM, Richard Y, Moustafa AA. Video/Computer Game Addiction among University Students in Ghana: Prevalence, Correlates and Effects of Some Demographic Factors. Psychiatry and Clinical Psychopharmacology 2020;30(1):17-23, DOI:10.5455/PCP.20200320092210.

[27] Bekir S., Çelik E., An Investigation of Online Gaming Addiction Level Among University Students in Terms of Emotional Schemas, Agentic Personality, and Various Variables, Malaysian Online Journal of Educational Technology 2020 ,Volume 8 (3).

[28] Müezzin E., An Investigation of High School Students' Online Game Addiction With Respect To Gender, TOJET: The Turkish Online Journal of Educational Technology – July 2015, Special Issue 1 for IETC 2015, pp. 55 – 60.

[29] Dennis O. Dumrique and Jennifer G. Castillo, (2018), "Online Gaming: Impact on the Academic Performance and Social Behavior of the Students in Polytechnic University of the Philippines Laboratory High School" in 4th International Research Conference on Higher Education, KnE Social Sciences, pages 1205–1210. DOI 10.18502/kss.v3i6.2447.

[30] Gómez-Gonzalvo F., Cardenal Herrera, José Devís-Devís, Dr. Pere Molina-Alventosa, Video game usage time in adolescents' academic performance, Comunicar, Vol. XXVIII,No 65 2020 | Media Education Research Journal | ISSN: 1134-3478; e-ISSN: 1988-3478 www.comunicarjournal.com.

[31] Almalki A., Aldajani, H., 2020, Impact of Playing Video Games On The Social Behavior And Academic Performance Of Medical Student In Taif City, International Journals of Sciences and High Technologies http://ijpsat.ijsht-journals.org Vol. 24 No. 1 December 2020, pp. 572-585.

Enhancing Business Process Modeling with Context and Ontology

Jamal EL BOUROUMI, Hatim GUERMAH, Mahmoud NASSAR
IMS Team, ADMIR Laboratory
Rabat IT Center, ENSIAS, Mohammed V University in Rabat, Morocco

***Abstract*—Business process is a sequence of events and tasks that encompass actions and people. Therefore, a company that pays much attention to its business processes has to clearly identify and define the procedures of their relations. However, with the exponential evolution in ubiquitous computing, the exploitation of information spread all over different devices has become essential to further improve business processes. Hence, in this paper we present a new approach for business process modeling that is based on context-awareness and ontology. We propose a set of meta-model for the elements that we find very important, taking into account the context and, in modeling, the business process. To validate our approach, we propose a concrete case study about transport system to provide a proof of the applicability as well as the utility of the model.**

***Keywords*—Business process; business process modeling; context; context-awareness; ontology**

I. Introduction

Companies are known by their constant search for competitive advantages in order to create new business opportunities and gain new areas of markets. Thus, they are permanently in need to perform digital transitions and continuous improvements which can only be achieved through implementing technology. One of the most known transformation is Business Processes that were introduced to help companies to regain their lost positions and powers. The use of Business Processes has a huge impact because it brings into existence new kinds of process-oriented organizations in which the structural hierarchy degraded. Under process-oriented organizations the contributors are assigned tasks and accomplishments which are sorted accurately to avoid any unexpected problems. Business Processes has made significant results in increasing the performance indicators (KPI, related to time, cost and quality of the organizations). So, the modeling of business processes is at the centre of an organization's analysis process. Whether as part of a global reorganization or a targeted improvement approach. Process modeling formalizes the function of an organization. It involves structuring and representing the activities of an organization, by using a graphical notation to visually represent the flow of activities. Modeling is based on specialized methods and tools as well as implements process reference frameworks.

In recent years, several researchers have proposed approaches for business process modelling.

It also displays new trends in ICT and the ubiquitous computing. The development of context-awareness service is a challenge in current research projects. In fact, since its appearance in the 1990s, several researchers have proposed many approaches to meet the challenge of context awareness, especially with the birth of service-oriented architectures.

However, several research questions remain to be addressed. This is due to a series of reasons, in particular the modeling of contextual information and the consideration of context in the business processes modelling.

However, our work allows to model contextual information, integrating this model in the modeling of business processes in order to obtain a contextual business process. With the importance and value that the semantic aspect gives to the model, we will present a semantic model of the contextual business process model based on ontology.

In this article, we integrate the context into the BP as follows:

- We define the context by the proposal of a context and sensor metamodel.
- We define a business process model that takes into account the context.
- We enrich this model by adding a semantic aspect.
- We apply our approach to a case study.

The rest of this paper is organized as follows: We start with a background and related works where we present some definitions and the state of the art on the BP modeling and the context. Section three presents the different models of our approach as well as the case study of transport. In section four we give a simple conclusion and a vision of our future work.

II. Background

A. Context

The notion of context has been a research subject for years, and until now, there is no general and standard definition. Apparently, several researchers have tried to give a definition, we are citing the most widely used definitions in the scientific community.

Let's start with the definition of [30] which categorized the context into three types, location, the identity of the person and the object in question. [31] defines the context as having three basic elements, the time, the date and the changes in that period.

But the most used definition, on which we rely is the definition of Dey et Al [32], it defines the context as follows:

"Context is any information used to characterize the situation of an entity. An entity may be a person, place or object considered relevant for the interaction between a user and an application, including the user and the application itself."

B. Context-Awareness

Regarding context Awareness, several visions are proposed to define a context-aware system. [1] Considers context-aware applications as "applications whose behaviour may vary depending on the context". Its applications must automatically extract information and perform actions based on the context of the user detected by the sensors. Similarly, [2] perceives context Awareness as the ability to adapt the functioning of a system in order to increase its usability and efficiency by taking into account the context of the user. Another interesting vision is provided by [3] which makes it possible to link the sensitivity and the context to the service offered by the application or the system. In fact, a system is said to be context-aware if it can automatically change these forms of services or if it is able to trigger a service as a response to changing the value of an information or a set of information which characterize the service.

These definitions may be considered limited, according to [4], since they exclude applications that use only the context and that do not make any adaptation of context-aware systems. To generalize these definitions, [4] considers that a system is context-aware if it uses context information to make information or / and services useful to the user available and considers that this utility is dependent of the user's task. This definition is also limited since a context-aware system deals with other aspects other than the context acquisition component, such as the interpretation of the context and the necessary adaptations. To overcome these shortcomings, [5] perceives that a system's sensitivity to context "is its ability to acquire, manage, interpret, and respond to changes in the context in order to provide the appropriate service". Thus we can consider that the two definitions of [4] and [5] complement one another and can be considered as reference to define a context-aware system.

C. Context-Awareness

For [6] ontology is the assumptions we make about the kind and nature of reality and what exists. Ontology is also defined as the nature of the world and what we can know about it.

In the field of computer science, ontology for [7] refers to "an explicit specification of a conceptualization" or a more refined way while [8] sees it as a "partial and formal specification of a shared conceptualization". Ontology is among the most used tools of the Semantic Web and Artificial Intelligence. It allows to model resources based on conceptual representations of the domain under study and allows systems and applications to draw inferences from them.

D. Business Process

A business process is a sequence of events that encompasses actions, people and the sequence of work. This concept comes from production lines but applies also to commercial transactions. Whatever the structure of a manufacturing, service or sales organization, it is underpinned by its business processes.

In the literature, several definitions are available for the Business Process. [9] Defines the business process as "a structured, measured set of activities and flows that use necessary resources of the organization to provide specified output for a particular customer". [10] Defines it as follows "the combination of a set of activities within an enterprise with a structure describing their logical order and dependence whose objective is to produce a desired result". [11] Defines a process as a productive activity which includes working for something, moving of people, materials, information and interaction.

III. Related Work

Business process modeling is a major challenge for all companies that aim to understand and improve their business. With the emergence of ubiquitous computing, consideration of the context has become paramount in the BP. So, in this section, we will present some approaches that are related to our work.

Modeling contextual information consists of representing the context in a well-defined format that is easily shared and used by the various stakeholders of the application. However, several approaches have been proposed to standardize the Context Modeling: [12][13] model the context as keys/value or each key defines the contextual information so the value presents its value, [14][15][16]are based on mark-up language, it is a more structured approach that presents information in a hierarchical structure, [17]presents the context in an object-oriented way, so the contextual information is encapsulated in an object in order to model it in a ubiquitous computing,[18] model is based on MDA, it allows for a model that is based on MOF (Meta-model facility) and[19][20] use ontology to model contextual information in a declaration form, so it allows for a semantic reasoning in these information.

Also for business process modeling, several methods and languages are proposed such as UML which allows for a modeling of any system in an object-oriented way, EPC (Event driven chains), IDEF0 that allows to model an organization's actions, Petri Net which is a mathematical graphical method to present systems, PSL (Process Specification Language) [21] that allows to represent model business process based on an ontology, etc. Thus, BPMN which is the standard OMG for business process modeling simplifies the understanding as well as manipulates and improves the quality of a business process.

In order to integrate the context in business process modeling several approaches have been proposed, including those based on the methods mentioned above, to model the context, the BP and other approaches that invent new methods. [22] extends BPMN meta-model in order to add elements (task, collector) to represent contextual elements. [23] has proposed business process and context meta-models, [24] is a layered architecture which is proposed to represent the domain, the context and the BP in order to produce a Contextual business process. [25] has proposed an approach to model the contextual business process based on BPMN. [26] proposes a new context model that takes into account the business process situations, Author in [33] has proposed a conceptual method of depicting the context of a business process, in [34] authors has proposed a method based on DMN for modelling business process .

Other methods have added semantic aspects to business processes: [27] have proposed an ontology of BPMN language that called BPMN Ontology, it allows to represent BPMN in the form of OWL-DL. [28] has proposed a framework that are called BPAL to manage ontologies based on BP modeling concepts (processes, atoms etc.).

IV. Case Study: Transport System

The system we have chosen as a case study is an example of an informational system made by a transport company that suggests the best routes in Rabat city (MyCity App). Indeed, MyCity is an application of transport services meant for residents and passengers of the city of rabat. A user of this application is anyone who wishes to have a vision on a journey inside the city of Rabat.

On the other hand, the current applications don't provide information about transports timetables and users' information (preferences, history, etc.) in order to make the service appropriate to the user. As a result, the services provided by these applications and the routes offered are not optimized for the users in question.

Indeed, a typical transport application includes, but is not limited to, the following services:

- Search Trip: the user enters his destination and,(if he wants, his location) otherwise, his location will be automatically collected by the sensory system.
- Display results: The system displays the different possible routes and orders them according to the history (preferred means of transport/departure time, etc.) and the user preferences (user context).
- Choose route: the user must choose the preferred route.
- Display route: the system must display the path, it must take into account the configuration of the user's device and it must follow the path with him in real time.

Depending on the scenario, several challenges arise:

- Context capture from various sources (user context, device, etc.).
- Considers the user's context and the context of transport in the various services (journey search, journey display, etc.).

Indeed, all these services and challenges are implemented in our system.

V. The Proposed Approach

Computers, smartphones, smart sensors are occupying more and more space in our life. However, the amount of information that circulates on these devices can help us in the enrichment of several systems that we will need daily, such as transport, health, telecommunication, sport and any system or service. Therefore, our idea is to take advantage of all this data (which is voluminous and rich) to improve the quality of services for users. As Mark Weiser says: "There is more information available at our fingertips during a walk in the woods than in any computer system, yet people find a walk among trees relaxing and computers frustrating. Machines that fit the human environment, instead of forcing humans to enter theirs, will make using a computer as refreshing as taking a walk in the woods".

However, with the place that Business Processes takes in the strategy of companies, and with this area of ubiquitous computing, the integration of contextual information into BP will increasingly improve the quality of service for end-users and provide a clearer vision for management to make the right decisions.

Our approach as in Fig. 1 is based on several components/models:

- Sensors: it is the model which is responsible for the representation of the sensors.
- Context: allows to model contextual information.
- CBPM: Contextual Business Process Model, it is the model that presents business process.
- CBPM Ontology: Contextual Business Process Model Ontology, it is an ontology that corresponds to the CBPM model.

Therefore, the ultimate objective of the proposed approach is to provide a general framework for modeling the context and to integrate it into the modeling of business processes, then add a semantic aspect to this model by proposing an ontology that corresponds to the proposed model.

Indeed, by the proposal of these models we were able to model the most important elements of the context and the BP, and to also bring them together in a single model that is CBPM which simplified the use of context in the execution of the BP model. This modeling approach facilitates the integration and the re-use of one or more models of our approach by other related systems / applications.

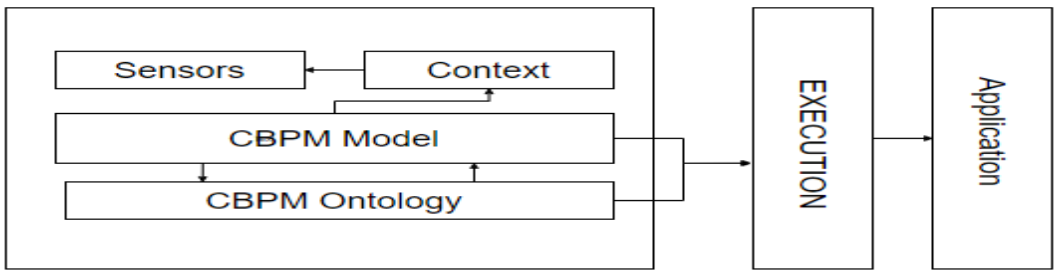


Fig. 1. The Proposed Approach.

A. Sensors

Sensors are components that are responsible for extracting contextual information. A sensor is any element capable of returning information about the user and the environment.

In general, there are two types of sensors: logical and physical:

- Physical sensors: they are the materials that can extract information related to the context of the organization, the user or the business process itself. Like GPS, localization, the state of the execution of the business process etc.
- Logical Sensors: they are intelligent programs that allow to retrieve information from the different possible sources: web services, user calendar, reasoning about user preferences.

In order to properly represent the different Sensors, we present in Fig. 2 a simplified meta-model of sensor. Fig. 3 presents the application of this meta-model to our scenario of transport system presented above. Among the different sensors related to this field, we have chosen the following sensors:

- Location Sensor: it is a physical sensor that captures the location of the user.
- Preference Sensor: it is a logical sensor that allows us to reason about the user's history in order to obtain his preferences.

B. Context

This model makes it possible to represent the context in a well-structured format in order to use it by other components. For the context in the Business process, it is any information that can give an added value to the modeling, execution or improvement of the BP.

So, and in order to give a generic formalization to the context that will be usable in any application, we propose in Fig. 4 a context metamodel based on the work [29].

Our meta-model is composed of the following elements:

- Context: represents the context, it is composed of sub Contexts.
- Sub Context: represents a sub-context, it is composed of Property context.
- Context Property: it is the contextual element; it can be attribute or Context Relationship.
- Sensor: represents the context provider.
- Context Relationship: represents the relationships between the different elements of the context.
- Attribute: is an atomic value that represents contextual information.
- Validity: represents the period of validity of a contextual information.

According to our scenario, we present in the Fig. 5 the context of our transport system, it is divided into the following sub Context:

- Use Context: Contains the attributes that describe the user's context (location, Preference).
- Environment Context: Contains the attributes of the environment.

Device Context: Contains the attributes which describe the device information (screen size, battery, ram, ...).

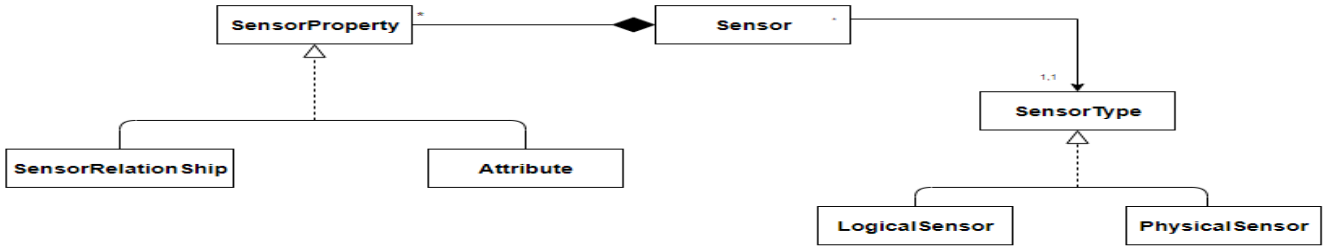


Fig. 2. Sensors Meta-Model.

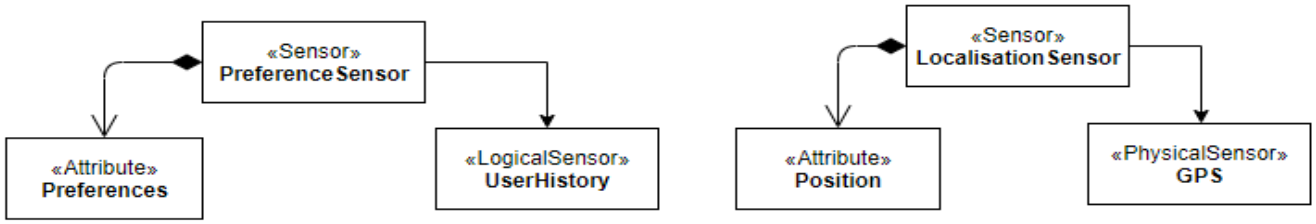


Fig. 3. Transport Sensors.

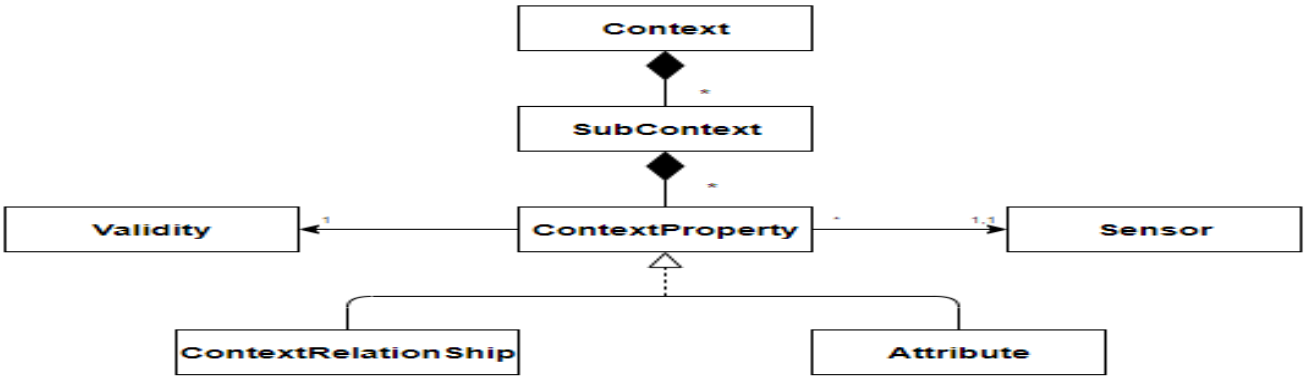


Fig. 4. Context Meta-Model.

C. CBPM Model

Taking the context into account in the business process significantly increases its representation and the productivity of the organization. And to respond to the different challenges of the BP, we have proposed the meta-model presented in Fig. 6 which represents the most important aspects of a business process, and it models the integration of context in this BP.

Our business process meta-model is composed of the following elements:

- Process: represent the business process.
- Sub Process: represents a subset of processes, it is composed of Process Element.
- Process Element: represents the main elements of a business process, it can be Gateway, Event or Service.
- Service: represents the element of Process which are services, in general these are the services offered by the organization.
- Event: represents the elements that can be triggered before, during or after the execution of the business Process.
- Gateway: it is a connection between the different services and events.
- Data: represents the data used in the business process, they can be Process Data or Context Data.
- Process Data: represents the process data.
- Context Data: represents contextual information.

Let's apply this meta-model to our case study, according to the scenario our application offers several services such as the search for a route, display of a route... the Fig. 7 presents the Contextual Business Process Model of the transport system.

D. CBPM Ontology

In order to enrich the different elements of the CBPM model with a semantic aspect, we have proposed the CBPM Ontology in Fig. 8 which is the result of a transformation of the CBPM model.

CBPM Ontology is made up of three main ontologies:

- Process Ontology: this is the ontology that describes the different elements of BP.
- Context Ontology: this is the ontology that describes the user's context, device and environment.
- Domain Ontology: This is the ontology of the studied domain, like Transport Ontology for the transport domain.

The mapping between the elements of CBPM model and CBPM Ontology is based on the classes and instances of the ontology and the classes and properties of the model CBPM. This mapping allows to give a semantic formalization to our CBPM model.

Fig. 8 presents also the mapping based on the case study described above.

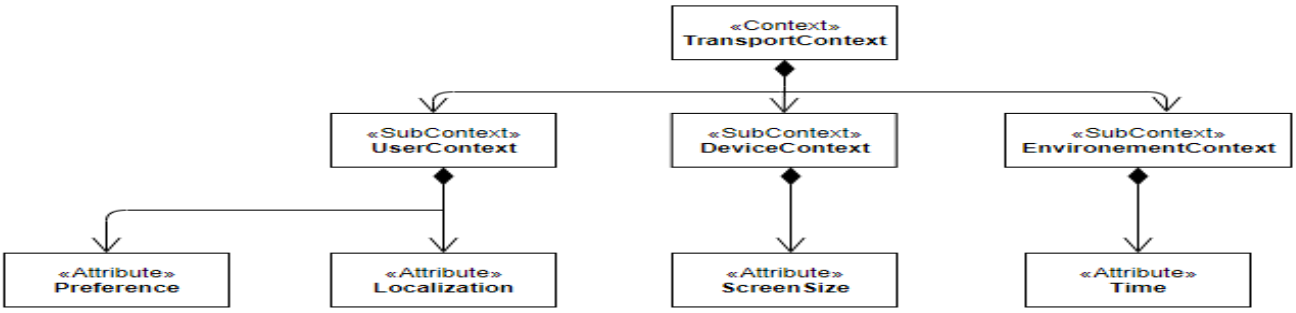


Fig. 5. Transport Context.

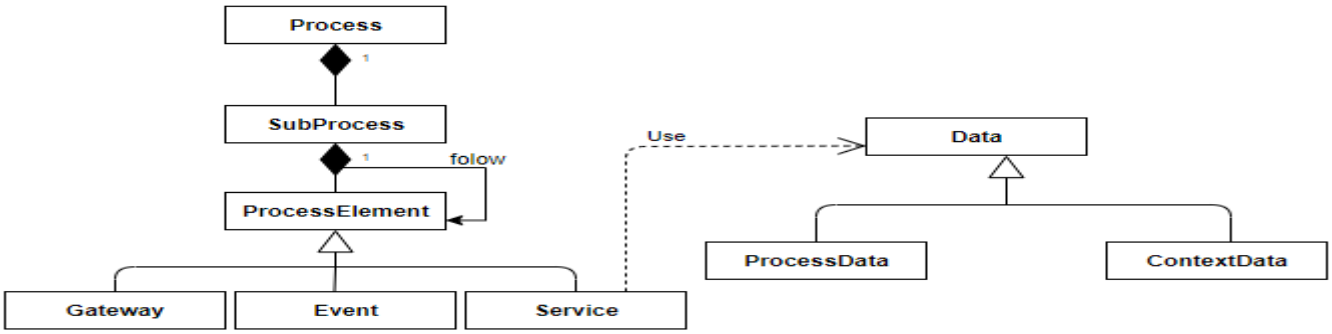


Fig. 6. CBPM Meta-Model.

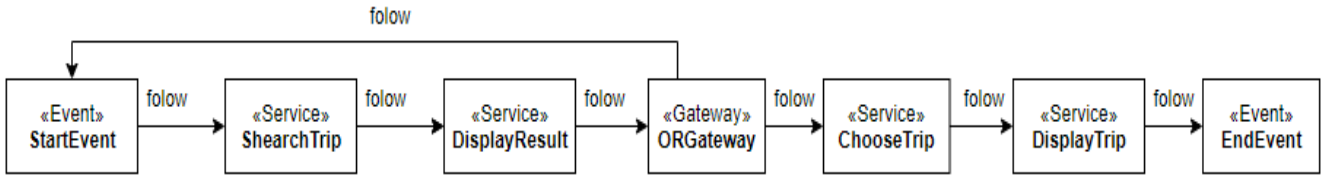


Fig. 7. Transport CBPM.

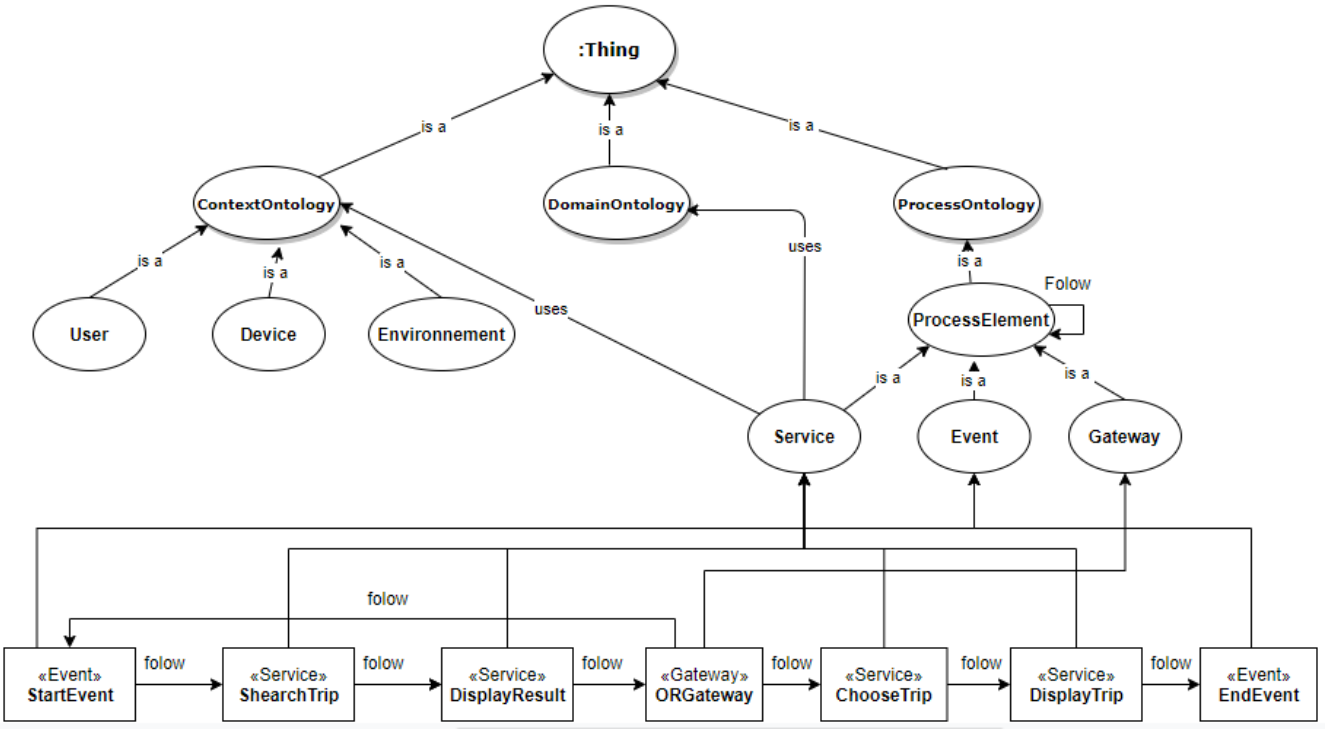


Fig. 8. CBPM Ontology.

VI. Result

Based on the different meta-models presented above, we have created a general framework to model not only the business processes, but also the contextual information, the sensors that allow us to extract this information. CBPM is a model that allows to model the business processes in a generic way independent of the field studied, and also to integrate the notion of context.

As mentioned above, our approach has been designed in such a way that it is easily integrated and reusable globally or partially with other systems. For example, for applications that rely only on the context without BP can use our sensor design model and context.

In fact, in order to validate our approach, we applied it to the transport study case which presents a challenge for the use of contextual information in the implementation of business processes.

Indeed, from the CBPM Model, we were able to extract the various services and events of our business process as well as the gateways. This has facilitated its execution and the adaptation of these services to the context of the user.

Indeed, for the search Trip service presented in Fig. 9, it allows to launch a search of the best routes between the departure and the destination of the user. This service is based on the geographical position, for the display Result service presented in Fig. 10, it is based on the preferences and the history of the user in order to propose routes.

In fact, all of these services are based on the hardware characteristics of the user's device (size, PC/Mobile).

Fig. 9. SearchTrip.

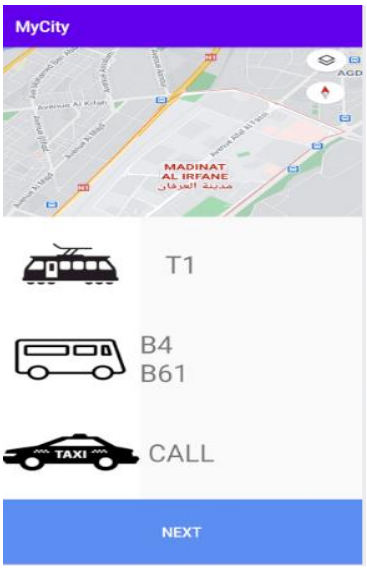


Fig. 10. Display Result.

VII. Conclusion

In this article we have presented a new approach to model the context and the business process. We proposed an architecture with several components, CBPM:

- Sensors: to model sensors according to their different types (logic, physics).
- Context metamodel: a context metamodel to represent contextual information.
- CBPM metamodel: the business process model that takes into account contextual information.
- CBPM Ontology: an ontology which is the result of a transformation of the CBPM model based on the rules for converting meta-models into ontology, which gives the latter a semantic aspect.

In addition, we presented the classical methods and language for business process modelling. We also stated approaches that have responded to the problem of taking context into account in the business process and others that have enriched business models with semantic aspects through approaches like ontology.

In our future work, we will use this model to propose an approach for business process improvement taking into account the different service events gateway presented in the CBPM Model.

References

[1] P. Brown, J. Bovey and X. Chen, Context-aware Applications: from the Laboratory to the Marketplace, IEEE Personal Communications, Vol 4, no 5, pp. 58–64, 1997.

[2] M. Baldauf, S. Dustdar, and F. Rosenberg, A survey on context aware systems, International Journal of Ad Hoc and Ubiquitous Computing, vol. 2, no. 4, pp. 263–277, 2007.

[3] M. Miraoui, and C. Tadj, A service oriented definition of context for pervasive computing, In The 16th International Conference on Computing, IEEE Computer Society, 2007.

[4] A. K. Dey, Supporting the construction of context-aware applications, In Dagstuhl Seminar on Ubiquitous Computing, Dagsthul, Germany, 2001.

[5] W. Xiaohang, The context gateway: a pervasive computing infrastructure for context aware services, Technical report, National university of Singapore, 2003.

[6] RICHARDS, Keith. Qualitative inquiry in TESOL. Springer, 2003.

[7] T.R. Gruber, A translation approach to portable ontology specifications, Knowledge Acquisition, Vol. 5, no. 2, pp. 199–220, 1993.

[8] N. Guarino, Formal ontology, conceptual analysis and knowledge representation, International Journal of Human-Computer Studies, Vol. 43, no. 5/6, pp. 625–640,1995.

[9] Laakso, T. *Process Assessment Method — an approach for business process development*. In: Plonka F., Olling G. (eds) Computer Applications in Production and Engineering. IFIP — The International Federation for Information Processing. Boston, MA: Springer. 1997.

[10] Aguilar-Saven R.S.. Business process modelling: Review and framework, *International Journal of Production Economics*. 90: 129–149. 2004.

[11] Anttila, J.; Jussila, K. An advanced insight into managing business processes in practice, *Total Quality Management* 24(8): 918–932. . 2013.

[12] Bettini., 2010: A survey of context modeling and reasoning techniques. Pervasive and Mobile Computing, vol. 6, no 2, pp. 161-180, 2010.

[13] Strang, A context modeling survey. In Ubicomp, First International Workshop on Advanced Context Modeling, Reasoning and Management, 2004.

[14] Venezia, Pervasive ICT Social Aware Services enablers. 14th International Conference on Intelligence in Next Generation Networks (ICIN). Berlin, Germany, 2010.

[15] Knappmeyer:ContextML: A light-weight context representation and context management schema. 5th IEEE International Symposium on Wireless Pervasive Computing (ISWPC), Modena, Italy, 2010.

[16] Klyne: G. Klyne, F. Reynolds, C. Woodrow, H. Ohto, J. Hjelm, M.H. Butler and L. Tran, Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies 1.0. W3C Recommendation, 2004.

[17] Cheverest: The role of adaptive hypermedia in a contextaware tourist guide, Communication of ACM, vol. 45 no. 5, pp. 47-51, 2002.

[18] Chihani: Enhancing Existing Communication Services with Context Awareness, Journal of Computer Networks and Communications, Volume 2012.

[19] Ferry: Adaptation Dynamique d'Applications au Contexte en Informatique Ambiante, Research Report I3S , number I3S/RR-2008-20-FR 2008.

[20] Chen: An ontology for context-aware pervasive computing environments. The Knowledge Engineering Review, vol. 18 no. 3, pp 197–207, 2004.

[21] Gruninger Michael, Schlenoff Craig, "Process Specification Language (PSL):results of the first pilot implementation", Proceedings of IMECE: International Mechanical Engineering Congress and Exposition, pp 1-10, 1999.

[22] Alaaeddine Yousfi, Christine Bauer, Rajaa Saidi, Anind K. Dey, uBPMN: A BPMN extension for modeling ubiquitous business processes, Information and Software Technology, Volume 74,Pages 55-68, 2016.

[23] Saidani, Oumaima & Rolland, Colette & Nurcan, Selmin. (2015). Towards a Generic Context Model for BPM. 2015. 10.1109/HICSS.2015.494.

[24] Santoro, Flávia & Baião, Fernanda & Revoredo, Kate & Nunes, Vanessa. (2017). Modeling and Using Context in Business Process Management: A Research Agenda. Modélisation et utilisation du contexte. 17. 10.21494/ISTE.OP.2017.0130.

[25] SANTRA, Debarpita et CHOUDHURY, Sankhayan. C-BPMN: A Context Aware BPMN for Modeling Complex Business Process. arXiv preprint arXiv:1806.01333, 2018.

[26] DA CUNHA MATTOS, Talita, SANTORO, Flávia Maria, REVOREDO, Kate, et al. A formal representation for context-aware business processes. Computers in Industry, vol. 65, no 8, p. 1193-1214, , 2014.

[27] Marco Rospocher and Chiara Ghidini and Luciano Serafini: An ontology for the Business Process Modelling Notation, FOIS,2014.

[28] De Nicola, Antonio & Lezoche, Mario & Missikoff, Michele. (2007). An Ontological Approach to Business Process Modeling. 1794-1813.

[29] Guermah: Context modeling and reasoning for building context aware services. In ACS International Conference on Computer Systems and Applications (AICCSA), IEEE, pp. 1-7, May, 2013.

[30] Bill N Schilit, Marvin M Theimer: Disseminating active map information to mobile hosts. IEEE network, vol. 8, no. 5, pages 22–32, 1994.

[31] Guanling Chen, David Kotzet al: A survey of context-aware mobile computing research. Rapport technique, Technical Report TR2000-381, Dept. of Computer Science, Dartmouth College, 2000.

[32] Anind Dey, Gregory Abowd, Peter Brown, Nigel Davies, Mark Smith & Pete Steggles. Towards a better understanding of context and contextawareness. In Handheld and ubiquitous computing, pages 304–307. Springer, 1999.

[33] Zhao, Xiaohui, et al. "Enabling intelligent business processes with context awareness." 2018 IEEE International Conference on Services Computing (SCC). IEEE, 2018.

[34] Song, Rongjia, et al. "A DMN-based method for context-aware business process modeling towards process variability." International Conference on Business Information Systems. Springer, Cham, 2019.

Hybrid Decision Support System Framework for Leaf Image Analysis to Improve Crop Productivity

Meeradevi, Monica R Mundada
Department of Computer Science and Engineering
M S Ramaiah Institute of Technology
Bangalore, India

***Abstract*—Crop disease is one of the major problems with agriculture in India. Identifying the disease and classifying the type of disease is most important which can be made possible using the deep learning technique. To perform this verified dataset is required which consists of healthy and disease leaf images of all crops. The proposed model uses a hybrid approach which integrates VGG16 classifier with an attention mechanism, transfer learning approach and dropout operation. The proposed model uses a rice disease dataset and using the proposed approach it achieves an train accuracy of 96.45 percent and train loss 0.09 and validation loss of 0.44. The dataset is collected from the plant village project for rice leaf which consists of 4955 images which include Brown Spot, Healthy, Hipsa, and Leaf Blast type of images. The proposed model use attention mechanism that focuses mainly on the part of the image rather than the whole part of the image using a glimpse ratio of 3:1. The traditional method of detecting crop diseases needs high experience and knowledge of experts in the field which is time consuming, ineffective, and high cost. In this study, Deep Convolutional Neural Networks (DCNN) and Transfer Learning with Attention models are used to detect diseases associated with rice plants without overfitting the model.**

Keywords*—Deep learning; activation function; attention mechanism; dropout operation; transfer learning; VGG16

I. INTRODUCTION

Country's economy depends on agriculture. One of the reasons that effect agro-economy is plant disease, which harms the entire crop by spreading the disease throughout the field. So, detection of disease becomes very important. Identifying leaf disease can be made possible with image dataset. The dataset comprises of both diseased and healthy leaf images which are the primary source of data for identifying the disease portion on the leaf. For early detection of leaf disease it is required to monitor the disease, in time treatment of disease by applying pesticide and minimizing the spread of disease and reduce the loss. Testing the model with pre-trained weights for better classification of disease is made using CNN [1]. The proposed approach uses convolution neural (CNN) network which is best approach for image segmentation. DCNN based technique is used to train the VGG16 classifier in order to distinguish healthy and infected rice leaf with multiple layers in CNN model. CNN based VGG classifier uses dense layer with convolution and max pool layer. The dataset consist of four types of leaf images three are diseased leaf images (Brown Spot, Hipsa and Leaf Blast) and healthy images. The preprocessing is done using pre-trained ImageNet weights. These pre-trained images are trained on millions of images with 1000 different categories. Using this pertained model in the proposed approach will reduce the time of the training new images. All Images are of three dimensions i.e., height, width and channels (RGB). Deep learning model requires high GPU. Every convolution layer is followed by ReLu activation function which increases non-linearity [2]. Designing a rice leaf disease detection model using deep learning algorithms is the main idea of the proposed model which is achieved using hybrid approach VGG16 with attention model. Vijay Kumar V [3] design a robot for monitoring the agricultural field condition such as soil moisture, crop quality, pesticide for good quality crops and supplying required amount of water. The robot is designed using Lattepanda which is Chinese board which runs intel processor integrated with machine learning model and an android application is designed for controlling the robot. The machine learning model integrated with robot performs clustering using mean shift vector and density estimation window is applied for converging to center of maximum dense area. For classification of disease SVM classifier is used which takes the input image labeled data and outputs the optimal solution. The robot comprises of two motors one for driving the robot and other for performing the operation of sensing soil moisture and humidity which is connected to robot. [4] Plant viruses are threat to agriculture productivity. Controlling the disease is based on two aspects i.e., genetic resistance (immunization) and prophylaxis to restrain virus dispersion by removing the infected plants. The paper discuss on the effect of plant viruses and genetic diversity. Plant disease identification methods are based on the plant DNA classified as polymerase chain reaction (PCR) and isothermal amplification. In PCR specific viral region of plant is identified and visualized by electrophoresis. [5] adapting the advanced decision support system, digitized and data driven technology identification of plant disease can be made easy. The paper proposes mathematical model based on deep learning technique. The region proposal network (RPN) technique is used where the images are segmented based on the RPN algorithm results. The segmented images are input into transfer learning the model is trained with various diseases and obtained the accuracy of 83%. The following sections in the paper discusses about the existing study made in this domain followed by methodology adapted for implementation next section shows the proposed VGG16 architecture followed by discussion on results obtained using proposed hybrid approach.

II. Literature Survey

The plant disease recognition and classification is done using image processing technique. The paper discuss about noise reduction and image segmentation technique for diseased part of leaf. Banana leaf is considered for disease identification. Noise is introduced in the image by various factors. Some of the main factors that induce noise are environmental conditions, variation in light and sensor temperature, transmission passage. Paper discuss about filtering technique like linear filters, adaptive filters, non-linear filter for removing noise [6].

The plant disease cause serious effect on countries economy [7]. This paper discusses the algorithm for image segmentation which detects and classifies the type of disease that appears on plant leaf. Digital camera captures the image and those images are used to identify the infected part of leaf. Various image processing techniques are applied on images to analyze the features. Workflow of the model is firstly image need to be acquired and preprocess the image to remove noise and improve the quality of the image. Compute the threshold value for the green colored pixels. Pixel value is compared with threshold value if the pixel value of green component is less than threshold value than zero is assigned to RGB components of the this pixel. Finally obtain the image segments to classify the leaf disease. The author performed experiments using MATLAB. Input is rose, banana, beans, lemon leaf image with bacterial disease.

The image classification is done using k-means clustering with 86.5% accuracy and detection accuracy is improved to 93.3%.

In [8] the proposed model use wavelet tool for image analysis. Maize leaf image is considered for noise removal in image. This noise is introduced in the picture due to variation in light and environmental conditions and also on image acquisition equipment. Adaptive local smoothing method is proposed in this paper. Wavelet tool is used for analysis.

In [9] image enhancement is done using filtering methods like Gaussian filter, Mean filter, median filterand wiener filter. Comparative study done among all filter and wiener filter gave better result with high signal to noise ratio. The proposed model is used to identify the rice plant disease brown spot. In [10], paper use random forest to identify the healthy and diseased leaf for dataset. The proposed model creates dataset and performs preprocessing of the image to bring all images to unified dimension and extract features of the image using Histogram of oriented gradient. The model is trained using Random forest classifier for classifying healthy and diseased leaf. Comparative study is done with Gaussian Naïve bayes, logistic regression, linear discriminant analysis, SVM and random forest has shown better accuracy for smaller dataset. The main objectives of the proposed model is identification of leaf disease using hybrid approach, implemented model must correctly classify the disease and finally evaluation of the performance. The plant is susceptible to disease due to soil quality also and soil health [11]. The frequent change in climate condition and the common practice in agricultural ecosystem is use of extensive pesticides and fertilizers and the effect of abiotic stresses have made the crop to degrade in quality and lead to reduced production [12]. The paper use arbuscular mycorrhizal fungi (AMF) for enhancing crop productivity. AMF are bio-fertilizers and prvides tolerance to the plants against various stressful situations like drought, heat, salinity and varying excessive temperature and weather condition. Nitrogen being one of the essential nutrient for plant productivity which is applied more but it has a negative impact on the plants and environment [13]. Nutritional exchanges between plant, arbuscular mycorrhizal fungi, and bacteria that help improve plant nutrition, including nitrogen (N) acquisition. Plant N acquisition can be improved in the presence of N2-fixing symbiotic and associative symbiotic bacteria and arbuscular mycorrhizal fungi (AMF). [14] ML based techniques have achieved a great attraction in digital image processing and prediction. Tough there are various challenging technologies this paper has come with hybrid approach by enhancing image, conversion of image and removing noise and applying GLCM technique and finally neuro-fuzzy logic classifier is used to train the model and extract features. The model is implemented using MATLAB and the average test accuracy obtained is 90%. The proposed algorithm in this paper uses a-priori information about the shape of plant leaves [15]. The model is compared with state-of-art segmentation technique. The model detects leaf tips to improve the segmentation accuracy of the leaf. The algorithm reduced the error of detecting the leaf tips accurately and increased the detection accuracy. Image processing and soft computing techniques are combined to improve the detection accuracy [16]. Automated plant species is identified in this paper using image data. The SVM and ANN technology is applied on the dataset for plant classification. With 32 different types of leaves the model could achieve the accuracy of 94% [17]. In [18] the hybrid approach is proposed for automatic detection of leaf disease based on CNN and convolutional autoencoder (CAE). This hybrid model could obtain the accuracy of 99%. ANN and CAE is integrated because they efficiently and reduce the image dimensionality and extract various spatial and temporal features from image data. The CAE network is used to reduce the dimensionality of the image and the output of the CAE is network is given as input to CNN model for classification of diseased and healthy plants. The normalized root mean square error is used to reconstruct the loss between original and reconstructed leaf image. This integrated model loss is less compared to other models.

III. Proposed System Architecture

As shown in Fig. 1, the flow of work is as follow, Firstly load the image dataset collected from plant village project and convert the image to grayscale and bring all image to unified dimension and create balanced dataset for all type of images. Generate the array of pixel for each image and normalize the value in the range 0 to 1 and dump in pickle file. Shuffle the dataset and visualize using openCV which is computer vision technique. Now split the dataset into train, test and validation. Train set is 60 percent and test set is 25 percent and validation set is 15 percent. Next train the dataset using VGG16 model using pre-trained ImageNet data rather than training from scratch and define the activation function ReLu after every convolution layer and define softmax function at fully

connected layer. Apply attention glimpse with ratio 3:1 to focus on sub part of image rather than whole image. Than validate the data using validation set and once model is finalized, test using test dataset and make prediction. Accuracy of the model is analyzed using ROC curve. The proposed model achieves accuracy of 96 percent with hybrid approach without any overfit in model.

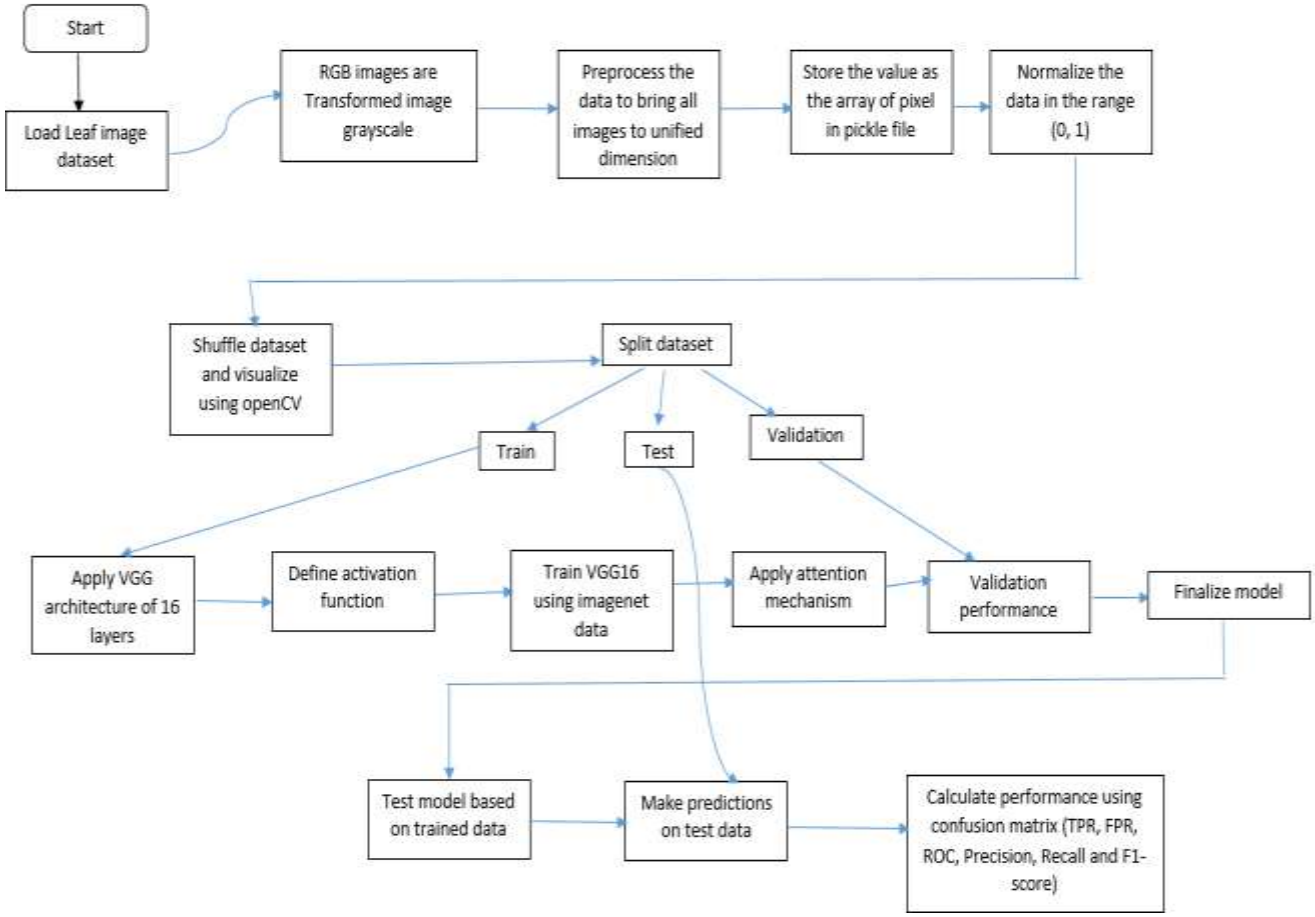


Fig. 1. Proposed System Architecture.

IV. Proposed VGG16 Classifier

Firstly the images are fed into neural network VGG16 architecture that contains five convolution blocks and three dense layers. First convolution block has 2 convolution layers followed by max pool layer. The image size is 180X180x64 height and width of image is 180X180 and 64 channels. Max pool layer reduce the dimension of the image and doubles the number of channels, so that more appropriate features can be extracted so, the max pool layer reduce the dimension to 90X90X128. Second convolution layer has two convolution blocks followed by max pool layer. The size of the image at this layer is 45X45X256. The third convolution block consists of three convolution layer followed by max pool. The size of the image is now at max pool layer 22X22X512. The fourth convolution block consists of three convolution layer with max pool layer. Max pool layer reduce the dimension of image to 11X11X512 and finally the last block has three convolution layers with max pool layer the dimension of the image becomes 6X6X512. The dropout operation is applied at dense layer which drops out some of the units randomly to avoid overfit in the model. The attention mechanism is applied with the glimpse ratio of 3:1 which focus of part of image rather than whole part of image at one time. Finally the fully connected layer uses softmax activation function which selects the most probable class among n classes. After every layer ReLu activation is used that will forward only positive weights to the next layer as shown in Fig. 2. The dataset considered for classification is shown in Table I. The main idea of using CNN is its advantage in image segmentation. Images can be of any size, it performs better than any traditional algorithms and each image is converted to array of pixel so, pixel wise predictions are done. Fig. 3 shows the grayscale images of various types of leaf disease used in proposed model Brown Spot, Leaf Blast, Healthy and Hipsa.

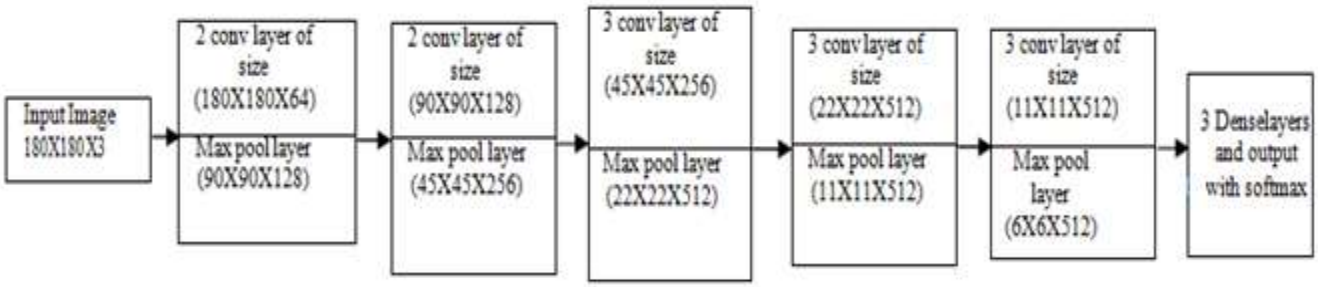


Fig. 2. VGG16 Architecture with Dimension of Image at every Layer

TABLE I. DATASET OF RICE LEAF IMAGES

Class	Number of Original Images	Number of Images inBalanced dataset
Brown Spot	923	400
Hipsa	1988	400
Leaf Blast	965	400
Healthy	1179	400

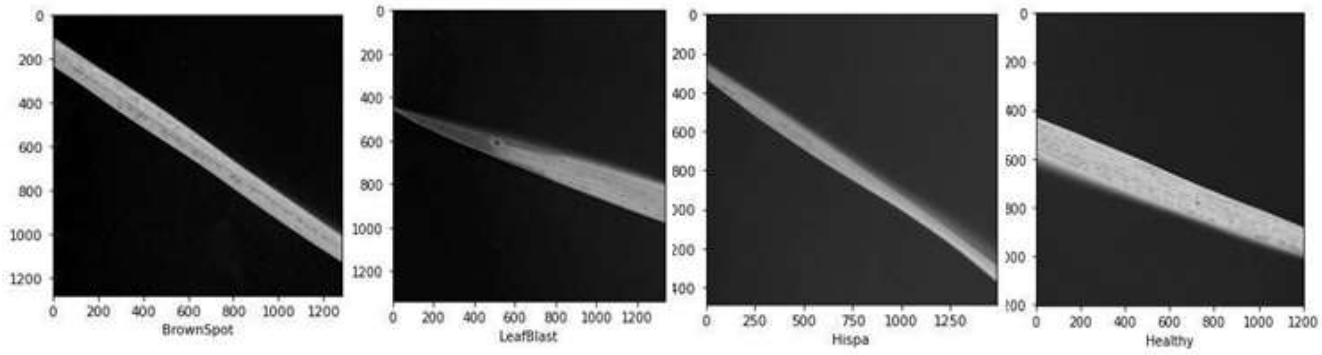


Fig. 3. (a) Brown Spot (b) Leaf Blast (c) Healthy (d) Hipsa.

The final Attention Model now has the following modules and properties:

- Dataset - Rice Plant Leaf Disease Dataset.
- Train split-0.60
- Test split-0.25
- Validation split - 0.15.
- Epochs - 100.
- Batch Size - 64.
- Transfer learning using pre-trained image net weights.
- Type of Classifier - VGG16Net.
- Dropout Regularisation - 0.4 (To Prevent Overfitting).
- Attention Layer - Image to Glimpse Ratio = 3:1. (To focus only on a subset of features on the neural network).
- Loss Function - Categorical Cross Entropy.
- Optimizer - Adam.
- Activation function-softmax.

V. RESULT AND DISCUSSION

The approach of transfer learning with pre-trained ImageNet weights is used in this model with Gaussian attention mechanism. Firstly the input layer takes the rice leaf image of size 180x180 and the convolution layer with ReLu activation function extract the low level features of the image and the convolution layer at the end of the VGG16 net extract high level features of the image. The proposed model works well even with small dataset with good accuracy. The hybrid approach is not proposed in state-of-art systems in literature study. In this work the hybrid model is applied to detect rice leaf disease of type Hipsa, Brown spot and leaf blast and healthy.

The proposed approach use transfer learning with ImageNet weights which is trained on millions of images with 1000 categories. In the proposed model four classes are considered with three disease class and one healthy class. The last layer is a fully connected layer with softmax function which identifies the most probable class. The learning rate of the VGG net model is 0.0001. The weights are updated using Adam optimizer and batch size is set to 64 and the model is run for 100 epochs. The accuracy obtained with this classification model is 0.9654 and model loss of 0.09 and validation loss is 0.44 and validation accuracy is 0.81 for 4955

images as shown in Fig. 4 and 5. Hence the validation loss is greater than train loss so there is no overfit in the model. The proposed model correctly classifies the images.

Fig. 6 represents the ROC curve. ROC curve shows x-axis with (1-specificity) and y-axis showing sensitivity. ROC is calculated on predicted scores and Fig. 7 shows the train loss train accuracy, validation loss and validation accuracy for 50th epoch. Fig. 7 clearly shows the train accuracy is greater than validation accuracy and loss is less than (<1) hence the model classifies the disease without any overfit with accurate prediction.

The accuracy of the model is cross-validated using confusion matrix as shown in Table II. The values f precision, recall and f-measure which gives the accuracy of the proposed model. Fig. 8 shows the ROC curve for all classes. ROC is showing good accuracy in the proposed model. X-axis shows the false positive rate and y-axis shows the true positive rate.

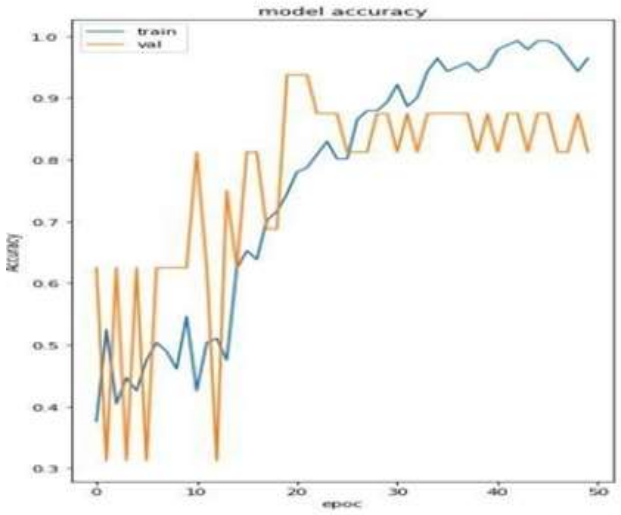


Fig. 4. VGG16 Model Accuracy.

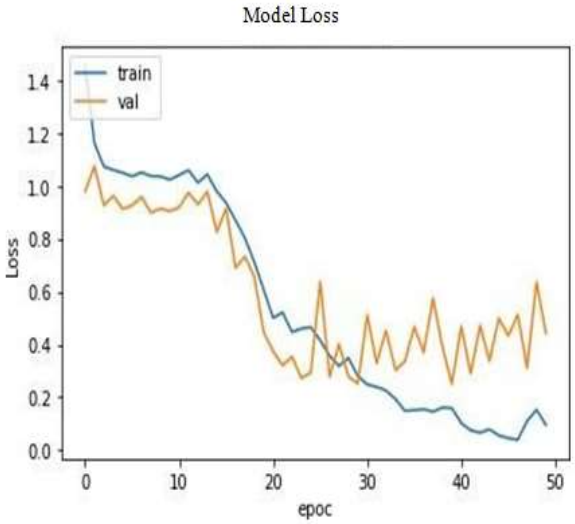


Fig. 5. Model Loss using VGG16 for Train and Validation Set.

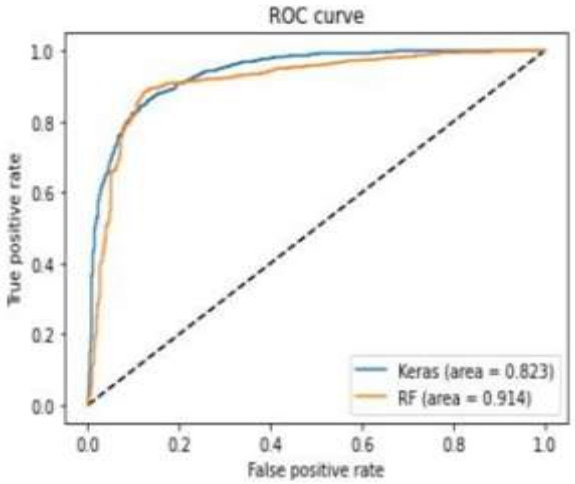


Fig. 6. ROC Curve for the Proposed Mode.

```
Epoch 49/50
128/141 [=========================>...] - ETA: 0s - loss: 0.1674 - accuracy: 0.9375WARNING:tensorflow:Can save best model on
ly with val_acc available, skipping.
141/141 [==============================] - 5s 38ms/sample - loss: 0.1535 - accuracy: 0.9433 - val_loss: 0.6356 - val_accurac
y: 0.8750
Epoch 50/50
128/141 [=========================>...] - ETA: 0s - loss: 0.1020 - accuracy: 0.9609WARNING:tensorflow:Can save best model on
ly with val_acc available, skipping.
141/141 [==============================] - 6s 42ms/sample - loss: 0.0962 - accuracy: 0.9645 - val_loss: 0.4433 - val_accurac
y: 0.8125
```

Fig. 7. Sample Model Run Showing Accuracy and Loss.

TABLE II. CONFUSION MATRIX

Class	Precision	Recall	F-measure
Brown Spot	0.963	0.948	0.955
Hipsa	0.958	0.657	0.780
Leaf Blast	0.808	0.937	0.868
Healthy	0.873	0.941	0.906

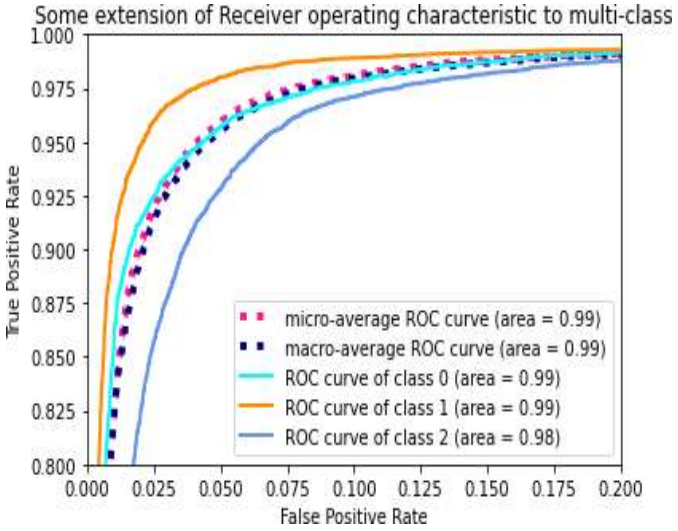


Fig. 8. ROC Curve for All Class.

VI. PERFORMANCE ANALYSIS

The VGG 16 model consists of 16 layers with one input layer and output dense layer. The dropout operation in the model randomly drops some of the units with the rate 0.4 to avoid overfit in the model. Fig. 9 shows total number of trainable parameters and non-trainable parameters. The omission of parameters at dense layer will lead to good accuracy without any overfit with 150 epochs. The performance analysis is done by varying various parameters like number of epochs, batch size, and dropout rate as shown in Table III. The performance of the model is evaluated by comparing with transfer learning using pre-trained ImageNet weights and next with dropout technique and finally dropout technique with attention model. The proposed model use five convolution blocks with transfer learning, dropout and attention mechanism hybrid approach has shown good accuracy as shown in Table III.

A. Comparative Study with other ML Algorithms

Table IV shows the comparative study of various ML models. The proposed model has shown good accuracy when compared to other models. The SVM does not work well with large dimension data and k-means is more sensitive to outliers and decision tree model is inaccurate because small change in data leads to large change tree structure. The proposed approach VGG16 net is a DCNN classifier which works for data of any size and fast in computation compared to other ML model. Hence, the accuracy is high. Fig. 10 shows the comparative study of various ML models and proposed model accuracy for precision, recall and f-measure with roc curve.

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
vgg16 (Model)                (None, 7, 7, 512)         14714688
_________________________________________________________________
flatten_2 (Flatten)          (None, 25088)             0
_________________________________________________________________
dense_4 (Dense)              (None, 1024)              25691136
_________________________________________________________________
dense_5 (Dense)              (None, 512)               524800
_________________________________________________________________
dropout_2 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_6 (Dense)              (None, 19)                9747
=================================================================
Total params: 40,940,371
Trainable params: 33,305,107
Non-trainable params: 7,635,264
_________________________________________________________________
```

Fig. 9. Total Number of Trainable and Non-trainable Parameters.

TABLE III. PERFORMANCE ANALYSIS

Parameters	Transfer Learning Model	Dropout Operation	Transfer Learning, Dropout and AttentionModel
CNN Blocks	5	5	5
Activation function	softmax	softmax	softmax
Dropout	-	0.5	0.4
Epochs	100	150	150
Batch Size	32	32	64
Loss Type	Categorical Cross Entropy	Categorical Cross Entropy	Categorical CrossEntropy
Optimizer	adam	adam	adam
Accuracy	0.9575	0.92	0.9645
Loss	0.15	0.26	0.09
Validation Accuracy	0.93	0.81	0.8125
Validation Loss	0.21	0.36	0.44

TABLE IV. COMPARATIVE ANALYSIS

ML Model	Precision	Recall	F-measure	ROC
K-Means	0.81	0.83	0.82	75.86
SVM	0.92	0.91	0.92	83.3
k-means with SVM	0.80	0.72	0.76	88.9
Decision tree	0.70	0.67	0.69	78.6
Proposed hybrid model VGG16 with transfer learning and attention	0.97	0.96	0.97	96.45

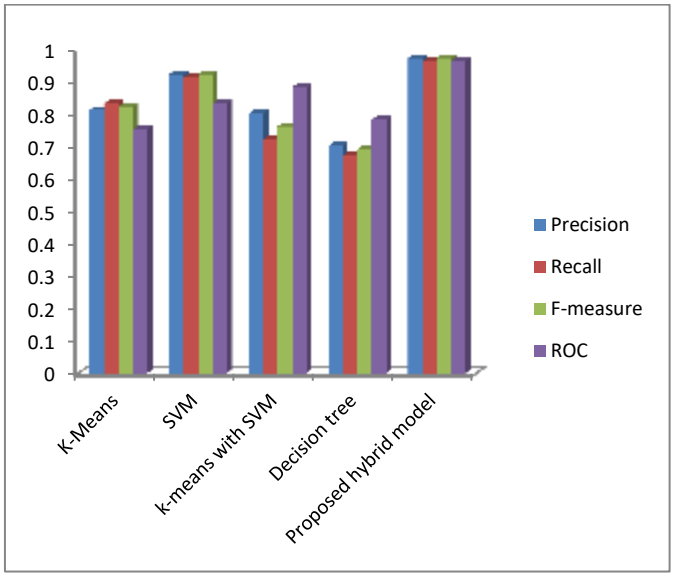


Fig. 10. Accuracy Values of Various ML Models.

VII. CONCLUSION

In the proposed model disease classification is done using rice leaf image dataset using pre-trained ImageNet weights. The proposed model achieves optimal accuracy. The performance of the model is evaluated by varying number of epochs, weights at every layer and batch size. The hybrid approach used in the model which integrates transfer learning, dropout and attention model helps in achieving high accuracy. The model will be using preprocessing techniques such as binarization of images and segmentation. The model will then be using CNN, Transfer Learning with Attention to classify images into their class labels. Results of the DCNN model will be compared with existing models to check which model has the better classification accuracy. This model when compared with the existing model provides a higher rate of accuracy and correct predictions as we are using a hybrid model. Plant disease classification can help farmers detect the disease in their crop before harvesting period. This allows them time to cure the crop from the disease before it can affect the majority of the crop. The model allows the farmer to get an accurate classification of the disease in order to take the appropriate countermeasures. Since farmers are able to detect and prevent spread of the disease, this helps to boost crop yield and furthermore helps in improving the economy.

In future more real time data can be collected using drones and robots and apply classification technique to get more accurate results and deploying the model in real time in agriculture field which will help farmers from huge loss due to diseases.

REFERENCES

[1] Aravind Krishnaswamy, Rangarajan Raja, Purushothaman Aniirudh Ramesh, “Tomato crop disease classification using pre-trained deep learning algorithm” International Conference on Robotics and Smart Manufacturing (RoSMa2018) ,Procedia Computer Science 133 (2018) 1040–1047 Elsevier.

[2] Kawcher Ahmed, Tasmia Rahman Shahidi, Syed Md. Irfanul Alam and Sifat Momen, “Rice Leaf Disease Detection Using Machine Learning Techniques”, 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), 24-25 December, Dhaka.

[3] Vijay Kumar V1 , Vani K S, “Agricultural Robot: Leaf Disease Detection and Monitoring the Field Condition Using Machine Learning and Image Processing”, International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 14, Number 7 (2018), pp. 551-561.

[4] Luis Rubio1, Luis Galipienso and In maculada Ferriol,”Detection of Plant Viruses and Disease Management: Relevance of Genetic Diversity and Evolution”, Frontiers in Plant Science 17 July 2020 Doi: 10.3389/fpls.2020.01092.

[5] Yan Guo, Jin Zhang Et.Al., “Plant Disease Identification based on Deep Learning Algorithm in Smart Farming”, Discrete Dynamics I Nature and Study, Hindawi, Volume 2020 DOI: 10.1155/2020/2479172.

[6] Ajitha N , Nandhini S, “Noise Reduction in Leaf Image by Fuzzy Based Filtering Technique”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 7, Issue 11, November 2018, ISSN (Online) 2278-1021.

[7] Vijai Singh, A.K. Misra, ”Detection of plant leaf diseases using image segmentation and soft computing techniques”, Informa tion Processing in Agriculture, 2017, publishing service by Elsevier.

[8] Y. Li, Y. Zhang, J. Zhu and L. Li, "Wavelet-based maize leaf image denoising method," 2010 World Automation Congress, Kobe, 2010, pp. 391-395.

[9] K.S. Archana , Arun Sahayadhas, “Comparison of various filters for noise removal in paddy leaf images”, International Journal of Engineering & Technology 7 (2.21) (2018) 372-374.

[10] Shima Ramesh Mr. Ramachandra Hebbar et al., “Plant Disease Detection Using Machine Learning”, 2018 International Conference on Design Innovations for 3Cs Compute Communicate Control, 978- 1-5386-7523- 6/18/$31.00 ©2018 IEEE DOI 10.1109/ICDI3C.2018.00017.

[11] Magdalena Frąc,Silja E. Hannula et.al., “Fungal Biodiversity and Their Role in Soil Health”, Frontiers Microbiology. 2018; 9: 707. 2018 Apr 13. doi: 10.3389/fmicb.2018.00707.

[12] Naheeda Begum, Cheng Qin, et.al., “Role of Arbuscular Mycorrhizal Fungi in Plant Growth Regulation: Implications in Abiotic Stress Tolerance”, Front. Plant Sci., 19 September 2019, Doi: 10.3389/fpls.2019.01068.

[13] Alia Dellagi, Isabelle Quillere, Bertrand Hirel, “ Beneficial soil-borne bacteria and fungi: a promising way to improve plant nitrogen acquisition”, Journal of Experimental Botany, Volume 71, Issue 15, 25 July 2020, Pages 4469–4479, https://doi.org/10.1093/jxb/eraa112.

[14] Anusha Rao, S.B. Kulkarni, “A Hybrid Approach for Plant Leaf Disease Detection and Classification Using Digital Image Processing Methods”, October 2020 International Journal of Electrical Engineering Education DOI:10.1177/0020720920953126.

[15] Joshua Chopin,Hamid Laga,Stanley J. Miklavcic, “A Hybrid Approach for Improving Image Segmentation: Application to Phenotyping of Wheat Leaves”, PLOS ONE Published: December 19, 2016 https://doi.org/10.1371/journal.pone.0168496.

[16] K. Subhadra, N. Kavitha, “A Hybrid Leaf Disease Detection Scheme Using Grayco-Occurance Matrix Support Vector Machine Algorithm”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019.

[17] Upendra Kumar Shashank Yadav Esha Tripathi,” A Hybrid Approach for Automated Plant Leaf Recognition Using Hybrid Texture Features and Machine Learning-Based Classifiers”, International Journal of Distributed Artificial Intelligence (IJDAI) 13(2) 2021 DOI: 10.4018/IJDAI.2021070103.

[18] Punam BediPushkar Gole,” Plant disease detection using hybrid model based on convolutional autoencoder and convolutional neural network Artificial Intelligence in Agriculture”, Volume 5, 2021, Pages 90-101 https://doi.org/10.1016/j.aiia.2021.05.002.

Weighted Clustering for Deep Learning Approach in Heart Disease Diagnosis

BhandareTrupti Vasantrao[1]
Department of Computer Science and Engineering
Alliance College of Engineering and Design
Alliance University, Bangalore, India

Dr. Selvarani Rangasamy[2]
Alliance College of Engineering and Design
Alliance University, Bangalore
India

***Abstract*—An approach for heart diagnosis based on weighted clustering is presented in this paper. The existing heart diagnosis approach develops a decision based on correlation of feature vector of a querying sample with available knowledge to the system. With increase in the learning data to the system the search overhead increases. This tends to delay in decision making. The linear mapping is improved by the clustering process of large database information. However, the issue of data clustering is observed to be limited with increase in training information and characteristic of learning feature. To overcome the issue of accurate clustering, a weighted clustering approach based on gain factor is proposed. This approach updates the cluster information based on dual factor monitoring of distance and gain parameter. The presented approach illustrates an improvement in the mining performance in terms of accuracy, sensitivity and recall rate.**

Keywords*—Learning approach; weighted clustering; heart disease diagnosis; gain factor

I. INTRODUCTION

Heart diseases are rapidly increasing in recent past due to uneven living style and a highly variant environment conditions. The automaton of heart diagnosis for an early alarming is hence a primal need in today's leaving. The advancement in recent technologies in data mining has brought out new possibilities of early diagnosis and alarming of heart disorderness based on vital parameter analysis. The existing system uses the past learning parameters from a large database in making decision to monitoring values. However the rapid increases in data set and the availability of new data exchange and storage facilities constraint the mining performance due to large search overhead. Hence, an enchantment to the existing approach of data mining is required in making decisions to Heart diagnosis. In developing approach for automation in heart diagnosis and early alarming various methods were presented in past. In context to the development of learning approach for heart diseases in [1] a new classification model based on feature fusion model is presented. The features of vital parameters in heart diagnosis are fused with medical monitored parameters for improvising the accuracy. In [2] a semantic co-ranking process for heart disease diagnosis is presented. This approach is however limited with the diversity in data base entry. In [3] a fast convergence learning approach based on distance parameter is presented. In [4] a decision system based of Bergman divergence condition in training and testing for heart disease diagnosis is presented. The distance parameters in feature vectors are used in developing a decision. In [5,6] a matrix learning approach for co-ranking of data base features is presented. The approach outlines a random walk method in deriving the best possible decision for given test parameters. A modified divergence approach is proposed in [7]. This approach develops a hyper graph model for classification using learning of feature vector matrix. The diversity of feature vector in database however limits the proposed approach of classification. In [8] a relevance coding for heart disease diagnosis is presented. This approach develops a classification model based in the distribution of feature vectors in a wider domain. In [9, 10] a decision system based on personalized information in cardiovascular disorderness is presented. This approach addressed the variation parameter of learning feature in making decision. In [11] an approach of diagnosis for heart disease using semantic features is proposed. This approach defines a new relevance coding based on feature correlation and re-ranking operation in large dataset for heart diseases diagnosis. A fusion approach for process features based on the variation of feature vector is outlined in [12, 13]. The presented approach defines a correlative function model based on the magnitude parameter of vitals given. The issue of scaling and varying diversity is not been addressed in the presented approach. In [14] an approach of random updation and mapping for feature vector in heart disease diagnosis is presented. The presented approach limits the misclassification performance by a robust online diagnosis of feature vector. In [15] a naive Bayes method for heart disease diagnosis is presented. This proposed approach develops a classification approach based on hidden knowledge based on continuous data in the database. In [16] content based represented of vital parameters for heart disease diagnosis is presented. This approach monitors the variation among the feature parameters in developing a rank value for decision making. The increase in data base feature is observed to be effective in learning performance, however, the constraint of search overhead and misclassification performance under semantic feature limits the current mining approaches in early diagnosis for heart disease. To improve the performance based on search overhead, delay and accuracy metric a new weighted clustering approach based on cluster gain parameter is proposed. In [19] new MAE approach is developed to estimate the heart disease estimation using machine learning algorithms. For prediction of health system and risk different machine learning approaches are outlined in [20]. The existing approach of machine learning based on feature values used for training. The training overhead of the classification system is based on the number of learning features used. The more details of

feature value results to higher accuracy however with a cost of processing overhead. Towards minimizing feature overhead, in a recent approach [18] a feature fusion approach is developed based on data and feature values. The fusion model is developed based on the magnitude of each parameter in observation. The fusion model minimizes the processing overhead by reducing the number of training features; however the proposed approach is developed based on discrete monitoring value of the vitals. Wherein fusion model minimizes the processing overhead, the fusion of feature based on magnitude is limited with external distortion. A deviation in magnitude value due to a spike or jitter results to a misclassification. The effect of distortion is hence needed to be minimized to improve estimation accuracy. With the objective of improving the fusion accuracy and noise suppression a new time variant feature value with medical significance characteristic is proposed. This proposed approach presented a new mean of feature fusion using a time line monitoring of feature variation and observing the classification improved due to the fusion. The rest of this paper is outlined in six sections. The existing feature clustering approach and heart diagnosis approach is presented in Section II. Section III outlines the approach of proposed heart diagnosis using weighted clustering approach. Section IV presents the results obtained for the developed system. A discussion to the observations obtained is presented in Section V with conclusion briefed in Section VI.

II. Heart Disease Diagnosis

In an early detection of heart disease the learning features of data base has a vital role. In this presented work Cleveland dataset [17] is used for learning and analysis of the developed approaches.

A. Dataset Description

As observed, there are multiple observing parameters for diagnosis for heart diseases. In [17] multiple physiological parameters is presented for early diagnosis of heart disease. For the implementation of the proposed work Cleveland Dataset is used. Most of the researchers used this Cleveland Dataset as the benchmark dataset. The dataset consists of 76 attributes out of which majority of Computational Techniques have chosen only 14 attributes. The 14 attributes that we have considered along with their details are as follows:

1) Age: A slow varying parameter specifying patient age.

2) Sex: A non-varying parameter specifying male or female patient.

3) Chest pain (CP): this is of four types, Typical angina (angina), Atypical angina (abnang), Non-anginal pain (notang), or Asymptomatic (asympt) pain. The conditions of this feature are defined by the past monitoring of patient history and pain in chest with illness conditions.

4) Trestbps: the BP condition of a patient under rest condition.

5) Chol: Defines cholesterol level in mg/dl in a patient.

6) Fbs: Fasting blood sugar level.

7) Restecg: This gives the patient ECG under rest condition.

8) Thalach: This defines the maximum heart rate observed.

9) Exang: A Boolean parameter defining the induced angina in exercise.

*10)*Oldpeak: Depression value obtained during exercise.

*11)*Slope: The slop parameter defines the ST region during exercise.

*12)*Ca: Indicate numbers of major vessels defined by fluoroscopy.

*13)*Thal: Indicates the heart status condition of a patient.

*14)*Num: (class attribute) values are 0 for healthy and 1, 2, 3, 4 for unhealthy.

The Cleveland heart disease dataset has five class attributes indicating either healthy or one of four sick types.

In the existing approach the parameters are developed for feature fusion to reduce the processing overhead, where the features are fused based on the data level and feature level of the observation. Wherein the fusion is based on the numerical and normalized values of the parameters, the fusion are magnitude based. This fusion limits the observation to a discrete time observation which has the probability of error in real time diagnosis.

In this work a new feature fusion approach is proposed based on the time line characteristic of measuring parameter in consideration with medical significance of the feature. This approach fuse the features based on the medical significance and time line variation of the monitoring parameter which results in more accurate decision at a lesser processing time compared to existing approach.

A single feature taken in isolation cannot figure out all individuals' risk of heart disease. Hence many features are required to diagnose it. Among the observed parameters features such as cholesterol, heart rate, hypertension (blood pressure), resting ECG, diabetes, blood sugar, stress, exercise induced angina and old age are significant in predicting heart disease. Out of these factors, eight of them are medically significant from the list of Cleveland heart disease dataset [18]. They are age, chest pain type, resting blood pressure, fasting blood sugar, cholesterol, maximum heart rate, resting heart rate and exercise induced angina. If medically significant features were neglected, then it has every chance to run into the risk of incorrect diagnosis. Cleveland heart disease data has got both continuous and discrete types of data. In the classification stage a deep learning approach is made to cluster the observing parameters and perform classification based on the modified feature set. For this cause, the testing accuracy of medical feature fusion with a weighted clustering is compared with the existing approach. The accuracy, sensitivity, specificity, recall, precision and F-measure parameters were used in evaluation of the classification performance.

B. Diagnosis Process

The process of heart disease diagnosis is developed based on the training feature and the classification method applied. In the presented diagnosis system a set of feature from Cleveland dataset is used for diagnosis. The process of decision making is developed based on the normalized distance metric as outlined in [1]. For a set of feature vector (Fv) given as,

$$Fv = \{fv_{1,1}, fv_{1,2}, \ldots\ldots, fv_{nl,nl}\} \quad (1)$$

The classification is obtained using distance metric. The decision is derived using the minimization function of distance vector given as,

$$Selected\ distance\ (D) = \min(dist_i) \quad (2)$$

Where,

$dist$defines the eclidian distance faotrgivne as,

$$\text{Euclidian Distance, } dist_i = \sqrt{\sum_{i=1}^{n} Qf - Tf_i} \quad (3)$$

Where, Qf is query sample feature, and

$Tf$is the trained database feature.

A normalized distance factor for the decision using Max-Min criterion is used given as,

$$Decision(f_{ti}, f_{li}) = \frac{dist(f_{ti}, f_{li}) - dist_{min}(f_{ti}, F)}{dist_{max}(f_{ti}, F) - dist_{min}(f_{ti}, F)} \quad (4)$$

Here, 'dist' represents Euclidian distance of feature of test sample and database trained features. The distance on testing feature (f_{ti}), selected feature set (f_{li}) and database feature (F) is computed. d_{max}, d_{min} indicate maximum and minimum Euclidian distance of testing signal with dataset feature. The decision is develop as a normalized ratio of maximum and minimum distance parameter of testing and training feature vectors.

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

III. Weighted Clustering with Time Line Monitoring

The distance based classification has a limitation of arbitrary variation in the magnitude of feature vector and a large search overhead due to linear search over the database. To overcome the addressed issue, a weighted clustering approach based on gain parameter is presented. In the representation of database information in a classification process, the raw database is mapped to a normalized feature set which has a normalized effect of magnitude variation. The process of feature monitoring is presented in "Fig. 1".

The updation normalized feature is derived as a mean difference of feature vector.

For a set of randomly distributed feature vector $fv_{i,}$ the normalization is given by,

$$F(fv_1, fv_2, \dots\dots fv_n) = \sum_{i=1}^{n} \sum_{fv \in F} \left\| fv - \overline{fv} \right\|^2 \quad (5)$$

Here, fv is the feature vector in database and $\overline{fv}$ defines the mean of dataset given by,

$$\overline{fv} = \frac{1}{n} \sum_{fv \in F} fv \quad (6)$$

In the retrieval process the updated features are correlated over a set of test sample feature in making distance following minimum distance criterion. The illustration of search process is shown in "Fig. 2".

The distance vector is computed as a correlation metric C, given by,

$$C(fv, cl) = \sqrt{\sum_{i=1}^{n} (fv_i - cl_i)^2} \quad (7)$$

Where, the classification is developed based on the correlation of the feature vector ($fv_{i,}$) and selected feature cluster (cl_i) for i[th] test feature vector. However, the search overhead is considerably high for a large database. To minimize the effect a clustering approach is proposed. This presented approach cluster the available dataset into k-clusters based on the distance metric and the gain parameter. The cluster gain (CGn) attained due to updation in cluster is given by,

$$CGn = \frac{R(fv_i) - R(fv_i | ft_i)}{R(ft_i)} \quad (8)$$

Here, R defines the redundancy of a feature. The redundancy of a feature in a class is defined by the redundancy factor (R) given by,

$$R(f) = -\sum_{i=1}^{n} Pr(C_i) log_2\big(Pr(C_i)\big) \quad (9)$$

where $Pr(.)$ is the probability function, and $Pr(C_i)$ defines the probability of i[th] information in a clusterC_i.

The magnitude variation has a direct impact on the clustering performance hence a weighted updated of the feature vector is used given as,

$$Upt(C, fv) = \sqrt{\sum_{i=1}^{n} \omega_i (C_i - fv_i)^2} \quad (10)$$

The weight parameter defines the tuning of the computed gain for an information fv under magnitude variation given by,

$$\omega_{i+1} = \omega_i + dist(fv_{i,}, C_i) \quad (11)$$

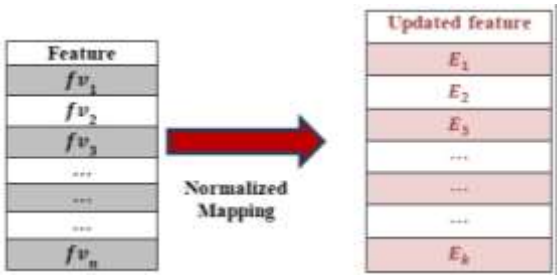


Fig. 1. Mapping Process for Feature Vector.

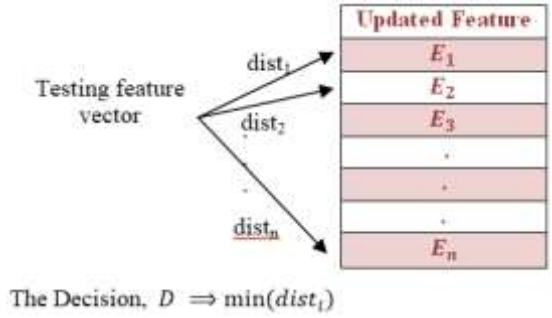


Fig. 2. Search Correlation over the Updated Dataset.

The weight parameter is updated by the distance parameter$dist(fv_i, C_i)$.

The convergence of the updation is given as a maximization function of cluster gain given as,

$$F_{convg} = \max\left(CGn_{\omega_i, fv_i, C)}\right) \quad (12)$$

The cluster values are updated based on the convergence of the feature value monitored for a period of time. For a time period the aggregated gain is computed and the cluster is updated based on the maximum cluster gain of a class. The classification of the developed system is performed using a multi class support vector machine (SVM). The decision is made for a set of training feature passed to the SVM architecture.

IV. Result Observation

The testing of developed approach is made over a Cleveland dataset where 1/3rd part of the database is used for training and remaining is used for testing. The approach presented is outlined in Matlab tool and validated for the retrieval Accuracy (Acc), sensitivity, specificity, Recall Rate, F-measure and computation time parameter. The parameters are computed as,

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

The sensitivity is given,

$$\text{sensitivity} = \frac{TP}{TP+FN} \quad (14)$$

The specificity is given as,

$$\text{specificity} = \frac{FP}{TN+FP} \quad (15)$$

Recall rate is given by,

$$Recall = TP/(TP+FN) \quad (16)$$

The precision is given as,

$$Precision = TP/(TP+FP) \quad (17)$$

And F_Measure is computed as,

$$F\_measure = (2 * Recall * Precision)/(Recall * Precision) \quad (18)$$

The observing parameters are developed with the observing factors listed in Table I.

The confusion matrix for the heart diagnosis is illustrated in Table II.

TABLE I. Observation Factor

(True Positive) TP	True match of test query feature and training feature
(True Negative) TN	False match of test query feature with true training feature
(False Negative) FN	False match of test query feature and training feature
(false positive) FP	True match of negative test query feature and training feature

TABLE II. Confusion Matrix for Heart Diagnosis

observation	True match	False match
True match	TP	FP
False match	TN	FN

The observation for the developed system is computed for different test cases measuring the evaluating parameters. The observation for the developed approach under different test cases of healthy, type-1, type-2, type-3 and type-4 attributes is listed in Table III.

The developed approach of weighted cluster based classification is compared with the existing approach of random clustering and linear searching method. Observation plots for the healthy case samples are shown in "Fig. 3-10".

"Fig. 3" shows the accuracy comparison obtained for the developed approaches in comparisons with the existing classification approach of linear search and random clustering. The linear search method correlates the test feature vector to the database feature in a linear manner and develops decision based on the maximum correlation value. The average distance of the mapping is considered in making the decision. Random clustering is developed based on the manual selection of features for classification. The set of features are manually picked from the test sample to pass to the classifier model for decision. The observation of accuracy illustrated an improvement of 8% as compared to random clustering and 18% as compared from linear search method.

The Sensitivity of the developed method is shown in "Fig. 4". The proposed method shows a Sensitivity of 91% wherein the existing approach of linear search and random clustering shows a Sensitivity of 87% and 77% respectively.

TABLE III. Analysis of Developed Classifications for Case Healthy Type

Method	Accuracy (%)	Sensitivity	Specificity	Recall	Precision	Time (sec)
Linear search method	70.42	0.77	0.45	0.77	0.79	0.64
Random cluster	86.09	0.91	0.48	0.91	0.92	0.43
Weighted Cluster	90.08	0.94	0.5	0.94	0.95	0.27

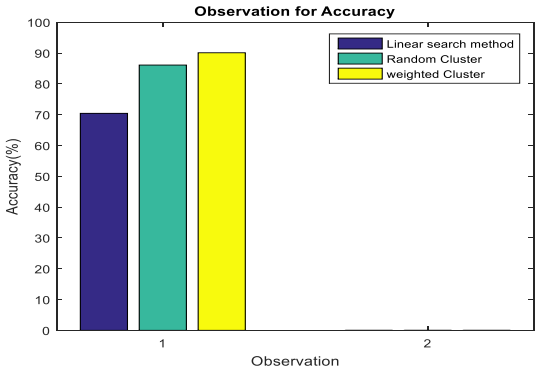


Fig. 3. Observation of Accuracy Plot for Healthy Case.

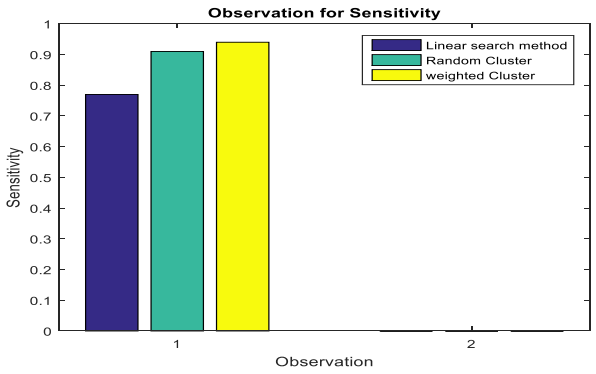


Fig. 4. Observation of Sensitivity Plot for Healthy Case.

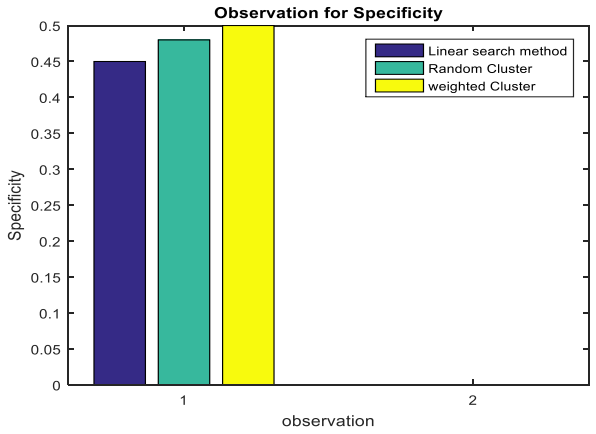


Fig. 5. Observation of Specificity Plot for Healthy Case

The specificity of the developed method is shown in "Fig. 5". The proposed method shows a specificity of 51%. The existing approach of linear search and random clustering shows a specificity of 46% and 44%, respectively.

The recall rate plot is shown in "Fig. 6". The proposed method obtain a recall rate of 91% and the existing approach of linear search and random clustering shows a recall rate of 89% and 77%, respectively.

The Precision of the developed method is shown in "Fig. 7". The proposed method shows a Precision of 93% wherein the existing approach of linear search and random clustering shows a Precision of 1% and 79% respectively. An analysis of the developed approach for different test types is summarized in Table IV.

Processing time defined as the computation time for classification is shown in "Fig. 8". The proposed method takes a computation time of 0.26 sec in classification, wherein the existing approach of linear search and random clustering observe a computation time of 0.43 and 0.65 sec, respectively.

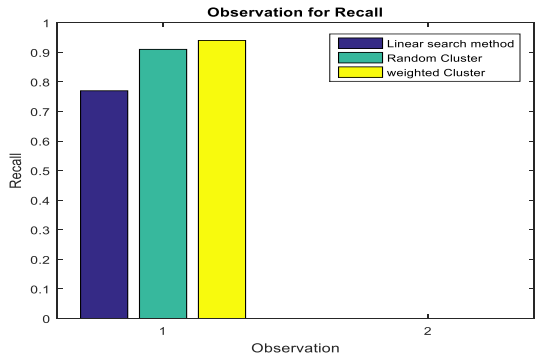


Fig. 6. Observation of Recall Plot for Healthy Case.

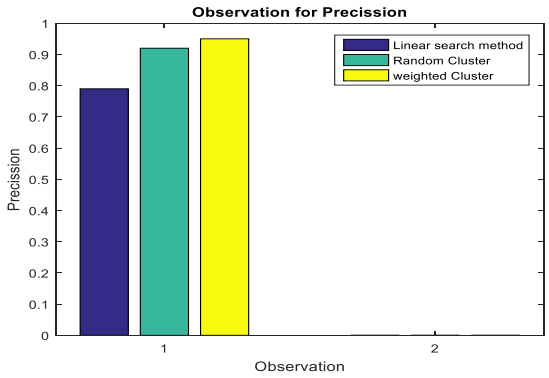


Fig. 7. Observation of Precision Plot for Healthy Case.

TABLE IV. ANALYSIS OF DEVELOPED CLASSIFICATIONS FOR CASES TYPE-1-4

Case Type	Method	Accuracy	Sensitivity	Specificity	Recall	Precision	Time (sec)
Type - 1	Linear search method	77.35	0.82	0. 81	0.87	0.52	0.64
	Random cluster	85.22	0.91	0.84	0.87	0.421	0.43
	Weighted Cluster	89.08	0.93	0.95	0.98	0.745	0.27
Type - 2	Linear search method	83.35	0.89	0.78	0.77	0.65	0.58
	Random cluster	86.22	0.86	0.84	0.87	0.72	0.41
	Weighted Cluster	90.08	0.94	0.95	0.91	0.85	0.20
Type - 3	Linear search method	79.35	0.86	0.82	0.83	0.77	0.78
	Random cluster	86.22	0.91	0.84	0.89	0.84	0.56
	Weighted Cluster	91.08	0.95	0.91	0.93	0.89	0.31
Type - 4	Linear search method	63.35	0.67	0.78	0.79	0.78	0.81
	Random cluster	88.22	0.88	0.84	0.88	0.90	0.53
	Weighted Cluster	92.08	0.95	0.92	0.91	0.94	0.33

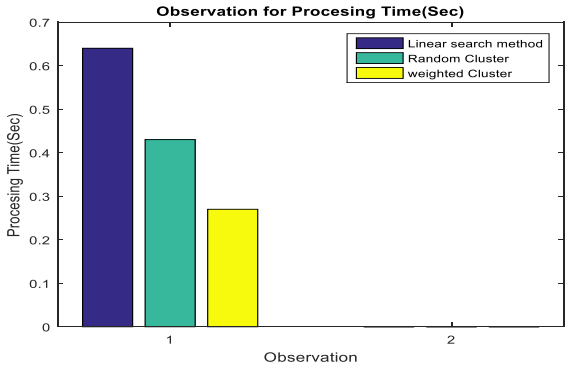


Fig. 8. Observation of Processing Time (Sec) Plot for Healthy Case

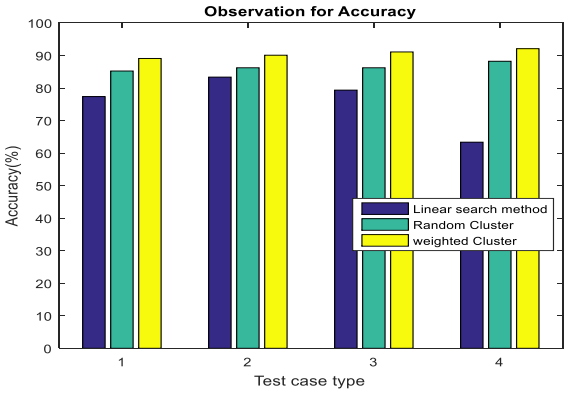


Fig. 9. Observation of Accuracy Plot for Type 1-4 Case.

Observation for four types of effects refereed in Cleveland data set is compared for accuracy. The observation plot for accuracy for the three developed classification approach shown in "Fig. 9". The observation shows an accuracy of 89% for type-190% for type-2, 91% for type-3 and 92% for type-4 cases.

Observation plot for Sensitivity for the three developed classification approach shown in "Fig. 10". The observations show a Sensitivity of 93% for type-1, 94% for type-2, 95% for type-3 and 95% for type-4 cases.

For Specificity for the three developed classification approach is shown in "Fig. 11". The observations show a Specificity of 95% for type-1, 95% for type-2, 91% for type-3 and 92% for type-4 cases.

Recall for the three developed classification approach is shown in "Fig. 12". The observations show a Recall of 98% for type-1, 91% for type-2, 93% for type-3 and 91% for type-4 cases.

Precision for the three developed classification approach is shown in "Fig. 13". The observations show a precision of 74% for type-1, 85% for type-2, 89% for type-3 and 94% for type-4 cases.

Processing time for the three developed classification approach is shown in "Fig. 14". The observations show a Processing time of 0.27 sec for type-1, 0.20sec for type-2, 0.31sec for type-3 and 0.33 sec for type-4 cases.

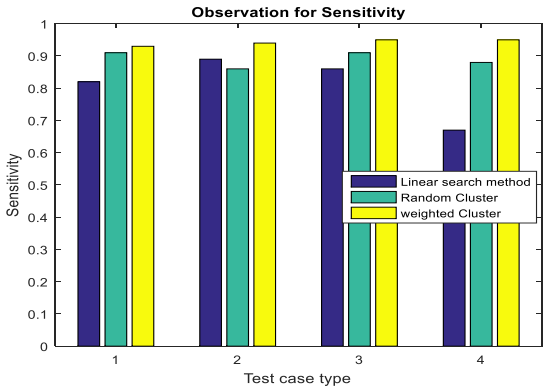


Fig. 10. Observation of Sensitivity Plot for type 1-4 Case.

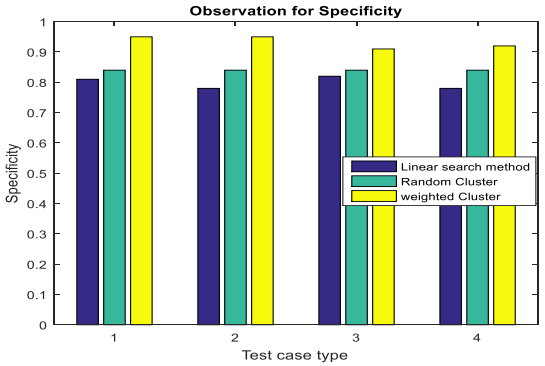


Fig. 11. Observation of Specificity Plot for type 1-4 Case.

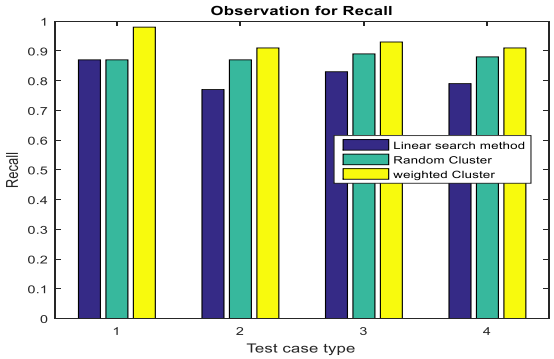


Fig. 12. Observation of Recall Plot for Type 1-4 Case.

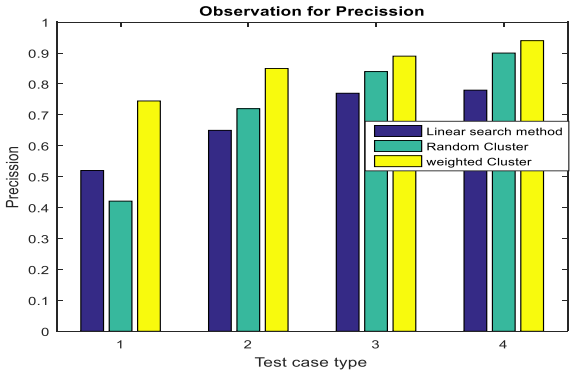


Fig. 13. Observation of Precision Plot for Type 1-4 Case.

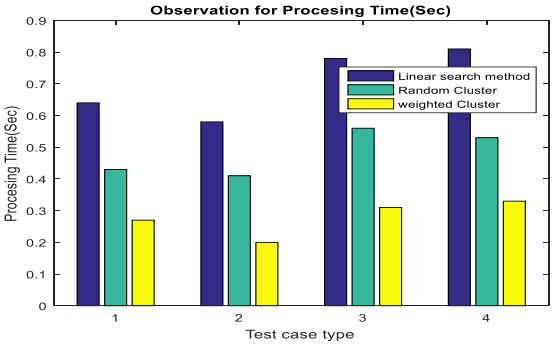


Fig. 14. Observation of Processing Time Plot for Type 1-4 Case.

V. Discussion

Diagnosis of heart disease based on the vitals measured is a primary need for automated diagnosis and classification system. In the presented work a weighted clustering approach for feature fusion is proposed. The analysis of the proposed approach for classification accuracy, sensitivity, specificity, recall, precision and computation time resulted in more efficient observations compared to the existing approach of linear search and random clustering method. The process of time line monitoring and denoising results in faster processing resulting g in a minimization of 0.4sec in processing time compared to linear search method. The accuracy of classification is observed to improve by 20% compared to linear search method. For varying test type parameters from 1-4 it is observed that higher accuracy for type 4 case is observed. The test time is lower for type-1 case. Wherein the classification due to linear search has high search time, clustering based approach is suitable in minimizing the processing time with higher accuracy. The random clustering selects the features from the list of observation based on the magnitude values. The distortion impacts the retrieval accuracy. However, the proposed approach illustrates higher classification accuracy with lower time of processing. The analysis is focus on testing the proposed method onto real time vital parameters as future work.

VI. Conclusion

The diagnosis of heart disease using weighted clustering is presented in this paper. The approach of weight updation and clustering of large database feature based on cluster gain factor is proposed. This approach developed a new clustering approach of a distributed feature set in Cleveland data set into specified cluster to improve the classification performance. The search overhead in mining of features for classification is minimized by the formations of sub clusters. The decision approach of feature update into cluster is made more accurate with the dual factor of class attribute and cluster gain factor. The observations of the developed system illustrated an improvement in accuracy for the proposed system compared to existing approach with a reduced time of computation due to cluster search approach.

References

[1] Farman Ali, Shaker El-Sappagh, S.M. Riazul Islam , Daehan Kwak, Amjad Ali f , Muhammad Imran g , Kyung-Sup Kwak h , “A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion,” Information Fusion, No.63, pp-208–222, 2020.

[2] Rohé M M, Datar M, Heimann T, et al. “SVF-Net: Learning deformable image registration using shape matching, International Conference on Medical Image Computing and Computer-Assisted Intervention,” Springer, Cham, 2017: 266-274.

[3] Gao Z, Xiong H, Liu X, et al., “ Robust estimation of carotid artery wall motion using the elasticity-based state-space approach,” Medical image analysis, 2017, 37: 1-21.

[4] Zhen X, Zhang H, Islam A, et al., “Direct and simultaneous estimation of cardiac four chamber volumes by multioutput sparse regression,” Medical image analysis, 2017, 36: 184-196.

[5] Gao Z Zhao S, Gao Z, Zhang H, et al., “Robust Segmentation of Intima–Media Borders With Different Morphologies and Dynamics During the Cardiac Cycle,” IEEE journal of biomedical and health informatics, 2017, 22(5): 1571-1582.

[6] E. J. Benjamin, S. S. Virani, C. W. Callaway, A. R. Chang, S. Cheng, S. E. Chiuve, M. Cushman, and F. N. Delling, et. al., "Heart disease and stroke statistics-2018 update: a report from the American Heart Association," Circulation, Vol. 137, No. 12, pp. 67- 492, 2018.

[7] Xu L, Huang X, Ma J, et al., “Value of three-dimensional strain parameters for predicting left ventricular remodeling after ST-elevation myocardial infarction,” The International Journal of Cardiovascular Imaging, 2017, 33(5):663-673.

[8] Thiago M. Nunes, Victor Hugo C. de Albuquerque,et. al., “Automatic microstructural characterization and classification using artificial intelligence techniques on ultrasound signals,” Expert Systems with Applications, 2013, 40(8):3096-3105.

[9] Khamparia A, Saini G, Gupta D, et al.,“Seasonal Crops Disease Prediction and Classification Using Deep Convolutional Encoder Network,” Circuits, Systems, and Signal Processing, 2019(May).

[10] Albuquerque V H C D, Nunes T M, Pereira D R, et al.,“Robust automated cardiac arrhythmia detection in ECG beat signals,” Neural Computing & Applications, 2018, 29(3):1-15.

[11] W. H. Organization and others, “World health statistics 2017: monitoring health for the SDGs, sustainable development goals,” 2017.

[12] R. I. Kementrian Kesehatan, “Riset kesehatan dasar,” Jakarta Kementerian. Kesehatan. RI, 2013.

[13] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, “Time series classification using multi-channels deep convolutional neural networks,” in International Conference on Web-Age Information Management, 2014, pp. 298–310.

[14] D. R. Chowdhury, M. Chatterjee, and R. K. Samanta, “An artificial neural network model for neonatal disease diagnosis,” Int. J. Artif. Intell. Expert Syst., vol. 2, no. 3, pp. 96–106, 2011.

[15] G. Subbalakshmi, K. Ramesh, and M. C. Rao, “Decision support in heart disease prediction system using naive bayes,” Indian J. Comput. Sci. Eng., vol. 2, no. 2, pp. 170–176, 2011.

[16] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, “Computational intelligence for heart disease diagnosis: A medical knowledge driven approach,” Expert Syst. Appl., vol. 40, no. 1, pp. 96–104, 2013.

[17] https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[18] Jesmin Nahar, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Chen, “Computational intelligence for heart disease diagnosis: A medical knowledge driven approach,” Expert Systems with Applications, No.40, pp- 96–104, 2013.

[19] Santosh A. Shinde, Dr. P. Raja Rajeswari, “A Novel Hybrid Framework for Cuff-Less Blood Pressure Estimation based On Vital Bio Signals processing using Machine Learning,” International Journal of Advanced Trends in Computer Science and Engineering, Vol.9, Issue 2, March-April 2020, pp. 1556-1561.

[20] Mr. Santosh A. Shinde , Dr. P. Raja Rajeswari, "Intelligent health risk prediction systems using machine learning: a review," International Journal of Engineering & Technology, Vol.7, Issue 3, 2018, pp. 1019-1023.

Research Efforts and Challenges in Crowd-based Requirements Engineering: A Review

Rosmiza Wahida Abdullah[1], Sabrina Ahmad[2]*, Siti Azirah Asmai[3]
Fakulti Teknologi Maklumat Dan Komunikasi
Universiti Teknikal Malaysia Melaka
Malaysia

Seok-Won Lee[4]
Dept. of Software and Computer Engineering
Dept. of Artificial Intelligence
Ajou University, Suwon, South Korea

Zarina Mat Zain[5]
Digital Group
PETRONAS Digital Sdn. Bhd
Kuala Lumpur, Malaysia

***Abstract*—Eliciting software system development requirements is a challenging task as the information is from various resources. The most constructive resource is the stakeholders of the system to be developed. It is critical yet time-consuming to capture essential requirements to realize a reliable and workable software system. The crowd-based Requirements Engineering (crowd-based RE) approach adapts the crowdsourcing technique to access an extensive range of stakeholders and save time, especially for the generic type system with no clear stakeholder. This paper presents current research efforts and challenges in crowd-based RE. A systematic literature review method is adopted to explore literature based on two specific research questions. The first question aimed at identifying research efforts on crowd-based RE, and the second question focused on the main challenges discovered in pursuing crowd-based RE. The findings from the literature review show that many efforts have been made to explore and further improve crowd-based RE. This paper provides a foundation to pursue research in improving crowdsourcing techniques for the benefit of requirements engineering.**

Keywords*—Crowd-based requirement engineering; requirements engineering; requirements elicitation; software engineering; crowdsourcing; review

I. Introduction

Requirement Engineering (RE) is the first and the most crucial phase in a software development project, and the process must be performed to ensure quality software requirements. A study [7] stated that a poorly engineered requirements process contributes immensely to the failure of software projects. It is also said that projects that undermine RE suffer or are likely to suffer from failures, challenges, and other risks [36].

Requirements collected must be correct, complete, and concise to ensure the success of the developed software system. To do that, requirement engineers need to specify the stakeholders and ensure they participate in providing the requirements [3]. The process is challenging as it needs to gather and translate the imprecise, incomplete needs and wishes of the stakeholders into complete, precise, and formal specifications. In the case of requirements from the crowd being welcome and deemed helpful in ensuring preferred system features are incorporated, the crowdsourcing technique is beneficial. The term crowdsourcing is introduced to portray the concept of outsourcing that describes a distributed problem-solving approach online with a large number of people [1]. Due to the advancement in Internet technology, crowdsourcing is now an emerging technique that has been actively studied and adapted in various domains such as software engineering, social innovation, and education. While the crowdsourcing technique is gaining popularity in multiple domains [2], requirements engineering should also benefit since crowdsourcing makes it possible to reach out to many stakeholders to voice out their needs and expectations towards a particular software system.

However, there is always a catch to benefit from such an emerging technique. While much information is good, issues like overloading, coverage, unknown source, and unreliable information need to be taken care of while eliciting requirements from the crowd. How are we going to ensure that the information we receive is enough? Is it from reliable resources? Is the information meaningful? Is it reliable? We have to deal with these challenges while eliciting requirements through crowdsourcing techniques since the Internet has no boundary.

Therefore, the ultimate aim of this paper is to provide an insight for further exploration and contribution towards strengthening crowd-based requirements engineering. The expected contributions of this research are:

1) To discover research efforts on crowdsourcing that have been done to empower requirements engineering in the years range from 2008 until 2021. We intend to discover the RE activities supported by crowdsourcing.

2) To present the chronology of research efforts to recognize issues and improvement in crowd-based requirements engineering thus far. The findings will be helpful to identify the research gaps that form a basis for future crowd-based RE research.

*Corresponding Author

(Universiti Teknikal Malaysia Melaka sponsors this paper through a research grant numbered PJP/2020/FTMK/PP/S01774)

Following the introduction, Section II explains the background of the study. This is followed by Section III, which elaborates on the systematic literature review method. Section IV presents the review results, and Section V elaborates on the discussion. Finally, Section VI concludes the paper.

II. Background

Traditional RE adopts conventional techniques such as interviews, surveys, document analysis, workshops, and brainstorming. To elicit requirements through these traditional techniques are challenging and costly in term of time and effort. The chances to miss out on essential requirements from the key stakeholders are also very high due to resource constraints to implement adequate RE. Therefore, many types of research are conducted to enhance the user involvement in the RE process within limited resources [6].

In line with the available Internet technology and how information is exchanged nowadays, it is only reasonable that the RE techniques have evolved. Besides, people are now very much exposed to doing things online, from communication to paying bills and even controlling smart facilities from a distance. The rapid rise in Internet, mobile and social media applications makes it even more possible to provide channels to link a large pool of highly diversified and physically distributed stakeholders, especially potential users, for the system to be developed [5].

Crowdsourcing is an evolving paradigm that provides help to gather enormous and functional software requirements. Crowdsourcing makes it possible to reach out to many stakeholders to offer or voice their needs and expectations towards a particular software system. By adopting crowdsourcing, we reduce the risk of missing essential requirements from specific key stakeholders. In [1], J. Howe introduced the term crowdsourcing, adapted from the concept of outsourcing that describes a distributed problem-solving approach online with the involvement of a large number of people. In [34], M. Hosseini et al. mentioned four critical features of crowdsourcing: the crowd; people participating in the crowdsourcing activity, the crowdsourcer; the party that owns the task, the crowdsourcing task, and the crowdsourcing platform; the setting where the mission is accomplished. Crowdsourcing is gathering works, information, and opinions from the public through the Internet, social media, and smartphone apps [8]. According to U.S. Ghanyni et al. in [3], crowdsourcing offers a wide range of expertise and talents, making it the best way to collect requirements and improve user involvement. In 2015, the term Crowd-Based Requirements Engineering, also known as CrowdRE, was coined [10]. After that, in [9], E.C. Groen et al. defined CrowdRE as an umbrella term for automated or semi-automated RE approach for gathering and analyzing information from a crowd to derive validated user requirements.

Due to the numerous benefits of crowdsourcing, crowd-based RE is becoming popular and a meaningful way to be applied in the RE process, especially in requirement elicitation activity. This is because every stakeholder will get the opportunity to propose their expectations of the software [11]. Hence, the gathered requirements will be complete in representing sufficient stakeholders' perspectives and perceptions compared to limited input from selected stakeholders.

In line with that, J.A. Khan et al. in [5] supported the fact that there is a growing interest in crowd-based RE. Therefore, further research to improve the crowd-based RE is relevant for better service to the software engineering community.

III. Methodology

This section describes our literature survey process based on a systematic literature review method [12], which drives research questions through searching, filtering, and analysis processes. The literature exploration is presented through two research questions.

A. Research Questions

Due to the growing interest in crowd-based RE, this paper presents current research efforts and challenges to explore further opportunities to improve RE through crowdsourcing. The research questions addressed by this study are as follow:

RQ 1 What researches have been done in crowdsourcing for RE?

To answer RQ1, we conduct a literature review aimed at identifying research efforts on crowd-based RE.

RQ 2 What are the challenges and limitations of current research in crowd-based RE?

To answer RQ2, we look at the issues encountered in the recent crowd-based RE researches.

B. Search Process

This sub-section explains the searching strategies of this literature survey. The search is done manually through popular and familiar digital libraries and databases as listed below:

1) IEEE Xplore (ieeexplore.iee.org).
2) ScienceDirect (sciencedirect.com).
3) Springer (Springerlink.com).
4) Google Scholar (scholar.google.com).
5) ACM Digital Library (dl.acm.org).

The searching included leading conferences, workshops, and journals that meet the search criteria. The search strings are based on the research questions and relevant keywords related to search areas such as requirements engineering, crowdsourcing, crowd-based and crowd-centric. We are aware that many articles about this topic are also posted on blogs, magazines, and newspapers, but we only focus on academic publications for this literature review. Besides, only papers written in English are covered.

C. Inclusion and Exclusion Criteria

Both research questions were answered by searching relevant research papers through meaningful keywords. The keywords from the primary studies were used to find more articles related to the research. Also, synonyms and alternative words were used to optimize the search of related works. The

general keywords used to search the associated articles were "crowd* AND requirement*." We used the combination crowd* AND requirement* search term to ensure we managed to obtain as many relevant results as possible. The search gave us various reliable journals and conference proceedings covering issues in crowdsourcing for requirements engineering. Fig. 1 shows the number of research articles from 2008 until 2021. It shows the ascending pattern in crowd* AND requirement* search terms. Hence, we can conclude that many researchers are interested in this area, and crowdsource in RE is gaining popularity year by year.

Upon completing the searching process, we filtered the findings to related works only. We have included 20 primary studies that proposed approaches to automate the RE activities through crowdsourcing techniques. Fig. 2 shows the distribution over the years the studies have been published. Referring to our search, no effort has been proposed in 2009, 2013, and 2016, but the number of proposed efforts spiked in 2019. In 2020, however, only one proposed effort was discovered. Fig. 2 shows that researchers never stop exploring this area and proposed solutions that make use of the advantages in crowdsourcing to overcome or at least minimize problems in RE. Therefore, we believe that it is worth the effort to explore this area for the benefit of RE.

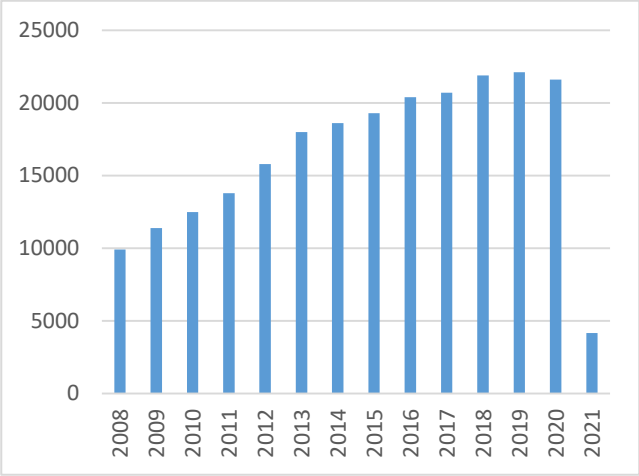


Fig. 1. Number of Research Articles.

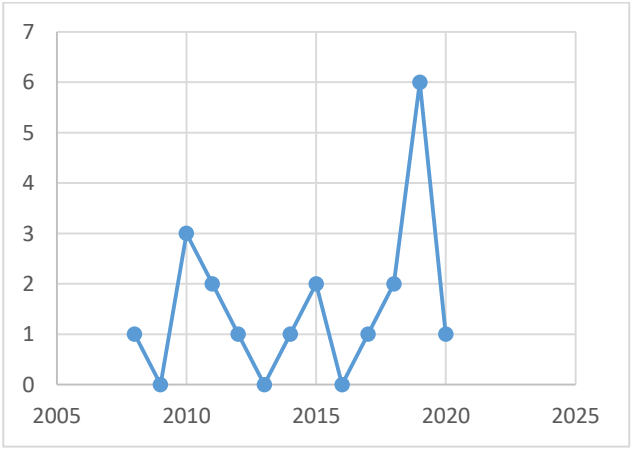


Fig. 2. Efforts on Approaches to Automate the RE Activities.

IV. Review Results

A. Research Question 1: What Researches have been done in Crowdsourcing for RE?

As we know, the ultimate idea of the crowd-based RE approach is to obtain input or feedback from the crowd who uses the software [35]. The crowdsourcing technique allows access to diverse stakeholders and can gain broader and up-to-date information about users' expectations toward the system to be developed [33].

To answer this research question, we conduct a literature review to identify the RE activities supported by crowdsourcing. We reviewed 20 primary researches that related to crowd-based RE efforts. Table I provides an overview of the efforts of crowd-based RE approaches, and the table also covers the RE activities supported by the efforts.

TABLE I. Crowd-based RE Approaches/Tools/Platform

Approach/ Tool/ Platform	Description	Supported RE Activity	Source
WikiWinWin	Wiki-based system for collaborative requirements negotiation	• Elicitation • Analysis • Specification	[13]
StakeNet	Using social networks to analyze the stakeholders of large-scale software projects	• Analysis	[14]
StakeSource	A web-based tool that automates stakeholder analysis	• Analysis	[15]
StakeSource 2.0	A web-based tool that utilizes social networks and collaborative filtering to identify and prioritize requirements	• Analysis	[16]
iRequire	An application that gathers requirements to develop new mobile apps	• Elicitation • Analysis	[17]
StakeRare	Use social network and collaborative filtering for large-scale requirements elicitation	• Analysis	[18]
CrowdREquire	A web-based platform that supports RE using crowdsourcing concept	• Elicitation • Validation	[4]
AppEcho	A user feedback approach for mobile platforms and applications	• Elicitation	[19]
Refine	A platform that uses gamification for participatory requirements engineering	• Elicitation • Analysis	[20]
FAME	a framework for the combined and simultaneous collection of feedback and monitoring data in web and mobile contexts to support continuous requirements elicitation	• Elicitation	[21]

CRUISE	➢ A platform for crowdsourcing Requirements Elicitation and evolution	• Elicitation	[22]
Effective classification methodology	An approach to classifying user requests into corresponding requirement types	• Elicitation • Analysis	[11]
GARUSO	➢ a gamification approach for involving stakeholders outside organizational reach in requirements engineering.	• Elicitation • Analysis	[23]
CREeLS	A framework of the necessary elements of crowdsourcing suggesting specific tools for each component, and a phased approach to implementing the framework in the requirement elicitation activity for eLearning System	• Elicitation • Analysis	[24]
SUPERSEDE	provides advances on end-user feedback, contextual data analysis, and decision-making support in software evolution and adaptation	• Elicitation • Analysis	[25]
Classification model	The proposed classification model collects feedbacks from the crowd, and the feedback will be analyzed by applying data mining techniques. Finally, the proposed model will classify the feedbacks into functional and non-functional requirements.	• Elicitation • Analysis	[28]
Continuous Requirements-Elicitation Methodology	The proposed methodology captures and analyzes user feedback and comments on social networks such as Twitter for software systems in use and then extracts potential requirements.	• Elicitation • Analysis	[29]
Automated Feature Identification	The proposed solution is a tool-supported approach that analyzes apps and associated feedback on the feature level. This information is used to inspire the development of new apps and, in particular, to suggest features for developing similar new apps.	• Elicitation • Analysis	[31]
Crowdsourced RE Platform for User Story (US) Authoring	investigate how a crowdsourced RE platform can enable the crowd to provide requirements by using one of the RE artifacts' type, the US.	• Elicitation	[32]
KMar Crowd	a CrowdRE platform called the KMar Crowd is applied in governmental organizations to discover the needs and wishes of user groups for a particular IT product.	• Elicitation	[30]

WikiWinWin was proposed by [13], which adapts the Wiki-based system that allows anybody to provide input to the platform. There are two types of primary users which are Shapers and Personal Knowledge Contributors (PKC). Shapers are the skilled stakeholders who contribute ideas, motivate PKC to express ideas, moderate the negotiation process, integrating, filtering, organizing, and rewriting contributions of others. While PKCs are the participants who contribute ideas and negotiate win conditions. The participants have to be invited to join in. At the end of the process, a software requirement description is produced.

StakeNet, StakeSource, and StakeRare concerned with stakeholder analysis. These tools are to carefully filter the stakeholder that participates in the project to contribute input for the requirements. Stakeholders are the source of the requirements, and we do not want to miss out on any crucial stakeholders to ensure the project's success [15].

iRequire and AppEcho enable mobile phone users to contribute feedback. These tools concern the end-user's involvement in obtaining the input and the context of the gathered information. For these tools, anybody who uses a mobile phone may participate in contributing the feedbacks.

Feedback Acquisition and Monitoring Enabler (FAME) also uses feedback and monitors the information to elicit new requirements. This tool is more to obtaining requirements for software evolution. FAME was developed as part of the SUPERSEDE EU project.

REfine, CrowdREquire, and CRUISE are concerned with stakeholder analysis and obtaining input from the invited stakeholder. Stakeholders can only participate through an invitation from the project owner. These tools incentivize the participant to motivate them to keep contributing to the project. REfine, and CRUISE applies game element while CrowdREquire offers a financial reward.

GARUSO approach uses a strategy for identifying stakeholders outside the organizational reach and a social media platform that applies gamification for motivating these stakeholders to participate in RE activities.

The SUPERSEDE (Supporting Evolution and Adaptation of Personalized Software by Exploiting Contextual Data and End-User Feedback; supersede.eu) project is developing multimodal-feedback functionalities that will let a crowd of users provide unobtrusive in situ feedback on software products. A runtime approach establishes comprehensive techniques to monitor software products and obtain environmental and context data through sensors. The received feedback and data will be analyzed to identify relevant information to support decision-making during software evolution. Informed decisions based on the feedback and monitoring data will lead to products that better meet user needs and improve the user experience.

The CREeLS, an effective classification methodology, and classification model are concerned with classifying the requirements. CREeLS adapt the approach proposed in the effective classification methodology, especially for the eLearning system. The classification model collects feedbacks from the crowd and then classifies the feedbacks into functional and non-functional requirements. CREeLS, effective classification methodology, and classification model apply text mining tools to analyze unstructured text because they use feedback as the input, which is usually natural language.

Continuous Requirement Elicitation Methodology and Automated Feature Identification also capture and analyze feedback from the crowd. Continuous Requirement Elicitation Methodology captures and analyzes user feedback and comments on social networks such as Twitter for a software system currently in use and then extracts the potential requirements. Automated Feature Identification examines feedbacks of mobile apps from the crowd to identify features for developing similar new apps.

Crowdsourced RE Platform for User Story (US) Authoring applies one of the RE artifacts: User Story. This research investigates how a crowdsourced RE platform can enable the crowd to provide requirements through four simple self-explanatory steps: Role, Goal, Benefit and Verification, and Category Selection.

KMar Crowd is a crowd-based RE platform applied in governmental organizations to identify the users' needs and wishes for a particular IT product. When this article is written, the researcher of KMar Crowd does not reveal how the platform works and what type of information is gathered from the crowd.

All the research efforts mentioned in this section applied diverse techniques and approaches to improve crowd-based RE in a specific area. More research should explore ways to utilize crowdsourcing and further improve RE.

B. Research Question 2: What are the Challenges and Limitations of Current Research in Crowd-based RE?

To answer this research question, we look into issues discovered in the current researches in crowd-based RE as listed in Table I. Through crowd-based RE, we can access a large pool of stakeholders to achieve the breadth of the requirements. However, some challenges need to be taken care of to guarantee the success of achieving the breadth of the requirements. As stated by D. Johnson et al. in [26], crowd-based RE has been argued to comprise four main activities: motivating crowd members, eliciting feedback, analyzing feedback, and monitoring context and usage data. These are essential elements to ensure that the information collected covers enough perspectives and to ensure if the information is reliable. In general, we found two main challenges in the existing research: stakeholders' coverage and information reliability.

The following sub-section is presented narratively to show efforts evolution to overcome the challenges:

1) Stakeholders coverage: Do we cover enough perspectives? The challenge here is whether or not the information obtained represents enough perception and perspectives to develop a quality system to fulfill the system's purposes. Discussed below are research efforts to improve stakeholders' involvement to improve the coverage.

WikiWinWin, proposed by [13], provides a platform for the stakeholder to vote and decide on the software requirements. As the stakeholder participation only through invitation, the issue of missing key stakeholders is still there. Moreover, the stakeholder who participates in a specific project must understand well about the project they are participating in to make sure they provide ideas according to the context of the project. Other than that, the stakeholder needs to vote for the ideas to make it a requirement. If the idea is not getting many votes, it will not be considered a requirement. Thus, the stakeholders involved in the project must understand and be well aware of the expectation for the software to be developed. It is indeed crucial to establish the right system and, at the same time to fulfill the end-users need. Therefore, in WikiWinWin, the challenge is to select the right and sufficient stakeholders to participate. Besides that, it is also a challenge to keep the stakeholders motivated to provide ideas and input to the project.

Many software projects fail because they overlook stakeholders or involve the wrong representatives of significant stakeholders' groups [14]. Knowing the importance of obtaining correct stakeholders in the software development project, S. L. Lim et al. in [14] proposed a tool called StakeNet for stakeholders' analysis. StakeNet requires experts to identify stakeholders, and then, the experts have to ask them to recommend other stakeholders individually. Consequently, a social network of stakeholders based on their recommendations will be built. The prioritization of the stakeholders is decided by using various social network measures. However, this tool will be very costly for a large project which involves many stakeholders since it requires the experts to approach stakeholders individually to ask for recommendations.

Aware of this issue, StackSource was introduced [15]. StackSource is a web-based tool that automates stakeholder analysis. StackSource identifies stakeholders by asking them to recommend other stakeholders, builds a social network of stakeholders from their recommendations, and prioritizes them using social network measures. Soon after that, S. L. Lim et al. in [16] proposed an enhanced version of the StackSource tool StackSource2.0. Besides stakeholders' analysis, this improved tool is introducing another feature to do requirement elicitation and prioritization. In the requirements elicitation and prioritization feature, the tool can identify requirements by asking stakeholders to suggest and rate the requirements, recommend other requirements of interest using collaborative filtering, and prioritize the requirements using their ratings weighted by their priority in the social network.

Later in 2012, S. L. Lim and A. Finkelstein [18] proposed a method known as StakeRare that uses social networks and collaborative filtering to identify and prioritize requirements in large software projects. StakeRare identifies stakeholders and asks them to recommend other stakeholders and stakeholder roles, builds a social network with stakeholders as nodes and their recommendations as links, and prioritizes stakeholders

using various social network measures to determine their project influence. It asks the stakeholders to rate an initial list of requirements, recommends other relevant requirements using collaborative filtering, and prioritizes their requirements using ratings weighted by their project influence. Recently, an approach named GARUSO has been proposed [23] to identify stakeholders outside the organizational reach. It is a social media platform that applies gamification to motivate related stakeholders to participate in RE activities. Compared to StakeNet, StakeSource, Stake Source2.0, and StakeRare, GARUSO is claimed to reach potential stakeholders from multiple online channels such as e-mail services and SNSs to identify stakeholders of a software system who are beyond the reach of an organization.

CrowdREquire, REfine, and CRUISE adapt stakeholder analysis and requirement elicitation. Stakeholders can only participate through an invitation from the project owner. These tools provide incentivize to motivate the participants to keep contributing to the project. REfine, and CRUISE adapts gamification while CrowdREquire offers a financial reward. Giving incentives and rewards to encourage the participation of the stakeholders may cause malicious and dishonest input. This is because the stakeholders might vote or offer information to gain reward, leading to incorrect requirements.

2) Reliable information: While much information is achievable through crowdsourcing, is the information useful? It is common knowledge that stakeholders especially end users are among reliable information sources from whom requirements are elicited. Traditionally, interviews, workshops, brainstorming, and survey will be conducted among the end-users to obtain requirements. It is clearly stated that software system users are an essential group of stakeholders, as reported by M. Bano and D. Zowghi [27]. End users' involvement in software development life cycle (SDLC) has been suggested to improve requirements' quality, accuracy, and completeness to ensure users' satisfaction. Presents below are research efforts that capture end-users input through crowdsourcing and introduce ways to ensure that the input is reliable.

Earlier in 2010, N. Seyff et al. in [17] stated that end-users involvement is particularly relevant for early software engineering activities such as requirements elicitation. In that particular study, iRequire is introduced to capture end-user requirements for mobile. In 2014, an app named AppEcho was introduced [19]. This app is a feedback approach that enables users to give feedback through the android platform. The method allows smartphone users to actively participate in continuous evolution and improvement by providing individual feedback to developers.

Furthermore, FAME was also introduced [21] to collect users' feedback on the software product. It is a stand-alone feedback app for mobile devices. FAME was developed as part of the SUPERSEDE EU project. SUPERSEDE project is a runtime approach to collect and analyze user feedback. It is also managed to identify relevant information to decide the essential requirements for the next release of a product. In 2019 Continuous Requirement Elicitation Methodology [29] and Automated Feature Identification [31] were proposed. Both of these researches use feedbacks as the primary source for the information. A study conducted by A. Alwadin and M. Asharagi in [29] collected feedback and comments from the crowd via Twitter for an in-use software system. They applied data retrieval and natural language processing (NLP) techniques to extract potential requirements. Automated Feature Identification was proposed by T. Iqbal et al. in [31] to identify features for developing new mobile apps. This research applied the app store mining technique, exploring crowd-generated data such as feedback on the existing apps to identify critical elements for creating new apps. Machine Learning is used to analyze the input.

Another proposed solution introduced by C. Li et al. in [11] is a framework that allows information related to the software to be developed gathered from various sources, including users' feedback on SNS, previous project documentation, and experts. AI techniques are applied to process the collected information, and finally, the requirements descriptions are produced. Later, N. M. Rizk et al. in [24] adapted the methodology introduced by [11] to proposed CREeLS. CREeLS is offered specifically for the eLearning System. A classification model was proposed in another research conducted by S. Taj et. al. in [28]. This model enables the crowd to actively participate in providing feedback which later, the feedbacks will be classified into functional and non-functional requirements.

In 2019, a study [32] proposed applying one RE artifact, User Story (US), in a crowd-based RE platform. It is reported that USs are estimated to be used by over half of the practitioners in the software industry to capture requirements. In this research, the participants are from the crowd. The participants have to provide information through four self-explanatory steps: Role, Goal, Benefit and Verification, and Category Selection. Finally, USs are formulated by the data the participants provided in these four steps. The requirement engineering will extract the potential requirements from the formulated USs.

V. DISCUSSION

In a study, Altug [35] stated that requirement engineering is a crucial stage in the software development life cycle process. During this stage, the requirement engineer must determine the optimum depth and breadth to obtain quality requirements. Crowd-based RE is an emerging approach that can help in securing quality software requirements. As depicted in Table I, we reviewed 20 primary researches that relate to crowd-based RE efforts.

In RQ1, we found that the researchers used no dominant techniques. However, we found that all 20 research efforts applied crowd-based RE in analysis, elicitation, or both. Analysis and elicitation are the early activities in the RE phase, in which the involvement of stakeholders to provide information is crucial. This is where crowd-based RE is adapted to improve and simplify the activities. We believe more research should be conducted to explore ways to utilize crowdsourcing techniques and improve the RE process.

As for RQ2, we discover two main challenges in the existing research: stakeholders' coverage and information reliability.

Through the review, we discover that stakeholders' input is crucial to ensure that the information obtained is from genuine sources and reliable. Since it is challenging due to numerous stakeholders, many researchers are exploring ways to involve the various stakeholders and ease the process of getting the information. One of the evolving initiatives is the attempt to provide the stakeholders' analysis tool. Other than that, there are also issues of malicious stakeholders that may respond for their benefit. Therefore, even if we already filter the source of information, which is the stakeholders, at the beginning of the project, it does not guarantee that we can obtain quality requirements. On top of that, by having a diverse set of stakeholders, we may get more relevant and meaningful requirements.

All of the crowd-based RE efforts require the users' involvement to obtain information. iRequire [17], AppEcho [19], FAME [21], SUPERSEDE [25], Classification Model [28], Continuous Requirement Elicitation Methodology [29] and Automated Feature Identification [31] fully rely on the users to provide data to developers. Other than that, Crowdsourced RE Platform for User Story (US) Authoring [32] also fully depends on the users' participation. Besides relying on the crowd input, iRequire, AppEcho, FAME, and Crowdsourced RE Platform for User Story (US) Authoring requires substantial effort to manually perform requirements extraction and refinement. Thus, there are research opportunities to improve the requirements extraction and refinement process after vast information is obtained from the crowd. Furthermore, iRequire [17], AppEcho [19], FAME [21], and Automated Feature Identification [31] can be enhanced in the future as their current capabilities only focus on mobile software.

Besides, there exist efforts that incorporating Artificial Intelligence techniques to process information obtained from the crowd to produce requirements descriptions such as SUPERSEDE [25], Effective Classification Methodology [11], CREeLS [24], Classification Model [28], Continuous Requirement Elicitation Methodology [29] and Automated Feature Identification [31] to handle reliability issues. This is important because the information gathered could be from anybody who gave their responses and feedback. We are well aware that anyone can hook on a web-based software system and mobile apps to give their responses and feedback in this Internet technology era. In their research, M. Bano and D. Zowghi [27] stated that user involvement in software development requires resources and careful management. If the user involvement is not carefully handled, it can cause issues and problems rather than benefits.

There is no single approach that could solve the problem in the traditional RE process. Current crowd-based RE researches span through methods, techniques, tools, and web-based platforms to assist requirements engineering process in many ways while utilizing crowdsourcing benefits. Each of the efforts is unique to solve a specific problem or address the explicit concern in any of the requirements engineering areas. The efforts are made to take advantage of the crowdsourcing technique and the assistance of the AI technique to obtain quality requirements.

VI. Conclusion

In summary, the researchers on crowd-based RE continuously explore the research area and propose new approaches, techniques, and tools to improve the crowd-based RE further. The progressing trend of the research in crowd-based RE proves that this field has more room to improve and to explore. With the rapid growth of technology, particularly the social web and mobile technology, crowd-based RE is becoming more relevant and important to elicit requirements because more people are using the technology to communicate and the software developed to be used by these people must be reliable meeting their needs.

A more comprehensive range of stakeholders can be accessed through crowdsourcing techniques to obtain valuable and meaningful information. Having access to a broader range of stakeholders can provide the breadth of knowledge that leads to quality requirements.

This paper presents the literature on crowd-based RE to complement conventional requirement elicitation techniques to obtain quality requirements. The benefits and advantages of crowd-based RE are worth exploring to strengthen requirement engineering in the future. As for the researchers exploring crowd-based RE, this paper also summarizes the challenges and limitations of crowd-based RE efforts to date.

For future works, it is beneficial to explore the utilization of crowd-based RE to obtain quality software requirements by optimizing the depth and breadth of information at a reduced cost of time and money. We believe crowd-based RE can simplify and improve the RE process to obtain quality software requirements that are later able to produce quality software systems.

Acknowledgment

We are thankful to Universiti Teknikal Malaysia Melaka for funding the publication of this paper through a research grant numbered PJP/2020/FTMK/PP/S01774.

References

[1] J. Howe, "Crowdsourcing: A definition," Retrieved 4 December 2019, from http://www.crowdsourcing.com/cs/2006/06/crowdsourcing_a.html, 2006.

[2] S.S. Bhatti, X. Gao, and G. Chen, "General framework, opportunities and challenges for crowdsourcing techniques: A Comprehensive survey," Journal of Systems and Software, 167, p.110611, 2020.

[3] U.S. Ghanyani, M. Murad and W. Mahmood, "Crowd-based Requirement Engineering," International Journal of Education and Management Engineering, 3, pp.43-53, 2018.

[4] A. Adepetu, A. A. Khaja, Y. Al Abd, A. Al Zaabi and D. Svetinovic, "Crowdrequire: A requirements engineering crowdsourcing platform," in 2012 AAAI Spring Symposium Series, March 2012.

[5] J.A. Khan, L. Liu, L. Wen, and R. Ali, "Crowd intelligence in requirements engineering: Current status and future directions," in International working conference on requirements engineering: Foundation for software quality (pp. 245-261). Springer, Cham, March 2019.

[6] A. Ahmad, K. Li, C. Feng, S. M. Asim, A. Yousif and S. Ge, "An empirical study of investigating mobile applications development challenges," IEEE Access, 6, pp.17711-17728, 2018.

[7] A. Inam, A Study of Requirements Engineering Practices Among Software Developers at UUM Information Technology, Ph.D. thesis, Universiti Utara Malaysia. 2015.

[8] M. Hargrave, "What Is Crowdsourcing," Retrieved 28 January 2020, from https://www. Investopedia. com/terms/c/crowdsourcing. Asp, 2019.

[9] E.C. Groen, N. Seyff, R. Ali, F. Dalpiaz, J. Doerr, E. Guzman, M. Hosseini, J. Marco, M. Oriol, A. Perini, and M. Stade, "The crowd in requirements engineering: The landscape and challenges," IEEE Software, 34(2), 2017, pp.44-52.

[10] E.C. Groen, J. Doerr, and S. Adam, "Towards crowd-based requirements engineering a research preview," in International Working Conference on Requirements Engineering: Foundation for Software Quality, Springer, Cham, March 2015, (pp. 247-253).

[11] C. Li, L. Huang, J. Ge, B. Luo, and V. Ng, "Automatically classifying user requests in crowdsourcing requirements engineering," Journal of Systems and Software, 138, 2018, pp.108-123.

[12] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," 2007.

[13] D. Yang, D. Wu, S. Koolmanojwong, A.W. Brown and B.W. Boehm, "Wikiwinwin: A wiki-based system for collaborative requirements negotiation," in Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), IEEE, January 2008, (pp. 24-24).

[14] S.L. Lim, D. Quercia, and A. Finkelstein, "StakeNet: using social networks to analyze the stakeholders of large-scale software projects," in 2010 ACM/IEEE 32nd International Conference on Software Engineering, Vol. 1, IEEE, May 2010, pp. 295-304.

[15] S.L. Lim, D. Quercia, and A. Finkelstein, "StakeSource: harnessing the power of crowdsourcing and social networks in stakeholder analysis," in 2010 ACM/IEEE 32nd International Conference on Software Engineering, Vol. 2, IEEE, May 2010, pp. 239-242.

[16] S.L. Lim, D. Damian, and A. Finkelstein, "StakeSource2. 0: using social networks of stakeholders to identify and prioritize requirements", in Proceedings of the 33rd international conference on Software engineering, May 2011, pp. 1022-1024.

[17] N. Seyff, F. Graf and N. Maiden, "End-user requirements blogging with iRequire," in 2010 ACM/IEEE 32nd International Conference on Software Engineering, Vol. 2, IEEE, May 2010, pp. 285-288.

[18] S.L. Lim and A. Finkelstein, "StakeRare: using social networks and collaborative filtering for large-scale requirements elicitation," IEEE transactions on software engineering, 38(3), 2011, pp.707-735.

[19] N. Seyff, G. Ollmann and M. Bortenschlager, "AppEcho: a user-driven, in situ feedback approach for mobile platforms and applications," in Proceedings of the 1st International Conference on Mobile Software Engineering and Systems, June 2014, pp. 99-108.

[20] R. Snijders, F. Dalpiaz, S. Brinkkemper, M. Hosseini, R. Ali and A. Ozum, "REfine: A gamified platform for participatory requirements engineering," in 2015 IEEE 1st International Workshop on Crowd-Based Requirements Engineering (CrowdRE), IEEE, August 2015, pp. 1-6.

[21] M. Oriol, M. Stade, F. Fotrousi, S. Nadal, J. Varga, N. Seyff, A. Abello, X. Franch, J. Marco, and O. Schmidt, "FAME: supporting continuous requirements elicitation by combining user feedback and monitoring," in 2018 IEEE 26th International Requirements Engineering Conference (RE), IEEE, August 2018, pp. 217-227.

[22] R. Sharma and A. Sureka, "CRUISE: A platform for crowdsourcing Requirements Elicitation and evolution," in 2017 Tenth International Conference on Contemporary Computing (IC3), IEEE, August 2017, pp. 1-7.

[23] M.Z. Kolpondinos and M. Glinz, "GARUSO: a gamification approach for involving stakeholders outside organizational reach in requirements engineering," Requirements Engineering, 2019, pp.1-28.

[24] N.M. Rizk, A.M. Zaki, E.S. Nasr, and M.H. Gheith, "CREeLS: Crowdsourcing based Requirements Elicitation for eLearning Systems," International Journal of Ad, 10(10), 2019.

[25] A. Perini, "Data-Driven Requirements Engineering. The SUPERSEDE Way", in Annual International Symposium on Information Management and Big Data, Springer, Cham, September 2018, pp. 13-18.

[26] D. Johnson, J. Tizard, D. Damian, K. Blincoe, and T. Clear, "Open CrowdRE Challenges in Software Ecosystems," in 2020 4th International Workshop on Crowd-Based Requirements Engineering (CrowdRE), IEEE, November 2020, pp. 1-4.

[27] M. Bano and D. Zowghi, "Users' involvement in requirements engineering and system success," in 2013 3rd International Workshop on Empirical Requirements Engineering (EmpiRE), IEEE, July 2013, pp. 24-31.

[28] S. Taj, Q. Arain, I. Memon and A. Zubedi, "To apply data mining for classification of crowdsourced software requirements," in Proceedings of the 2019 8th International Conference on Software and Information Engineering, April 2019, pp. 42-46.

[29] A. Alwadin and M. Asharagi, "Crowd, generated data mining for continuous requirement elicitation," Journal of Advanced Computer Science and Application, 10, 2019, pp. 45-50.

[30] J. Wouters, CrowdRE@ KMar-Case study of the implementation of CrowdRE at the Royal Dutch Marechaussee, Master thesis, Faculty of Science, Utrecht University, Netherlands. 2020.

[31] T. Iqbal, N. Seyff and D. Mendez, "Generating requirements out of thin air: Towards automated feature identification for new apps," in 2019 IEEE 27th International Requirements Engineering Conference Workshops (REW), IEEE, September 2019, pp. 193-199.

[32] A. Menkveld, S. Brinkkemper, and F. Dalpiaz, "User story writing in crowd requirements engineering: The case of a web application for sports tournament planning," in 2019 IEEE 27th International Requirements Engineering Conference Workshops (REW), IEEE, September 2019, pp. 174-179.

[33] M. Hosseini, A. Shahri, K. Phalp and R. Ali, "Recommendations on adapting crowdsourcing to problem types," in 2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS), IEEE, May 2015, pp. 423-433.

[34] M. Hosseini, K. Phalp, J. Taylor, and R. Ali, "The four pillars of crowdsourcing: A reference model," in 2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS), IEEE, May 2014, pp. 1-12.

[35] C. Altug, "Crowd-based Requirements Engineering," Retrieved 4 December, 2019, from https://medium.com/thinkerfox/crowd-based-requirements-engineering-e7e9d044bf71, 26 September 2019.

[36] A. Hussain, E.O. Mkpojiogu and F. Mohmad Kamal, "The role of requirements in the success or failure of software projects," International Review of Management and Marketing, 6(S7), 2016, pp.306-311.

Application of Convolutional Neural Networks for Binary Recognition Task of Two Similar Industrial Machining Parts

Hadyan Hafizh[1]*, Amir Hamzah Abdul Rasib[2]
Rohana Abdullah[3], Mohd Hadzley Abu Bakar[4], Anuar Mohamed Kassim[5]
Faculty of Mechanical and Manufacturing Engineering Technology[1, 2, 3, 4]
Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia[1, 2, 3, 4]
Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia[5]

***Abstract*—Misclassifying parts in the small-medium manufacturing enterprise can lead to serious consequences. Manual inspection, as currently practiced, allows for compromises in product traceability. Due to this condition, inspection of the part's number is not digitally visible. Due to a lack of modern traceability, customers receive incorrect parts, and the same incidents continue to occur. It is essential to transform manual inspections into digital and automated ones. AI-based technologies have recently been employed to enable a smart and intelligent recognition system for industrial machining parts. Convolutional Neural Networks (CNN) are widely used for image recognition tasks and are gaining popularity as deep learning algorithms. In this paper, a CNN model is used to perform binary recognition on two similar industrial machining parts. The model has been trained to recognise two classes of machining parts: Parts A and B. The dataset used to train the model includes both original and augmented images, with a total of 2447 images for both classes. The performance metrics have been measured during the training process, and 10 experiments have been conducted to evaluate the performance of the model. The test results reveal that the CNN model achieves 98% mean accuracy, 97.1% precision for Part A, 99% precision for part B and 0.982 AUC value. The results demonstrate the effectiveness of the CNN-based recognition of parts. It offers an effective alternative and is a compelling method for quality assurance in small-medium manufacturing enterprises.**

Keywords*—Convolutional neural networks; binary recognition; machining parts; deep learning

I. Introduction

Computer vision and automation are now being used by modern industrial firms to achieve higher quality and more accurate inspection of parts. Deep learning, for example, is an AI-based technology that assists the industry regarding automation with minimal human intervention. The convolutional neural network (CNN) is one of the most important deep-learning-based computer vision methods. CNN has a wide range of applications. One of the most popular uses of CNN is image classification and defect detection in industrial products. Zhang et al. [1] used CNN to study defects detection for aluminium alloy in robotic arc welding. In this study, data augmentation and noise addition have been applied to boost the CNN dataset. It was found that the CNN model was able to attain a 99.38% accuracy. Westphal et al. [2] employed CNN to detect irregularities in selective laser sintering (SLS). Two transfer learning CNN models were used with pre-trained weights to classify good and defective images during the manufacturing of parts. The VGG16 transfer learning CNN model achieved the best results with 95.8% accuracy. Furthermore, a study on weldment classification using a vision system by Bacioiu et al. [3] had reported that the CNN model can achieve the highest accuracies of 71%, 89% and 95% for 6-class, 4-class and 2-class, respectively. A similar study utilising the SS304 TIG welding process had summarised that CNN is capable of learning powerful representations of welding defects [4]. Approximately 89.5% of accuracy was reported in classifying good vs defective welds with the use of CNN.

The application of CNN for the inspection of metal additive manufacturing parts has been examined by Cui et al. [5]. In this study, the regularisation and dropout layers were added to the CNN architecture in order to avoid overfitting problems. With the help of data augmentation, about 92.1% of accuracy was obtained for the CNN model. A similar study was also conducted by Xiaoming et al. [6] where CNN was utilised to detect defects in metallic surfaces. Data augmentation was further applied to enrich the training data. The study also contributed to the new dataset, called GC10-DET, for metallic surface defect detection. By using this dataset and the CNN model, the proposed method successfully met the accuracy requirements for the detection of metallic defects. The application of CNN to metal manufacturing parts was also conducted by Ma et al. [7]. Four CNN models were utilised in this study to detect the weld defects of galvanised steel sheets. The study had found that the VGG16 transfer learning CNN model combined with the data augmentation method made the best model to achieve a state-of-the-art performance in detecting weld defects.

Due to high classification abilities, CNN is gaining attention in various industrial fields, particularly in metal and welding defect detection. CNN has proven to be effective in performing recognition and classification tasks [8-11]. CNN was also used in casting applications to detect and investigate the defects of casting products. A study conducted by Mery et al. [12] had used synthetic defects in order to improve the performance of the CNN model. This study proposed a CNN

*Corresponding Author

architecture called Xnet-II which has 30 layers and more than 1,350,000 parameters. Another study conducted by Jiang et al. [13] employed X-ray images of casting products as inputs to the CNN model. The test accuracy achieved was reported to reach up to 95.5%. Various defects in the casting product, such as blowholes, chipping, cracks and wash automatically, were investigated by Nguyen et al. [14]. The study used 6000 images with 768 × 768 px resolution as input to the CNN model. The training model was reported to attain an experimental accuracy of more than 98%.

Although previous researchers have made significant efforts in detecting defects in industrial products using the CNN model [15-20], little attention has been paid to recognise similar industrial machining parts. The issue that the human operator faces on the manufacturing floor is not only related to defects, but also to misclassification of machining components. This problem arises due to the similarity of two machining parts, and when handled by a human operator, it leaves room for human error. Misclassification of machining parts is a real issue that occurs on the manufacturing floor. Due to lack of modern traceability, incorrect parts have been delivered to customers and the same incident is repeatedly occurring. The company's image becomes tarnished, decreasing its reputation in the eyes of existing and potential customers and vendors. It is an urgent matter to transform manual inspections into digital and automated ones. Therefore, the current work proposes a CNN model to recognise and classify two similar machining parts. The proposed CNN model can be integrated into a machine-vision system and perform automatic recognition tasks.

II. Methodology

The dataset used for training the model and testing the results is described in this section. Subsequently, the process of data augmentation is discussed in order to enhance the performance of the CNN model. The CNN model used in this work is also presented and discussed.

A. Data Structure

The original images of the machining parts dataset were captured by using an android-based smartphone, with a resolution of 750 × 1000 px. A total of 160 images were taken for both Parts A and B, with each part consisting of 80 images. These images are then resized to 224 × 224 px before being fed into the CNN model as input. Fig. 1 presents a sample of the resized original images taken with an android-based smartphone for both Parts A and B. It can be seen that these two parts are similar. There is a high possibility that these two parts will be misidentified by a human operator.

The original images were then used to generate an augmented dataset in order to improve the CNN model's performance. By performing various augmentation processes such as rotation, translation, zoom and brightness adjustment, a total of 2317 augmented images were generated. The original and augmented images have been combined to produce a total of 2477 images, 1234 of which belongs to Part A and 1243 to Part B. Among 1234 images from Part A, 980 images were used to create the training dataset and the remaining 254 were used to create the test dataset. As for Part B, 987 images were used to create the training dataset, while 256 images were used to create the test dataset. A balanced dataset was used to train the CNN model, with nearly equal numbers of images for training and testing of both classes. The data structure applied in the current work is presented in Fig. 2.

B. Data Augmentation

Image data augmentation is an alternative method to expand a training dataset by creating new versions of images. It can improve the performance of deep learning models by creating variations of the images they learn. Data augmentation is a regulatory mechanism designed to prevent model overfitting. This procedure works by performing the following operations, as shown in Table I.

The augmented dataset is generated by randomly selecting images from the original dataset. The process shown in Table I is then applied to generate a total of 2317 augmented images. In order to create the augmented dataset, four processes were applied to the original images. These processes have been selected based on the common scenario encountered on the manufacturing floor when performing the recognition task. The images can be arbitrarily placed under the camera before performing the recognition task; therefore, the rotation and translation processes are applied to generate a series of augmented images with random placement. Furthermore, the camera's zoom and brightness can be adjusted. As a result, the augmented dataset with different zoom and brightness settings is essential for training the model.

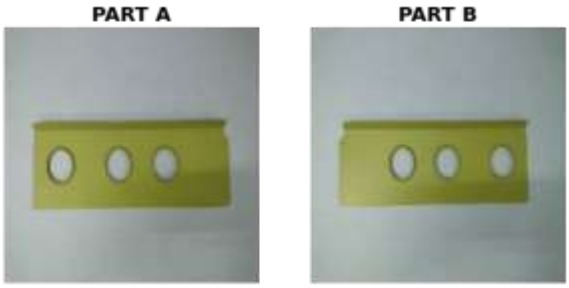


Fig. 1. Sample of Original Images.

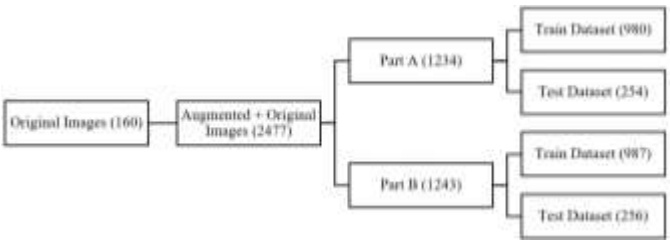


Fig. 2. Data Structure Employed in the Current Work.

TABLE I. Process Applied in Data Augmentation

Process	Description
Rotation	Randomly rotate image in the range of 00 to 450
Translation	Randomly shift image horizontally and vertically in the range of 0 to 0.1 (as a fraction of total width and height)
Zoom	Randomly zoom the image in the range of 0 to 0.2
Brightness	Randomly adjust image brightness in the range of 0.1 to 1.0

C. Convolutional Neural Network

A convolutional neural network (CNN) is a deep learning algorithm that is widely utilised in the area of image recognition. CNN can be regarded as a special type of feed-forward neural network in AI technology. CNN's main advantage over its predecessors is that it automatically detects significant features without the need for human intervention, making it the most widely used [21, 22]. As in a standard multi-layer neural network, a CNN has at least one convolutional layer followed by at least one fully connected layer. The CNN architecture applied in this paper consists of three convolutional layers and two fully connected layers.

The proposed CNN architecture is based on the LeCun model [23]. The model consists of three convolutional layers followed by fully connected layers, as illustrated in Fig. 3. The machining parts image was captured with a resolution of 750 × 1000 px. Before feeding the original and augmented images into the CNN model, they were resized to 224 × 224 px. These images were then transformed into grey-scale and with the dimension of 224 × 224 × 1. The grey-scale images were then passed through a block of convolution layers with a kernel size of 3 × 3 and a stride of 1 px.

In the convolutional layers, the number of output filters was set to 8, 16 and 32, respectively. Following the convolutional layers, three max-pooling layers with window sizes of 2 × 2 and strides of 2 px were added to compress the spatial representation of the input data [24]. Furthermore, the Rectified Linear Unit (ReLU) function was used as the activation function in the convolutional layer.

The fully connected layer is the primary building block of traditional artificial neural networks. It converts the high-level filtered machining parts image into votes. This layer's primary goal is to perform classification using the features extracted by the convolutional layers. Because the current work's class is binary, the model must only choose between two classes, Parts A and B. Due to the flattening process, the input is treated as a single list. The flattened layer is 1 × 25088 in size. The flattened output is then fed to a feed-forward neural network. The backpropagation algorithm is applied to every iteration of training in the dense layer. In the last layer of the CNN model, the sigmoid activation function was used to estimate the probability of the sample belonging to each class.

D. Experimental Procedure

A series of numerical experiments were conducted following the procedure depicted in Fig. 4. The first step in the process is to collect true label data, which was accomplished by taking 160 images of Parts A and B with an android-based smartphone. The original images were then randomly selected to generate an augmented dataset. This process yields a database of machining part images, which are saved for later use to train the CNN model. Having a sufficient dataset, a CNN model can then be developed. The architecture of the CNN model is shown in Fig. 3.

The model is trained by using the augmented and original dataset until the accuracy achieves a value of more than 95%. Subsequently, a series of numerical experiments are performed. The model was run 10 times and its performance was measured. The loss and accuracy values per epoch during the training and testing were also measured. A confusion matrix was further computed for each numerical experiment in order to measure the performance of the CNN model. The model was then applied to perform recognition and prediction tasks using a random image from the test dataset. Finally, the model was saved, and the experiment was successfully completed.

The training and recognition tasks were repeated ten times. The performance of the CNN model was measured and visualised in the form of a confusion matrix for each training process. The accuracy, precision, sensitivity and specificity of the model can be calculated from this matrix using the following equations [25]:

$$\text{Accuracy}=(TN+TP)/(TN+FP+TP+FN) \quad (1)$$

$$\text{PPV}=TP/(TP+FP) \quad (2)$$

$$\text{NPV}= TN/(TN+FN) \quad (3)$$

$$\text{Sensitivity}=TP/(TP+FN) \quad (4)$$

$$\text{Specificity}=TN/(TN+FP) \quad (5)$$

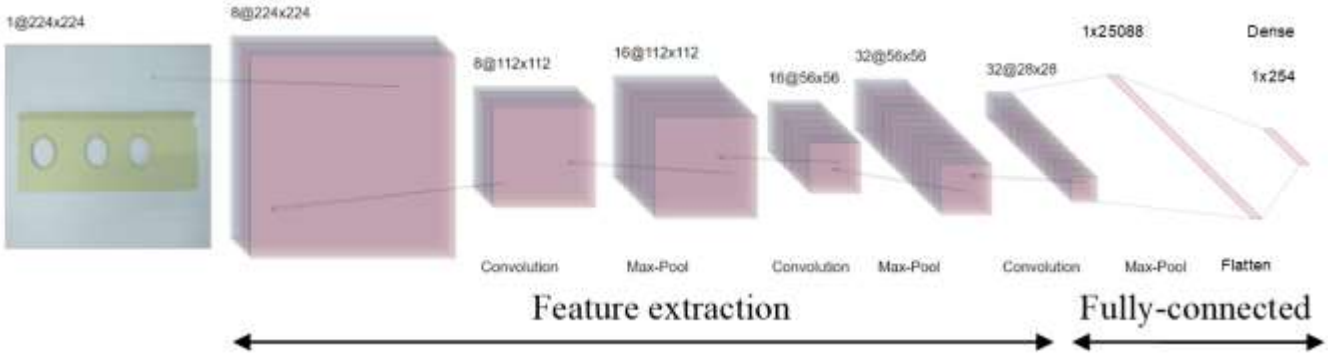


Fig. 3. Architecture of the CNN Model.

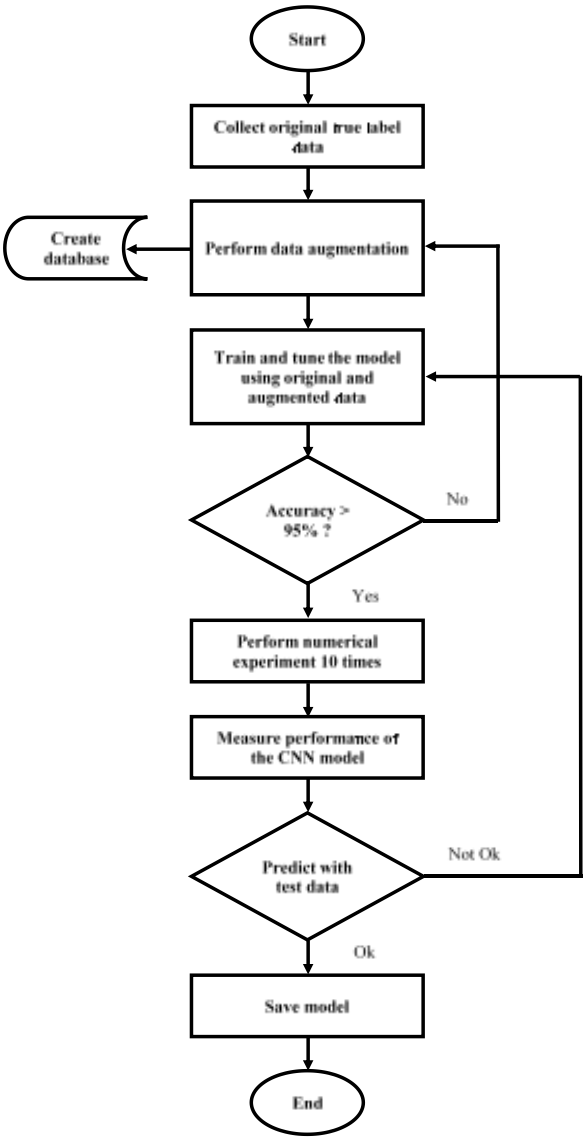


Fig. 4. Flow Chart of the Experimental Procedure.

where TP, TN, FP and FN are true positive, true negative, false positive and false negative, respectively. The TP value in the confusion matrix represents the number in which the CNN model has predicted as Part A and the true label is actually Part A. The TN value represents the number in which the CNN model has predicted as Part B and the true label is actually Part B. The FP value means the number in which the CNN model has predicted as Part B, but the true label is actually Part A. Lastly, the FN value means the number in which the CNN model has predicted as Part A, but the true label is actually Part B.

Precision values for Parts A and B are also referred to as Positive Predictive Value (PPV) and Negative Predictive Value (NPV), respectively. The PPV value counts the number of observations that are predicted to be positive (Part A) and are, in fact, positive. Similarly, the NPV value indicates how many predictions are correct out of all negative predictions (Part B). Furthermore, the Receiver Operating Characteristics (ROC) curve and the Area Under the Curve (AUC) value are also measured during the training and testing processes.

III. Results and Discussion

A. Binary Class Test

The machining parts dataset contains original and augmented images. By using this combined dataset, the experiment was conducted, and the recognition task was performed for two classes of machining parts, i.e., Parts A and B. For each experiment, the training dataset was shuffled but the random seed parameter was kept constant to ensure all algorithms used the same samples as the testing and training data. The test dataset utilised in the current work was never used to train the model, therefore it represents new data for the trained model. After the training process, a recognition task was performed by randomly selecting the image from the test dataset for both classes. The recognition task was conducted 10 times and the results are presented in Fig. 5. The true label and the prediction results are visualised on the images. From this figure, the CNN model correctly recognised all of the test images. These images are randomly rotated and translated to simulate the real-world situation in which a human operator attempts to perform a recognition task by placing machining parts under the sensor. As discussed in the data augmentation section, the brightness and zoom level of the test images were also randomly assigned within the prescribed range.

B. Performance Measures

The CNN model's performance metrics were measured using the Confusion Matrix. Equations (1–5) were used to calculate the Accuracy, Precision, Sensitivity and Specificity values from the Confusion Matrix. The results are displayed in Table I. The individual experiment has been considered. The first and second experiments achieved accuracy values of 0.992 and 0.986, respectively. When compared to the first experiment, the precision values (PPV and NPV) in the second experiment are lower. Despite the fact that the accuracy values for both experiments are similar, the second experiment has a higher false negative value (4 Parts A are wrongly recognised as Part B). In consequence, the precision value of Part B (0.984) in the second experiment is lower than in the first.

The third and fourth experiments have the same accuracy (0.984), but they are smaller in magnitude than the former experiments. The CNN model, in particular, was able to correctly recognise all the images in Part A in the fourth experiment. As a result, the precision value of Part B is 1. Although the CNN model performed well with the image of Part A, there are 8 images of Part B that were incorrectly identified as Part A. When compared to the third experiment, this condition leads to a lower precision value for Part A. The fifth through tenth experiments showed a fluctuation in the accuracy value.

The CNN model achieved the lowest accuracy in the tenth experiment. In this experiment, 14 Parts B were incorrectly identified as Part A, while 6 Parts A were incorrectly identified as Part B. Although the instantaneous accuracy values fluctuated, the CNN model was able to achieve 0.980 mean accuracy across 10 experiments.

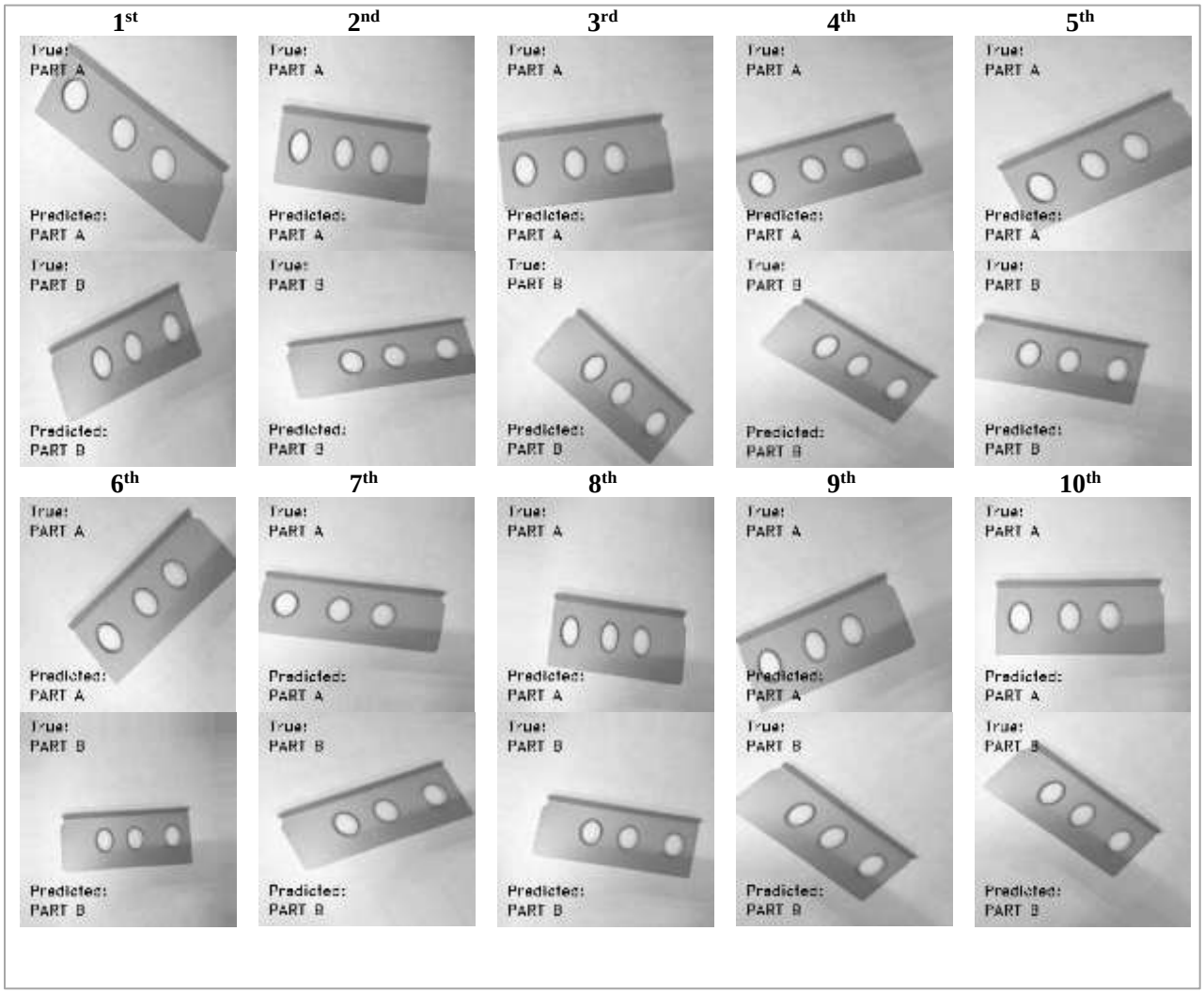


Fig. 5. Recognition Results of the CNN Model for 10 Experiments.

Table II reveals an intriguing pattern: the number of images of Part B that were incorrectly classified as Part A is significantly higher than the other way around. In other words, the precision value for Part A is lower than the precision value for Part B. This condition indicates that the Part A training dataset should be improved.

Enhancing the augmentation process, such as increasing the variation of image rotation and translation, can result to an improvement. Furthermore, the ROC curves and AUC values can be calculated using the sensitivity and specificity values shown in Table II. The ROC curve is a plot of the True Positive Rate (TPR) versus the False Positive Rate (FPR) for each possible prediction threshold. TPR is another name for sensitivity, and FPR can be calculated from 1 – specificity. The ROC curve illustrates the trade-off between correctly recognised positive samples (Part A) and misclassified negative samples (Part B).

The instantaneous AUC value obtained was 0.982. A perfect CNN model has an AUC value equal to one, indicating that it has a high level of separability. This is illustrated by a solid line (perfect classifier) in Fig. 6. A poor CNN model has an AUC value close to zero, indicating that it has the worst measure of separability. When the AUC value is 0.5, the model has no class separation capacity at all (random classifier), as indicated by the linear dashed line in Fig. 6. Based on the computed AUC value, the current model is considered to perform well in recognising and classifying two similar machining parts.

Fig. 7 shows the evolution of loss and accuracy values during the training process. The instantaneous value of loss and accuracy were recorded for 30 epochs. Both loss and accuracy diagrams characterise the training process. They provide initial information regarding the effectiveness of the selected hyperparameters. The current work uses binary cross-entropy as a loss function since it is widely employed for binary classification tasks [26]. Furthermore, the Adaptive Moment Estimation (Adam) algorithm was applied for the optimisation process and the learning rate was set to 0.001. The number of epochs and the batch size of the CNN model were set to 30 and 32 respectively.

TABLE II. PERFORMANCE METRICS OF CNN MODEL

Experiment	Confusion Matrix		Accuracy	Precision		Sensitivity	Specificity
				PPV	NPV		
1st	253	1	0.992	0.988		0.996	0.988
	3	253			0.996		
2nd	250	4	0.986	0.988		0.984	0.988
	3	253			0.984		
3rd	252	2	0.984	0.977		0.992	0.977
	6	250			0.992		
4th	254	0	0.984	0.969		1.000	0.969
	8	248			1.000		
5th	253	1	0.990	0.984		0.996	0.984
	4	252			0.996		
6th	251	3	0.978	0.969		0.988	0.969
	8	248			0.988		
7th	250	4	0.973	0.962		0.984	0.961
	10	246			0.984		
8th	253	1	0.980	0.966		0.996	0.965
	9	247			0.996		
9th	251	3	0.973	0.958		0.988	0.957
	11	245			0.988		
10th	248	6	0.961	0.947		0.976	0.945
	14	242			0.976		
Mean			**0.980**	**0.971**	**0.990**	**0.990**	**0.970**

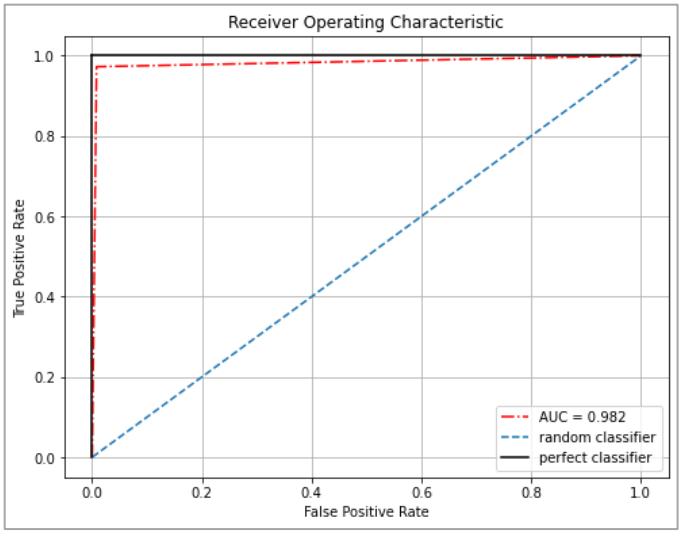


Fig. 6. ROC Curve and AUC Values.

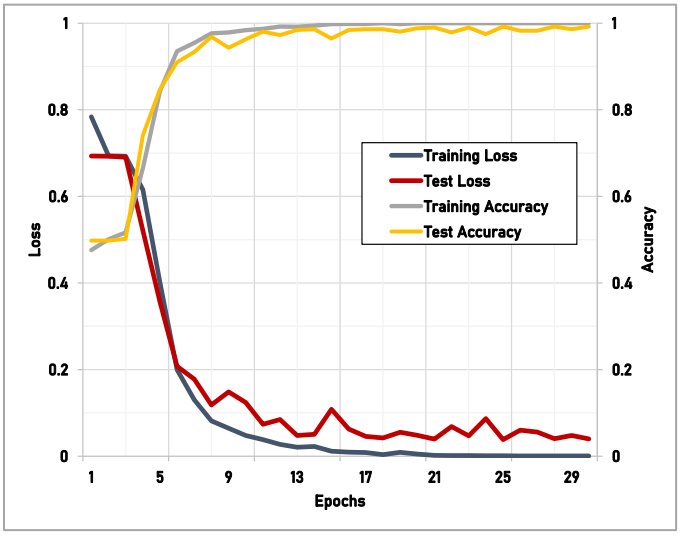


Fig. 7. Loss and Accuracy Evolution during Training.

From Fig. 7, it can be seen that there is a slight gap between training and test loss. This indicates an unrepresentative dataset, which means that the training dataset used to train the CNN model does not provide sufficient information to learn the recognition problem [27]. This situation is consistent with the condition of the precision value obtained in Table II. It can be observed that the precision of Part A is lower compared to Part B. This suggests that the training dataset of Part A is moderately unrepresentative.

In Fig. 7, the accuracy has been recorded for one experiment and plotted over 30 epochs. The accuracy values for ten experiments are shown in Fig. 8, and the mean value has been calculated accordingly. The accuracy values ranged between 96% to 99%. The mean accuracy calculated from the instantaneous value was found to be 98%. This indicates that the CNN model has good recognition performance and can provide an alternative method for manual inspection of industrial machining parts.

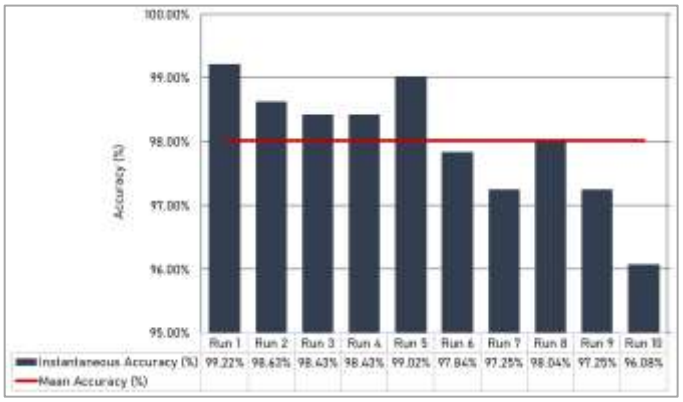


Fig. 8. Instantaneous and mean Accuracy for 10 Experiments.

IV. Conclusion

In this paper, a CNN model was employed to perform binary classification tasks of two similar machining parts. The model consists of three convolutional layers and three max-pooling layers for feature extraction, followed by two fully connected layers for recognition and classification. Two classes, Parts A and B, have been assigned for the recognition task. The dataset used to train the model consists of 160 original images and 2317 augmented images. Four types of data augmentation processes were applied in order to improve the performance of the model. Rotation, translation, zooming and brightness adjustments were all part of the augmentation process. These images were then assigned to one of two classes: Part A (1234 images for training and 254 for testing) and Part B (1243 images for training and 256 for testing).

The model was run 10 times, and the performance metrics in the form of loss and accuracy values were measured for each experiment. The confusion matrix was also recorded, as well as the model's accuracy, precision, sensitivity, specificity and AUC value. Experiment results show that the CNN model achieved a mean accuracy of 98%. The test outcomes also show that the mean precision values for Parts A and B are 0.971 and 0.99, respectively. The instantaneous ROC curve and AUC value (0.982) indicate that the current CNN model performs well in recognising and classifying two similar machining parts. The results further demonstrate the effectiveness of part recognition based on CNN. It offers a compelling alternative to replace manual inspections currently practiced in small-medium manufacturing enterprises. In the future, the CNN model's results should be compared to those of other well-known CNN architectures, such as MobileNet, ResNet50, and VGG16, to investigate their performance in recognising two similar machining parts.

References

[1] Z. Zhang, G. Wen, and S. Chen, “Weld image deep learning-based on-line defects detection using convolutional neural networks for Al alloy in robotic arc welding,” J. Manuf. Process., vol. 45, no. March, pp. 208–216, 2019.

[2] E. Westphal and H. Seitz, “A machine learning method for defect detection and visualization in selective laser sintering based on convolutional neural networks,” Addit. Manuf., vol. 41, no. November 2020, p. 101965, 2021.

[3] D. Bacioiu, G. Melton, M. Papaelias, and R. Shaw, “Automated defect classification of Aluminium 5083 TIG welding using HDR camera and neural networks,” J. Manuf. Process., vol. 45, no. August, pp. 603–613, 2019.

[4] D. Bacioiu, G. Melton, M. Papaelias, and R. Shaw, “Automated defect classification of SS304 TIG welding process using visible spectrum camera and machine learning,” NDT E Int., vol. 107, no. November 2018, 2019.

[5] W. Cui, Y. Zhang, X. Zhang, L. Li, and F. Liou, “Metal additive manufacturing parts inspection using convolutional neural network,” Appl. Sci., vol. 10, no. 2, 2020.

[6] X. Lv, F. Duan, J. J. Jiang, X. Fu, and L. Gan, “Deep metallic surface defect detection: The new benchmark and detection network,” Sensors (Switzerland), vol. 20, no. 6, 2020.

[7] G. Ma, L. Yu, H. Yuan, W. Xiao, and Y. He, “A vision-based method for lap weld defects monitoring of galvanized steel sheets using convolutional neural network,” J. Manuf. Process., vol. 64, no. December 2020, pp. 130–139, 2021.

[8] Y. Yang et al., “A lightweight deep learning algorithm for inspection of laser welding defects on safety vent of power battery,” Comput. Ind., vol. 123, p. 103306, 2020.

[9] R. Miao et al., “Real-time defect identification of narrow overlap welds and application based on convolutional neural networks,” J. Manuf. Syst., no. June 2020, pp. 1–11, 2021.

[10] [A. M. Deshpande, A. A. Minai, and M. Kumar, “One-shot recognition of manufacturing defects in steel surfaces,” Procedia Manuf., vol. 48, pp. 1064–1071, 2020.

[11] X. He, T. Wang, K. Wu, and H. Liu, “Automatic defects detection and classification of low carbon steel WAAM products using improved remanence/magneto-optical imaging and cost-sensitive convolutional neural network,” Meas. J. Int. Meas. Confed., vol. 173, no. August 2020, p. 108633, 2021.

[12] D. Mery, “Aluminum Casting Inspection Using Deep Learning: A Method Based on Convolutional Neural Networks,” J. Nondestruct. Eval., vol. 39, no. 1, 2020.

[13] L. Jiang, Y. Wang, Z. Tang, Y. Miao, and S. Chen, “Casting defect detection in X-ray images using convolutional neural networks and attention-guided data augmentation,” Meas. J. Int. Meas. Confed., vol. 170, no. November 2020, p. 108736, 2021.

[14] T. P. Nguyen, S. Choi, S. J. Park, S. H. Park, and J. Yoon, “Inspecting Method for Defective Casting Products with Convolutional Neural Network (CNN),” Int. J. Precis. Eng. Manuf. - Green Technol., vol. 8, no. 2, pp. 583–594, 2021.

[15] B. Staar, M. Lütjen, and M. Freitag, “Anomaly detection with convolutional neural networks for industrial surface inspection,” Procedia CIRP, vol. 79, pp. 484–489, 2019.

[16] Z. Snow, B. Diehl, E. W. Reutzel, and A. Nassar, “Toward in-situ flaw detection in laser powder bed fusion additive manufacturing through layerwise imagery and machine learning,” J. Manuf. Syst., vol. 59, no. February, pp. 12–26, 2021.

[17] H. Pan, Z. Pang, Y. Wang, Y. Wang, and L. Chen, “A New Image Recognition and Classification Method Combining Transfer Learning Algorithm and MobileNet Model for Welding Defects,” IEEE Access, vol. 8, pp. 119951–119960, 2020.

[18] F. Nagata et al., “Defect detection method using deep convolutional neural network, support vector machine and template matching techniques,” Artif. Life Robot., vol. 24, no. 4, pp. 512–519, 2019.

[19] X. Ji et al., “Filtered selective search and evenly distributed convolutional neural networks for casting defects recognition,” J. Mater. Process. Technol., vol. 292, no. June 2020, 2021.

[20] R. Perera, D. Guzzetti, and V. Agrawal, “Optimized and autonomous machine learning framework for characterizing pores, particles, grains and grain boundaries in microstructural images,” Comput. Mater. Sci., vol. 196, no. April, p. 110524, 2021.

[21] G. Yao, T. Lei, and J. Zhong, “A review of Convolutional-Neural-Network-based action recognition,” Pattern Recognit. Lett., vol. 118, pp. 14–22, 2019.

[22] A. Dhillon and G. K. Verma, “Convolutional neural network: a review of models, methodologies and applications to object detection,” Prog. Artif. Intell., vol. 9, no. 2, pp. 85–112, 2020.

[23] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[24] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," Proc. - 3rd IAPR Asian Conf. Pattern Recognition, ACPR 2015, pp. 730–734, 2016.

[25] L. Alzubaidi et al., Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, vol. 8, no. 1. Springer International Publishing, 2021.

[26] Y. Ho and S. Wookey, "The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling," IEEE Access, vol. 8, pp. 4806–4813, 2020.

[27] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7700 LECTU, pp. 437–478, 2012.

Design and Evaluation of an Engagement Framework for e-Learning Gamification

Mohammed Abdulaziz Alsubhi, Noraidah Sahari Ashaari, Tengku Siti Meriam Tengku Wook
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Malaysia

***Abstract*—Recently, gamification in education software development to improve student engagement and performance has become prevalent. Gamification is used to counter attrition and dropout issues in e-learning. A handful of methods are presented for the gamification of e-learning systems in the literature. However, the e-learning gamification methods proposed in the literature lacked consistency. The number and types of game elements used in the methods are varied. In addition, there is a lack of an engagement framework that can be used in applying game elements to e-learning systems. Therefore, this paper provides insights into gamification and how it is used in e-learning systems. Then, the study proposes and evaluates an engagement framework that can be used to guide developers on how to add game elements to e-learning to improve student engagement and performance. The framework consists of three components: game elements, learning activities, and engagement factors components. Two experts evaluate the engagement framework via a semi-structured interview. The evaluation results indicate that developers can efficiently and effectively use the framework to gamify e-learning systems for improved student engagement and performance.**

Keywords*—e-learning; learning activities; gamification; game elements; engagement framework*

I. Introduction

Student engagement plays an essential role in minimising student attrition and dropout rates. However, it also represents one of the main challenges in teaching and learning, especially in online education or e-learning. The level of engagement is influenced not only by how lessons are delivered but by the students' readiness and enthusiasm to engage with their studies. Hence sustaining student engagement is a daily struggle that leaves higher education lecturers in a quandary because the students' energy into immersing themselves in their courses plays a key role in their engagement [1].

Several studies have proposed the use of gamification to improve student engagement [2, 3]. Gamification adds game elements to e-learning systems to engage students and achieve the desired learning behaviours and outcomes [4, 5]. The researchers in the domain have proposed a wide array of gamification guidelines in different research areas. Most of the guidelines provide strategies for gamifying other systems, including business systems like Facebook, Twitter, etc. [6, 7]. Likewise, gamification strategies proposed in education intends to improve student performance, student outcome, or student engagement. Yet, the strategy adds game elements to educational systems in an ad-hoc manner. The wide variety of e-learning tools utilised in education, such as Moodle, Adaptweb, etc., makes the gamification process vague and unclear to developers [7, 8].

However, the literature has not adequately investigated an engagement framework that can help educational software developers gamify e-learning systems. Therefore, this paper aims to propose an engagement framework that can address the lack of methods or standards for the gamification of e-learning systems. The gamified e-learning system ensures students become engaged in their learning, which will indirectly lead to minimising the number of dropouts and attrition, which is a suffered by most higher education institutions (HEIs).

Subsequently, the framework is evaluated by submitting the framework and a document that explains the usage to experts. The experts assessed the selected game elements and their completeness. The experts were asked to suggest any additional game elements apart from those commonly used elements selected by the authors. Furthermore, the experts were asked to comment on the learning activities included in the framework for the gamification of e-learning systems. In this light, the experts were requested to evaluate the framework for its usability and whether the mapping between the game elements, learning activities, and engagement factors are easy to understand and can improve student engagement in e-learning systems. The overall result of the evaluation was positive. They agreed with the authors that the selected game elements were enough and needed no additional elements. Similarly, the experts approved that the learning activities were enough and could represent learning activities required for e-learning systems. The experts, as a result, agreed with the usability of the proposed engagement framework of the gamification in e-learning systems to improve engagement of higher education students.

The rest of the paper is organised as follows: Section II provides an overview of related work, Section III briefly describes learning management systems (LMSs), Section IV explains the reasons for attrition and dropouts in learning, Section V introduces the proposed engagement framework, Sections VI evaluates the framework, and VII concludes the paper.

II. Related Work

Based on the literature on gamification in education, several studies have employed game elements in e-learning systems for various purposes. Some have used gamification to improve student outcomes (performance), while others have utilised gamification to improve student engagement [1, 9, 10]. Given the scope of the current research, only those works that focus on gamification in e-learning systems for student engagement are reviewed here.

In [11], the researchers proposed a model for use in the gamification of e-learning systems for better user engagement. Several commonly used game elements were selected in gamifying e-learning systems. However, the researchers did not implement the model to investigate whether it improves user engagement.

The researchers in [12] also proposed an e-learning gamification model, but they further used it in the gamification of a newly developed e-learning platform. The model was validated via experimentation, where a group of students used the platform. Based on the data collected on the students' experiences with the gamified platform, the researchers were able to realise the extent to which the model was able to improve student engagement.

In another study [13], the researchers investigated the influence that a gamified learning environment may have on student learning achievements when used in conjunction with social media. The experiment showed that there was improved user engagement. However, in this study, the employment of gamification did not involve the use of an e-learning platform.

In [14], the researchers investigated which game elements could improve user engagement when added to LMSs. In addition, they explored the effects that the deployment of gamified e-learning platforms may have on users. In [15], the researchers assessed the level of user engagement when users in higher education institutions used a gamified environment. However, the researchers failed to identify which game elements improved which engagement factors. Moreover, they could not determine which game element needed to be used in which e-learning activity for improved user engagement.

In [16], the researchers explained the influence of gamification in online discussion systems on user engagement. The results showed that gamified online discussion platforms have a positive impact on user engagement. Furthermore, they also revealed that some technical issues and classmate behaviours were impediments to user engagement.

In another study [17], the researchers examined the impact of gamified e-assessments on user engagement. The researchers studied the effect of gamification on one learning activity, such as discussion and assessment, respectively. In other words, learning activities such as assignments and learning material were not studied in the research.

For ease of reference, the game elements that were used in e-learning systems to improve student engagement in the studies reviewed in this section are summarised in Table I.

TABLE I. Game Elements used in Studies on Gamification for Student Engagement

Source	Game elements
[18]	Badges
[19]	Points, badges, customisation, leaderboards, levels, challenges, quests, feedback, freedom to fail
[11]	Points, badges, levels, rankings, , challenges, rules, message board
[12]	Badges, leaderboard, storyline, backslash, challenges, sharing via Facebook, points, levels
[20]	Badges
[16]	Badges, experience points, leaderboard, progress bar, rewards
[21]	Points, leaderboard, levels, module division
[17]	Points, badges, avatar, Themes, music, and leaderboards

III. Learning Management System

In the context of HEIs, an LMS can be defined as "a server-based or cloud-based software program containing information about users, course and content which provides a place to learn and teach without depending on the time and space boundaries" [22]. An LMS can also include applications and web-based technologies that can be utilised by HEIs and students to access, plan, implement, supplement, monitor, and assess learning [23]. Furthermore, an LMS can provide a discussion platform to facilitate communication among the students [23]. Different LMSs are commercially available, some of the most well-known are Moodle, Canvas, Blackboard, and Desire2learn. These LMSs may employ standard components for implementing learning activities. Learning management systems can hold course contents, provide discussion and chat forums, administer e-quizzes and e-assignments, create activity logs, summarise grades, to schedule, provide calendar reminders, and store multimedia files.

Hence, it can be concluded that LMSs can play a significant role in creating, distributing, tracking, and managing the learning activities widely used in HEIs to provide an innovative technology-based teaching and learning experience. In addition, LMS platforms are believed to have the potential to expand reach, reduce cost, and improve the quality of education, and thereby help HEIs to meet the demands of a growing student population. Thus, LMSs can be considered to be one of the numerous positive outcomes of digitising conventional campus teaching and learning functions.

However, despite the growing use of LMS solutions in higher education, researchers have also found some disadvantages associated with their service. Among these, lack of user engagement, poor intention to use, and low satisfaction are the major drawbacks cited in the literature. Other issues raised concerning LMS usage include ease of use, usage behaviour, expected performance, and attitude towards usage. These are essential factors that require the immediate attention of the research community. They have not been exhausted enough in terms of how they relate to the user performance and impediments that need to be handled to improve LMS usage.

IV. Attrition and Dropouts in E-Learning

The attrition rate in HEIs has become a noticeable problem since e-learning systems have been adopted. Although there is no concrete evidence as yet, according to the available literature, there are strong indications from the higher education sector that student attrition is higher because of e-learning systems [3, 24]. Also, somewhat ironically, although many institutions have implemented e-learning to meet learners' needs, a large percentage of learners do not complete e-learning courses [3, 24].

Some studies have taken time on covering various studies that have looked at the problem of high attrition in e-learning [25, 26]. Such studies have identified that the factors that cause attrition in e-learning are related to the students' characteristics and the design of the e-learning system. To date, the literature has focused primarily on investigating the student-related factors [26, 27], and less research has been dedicated to assessing whether there is a link between attrition and dropout rates and e-learning system designs. Moreover, the studies that do exist do not provide any overarching solutions [26, 27].

Nevertheless, it remains crucial to develop strategies to reduce the attrition rate in e-learning from an economic and a quality perspective. High attrition rates harm the financial performance and academic reputation of HEIs. Therefore, some studies have proposed student-centred learning as a solution for reducing attrition. In [28], the researchers studied the relationship between poor course design in e-learning and attrition. The results of their analysis indicated that course workload was a significant barrier and caused student dropout. In another study [29], the same researchers highlighted several strategies that could be used to overcome attrition. The learning systems should include good course design with greater flexibility to provide activities, structured course formats, time management support, syllabus quizzes, a feedback strategy, and collaboration between students and lecturers.

The technical aspect is believed to help the readiness of a student to accept online learning can improve their success in e-learning. For example, in a previous study [30], it was found that student success in online programs depends on the level of student engagement with information and communication technologies.

It can be argued that student success in e-learning has more to do with the student's relationship with technology. Technology's role in a student's life can determine how engaged they will behave in an e-learning environment [31]. It is essential for gamification comes into play. The researchers in [32] argued that gamified e-learning courses might build confidence and engage students in a course if they were familiar with the technology. Therefore, in the following paragraphs, student engagement factors are discussed.

In the literature, little has been said about the theoretical connections between gamification and engagement. However, in her book, "Reality is Broken", Jane McGonical has argued that games are not just for entertainment; but rather, the skills developed during games can be useful for solving real-life problems [33]. In other words, gamification can promote the engagement of users and help them to solve problems. For example, a study on gamification in the business domain showed that customers became engaged by combining the power of games with business strategies [34]. Moreover, in their book, "Total Engagement", Leighton and Baron explored the idea of using gamification to improve enthusiasm for work among employees [35]. That is to say that gamification can be used to engage people in the workplace. Furthermore, in the context of education, the authors of 24 peer-reviewed empirical studies on gamification found that most of the investigations yielded positive results when they applied gamification to improve learners' engagement [36].

V. Proposed Engagement Framework

In the literature, studies have used different game elements in e-learning for student engagement [37]. Hence, there is no consistency or consensus on which type of game element influences student engagement. However, from the literature review, several game elements are commonly used in such studies. As shown in Table II, these game elements are points, progress bars, levels, badges, Leaderboards, dashboards, content unlocking, teams, avatars, and timers. These ten game elements are therefore used in the proposed engagement framework. Descriptions of the selected game elements are provided in Table III.

The literature also shows that different learning activities in e-learning systems have been gamified for various purposes [24, 38, 39]. e-Learning systems offer several learning activities, but some are less frequently used in LMSs. Therefore, only those learning activities that may directly impact student engagement have been included in the proposed framework, namely, learning materials, assessment, assignment, and discussion.

TABLE II. Common Game Elements used for Engagement

Game elements	**Sources**							
	[18]	[19]	[11]	[12]	[20]	[16]	[21]	[17]
Badges	✓	✓	✓	✓	✓	✓		✓
Dashboard			✓					
Points		✓	✓	✓		✓	✓	✓
Levels		✓	✓	✓			✓	✓
Avatars		✓						✓
Teams			✓		✓			
Content unlocking		✓	✓	✓			✓	
Leaderboard		✓		✓		✓	✓	✓
Progress bar						✓		
Timer			✓					

TABLE III. DESCRIPTION OF THE COMMON GAME ELEMENTS

Game element	Description
Points	Points indicate the level of student achievement in various learning activities (e.g., quizzes, examinations, assignments, and discussion fora in e-learning systems [35].
Levels	Levels usually correspond to different modules or chapters of a course composed of activities such as course material, assignments, and assessments.
Badges	Virtual badges are given to students when they have completed specific activities, such as quizzes, assignments, and examinations.
Leaderboard	The Leaderboard is shown at the system level in a scoreboard that displays students' results based on the number of points and badges earned [11, 71].
Dashboard	The dashboard provides almost immediate feedback by presenting a summary of all the activities that have been completed by the student and those that the student has not yet completed. It also helps the student identify his or her expected performance outcomes by displaying analyses of the activities they have done on the e-learning system.
Progress bar	The function of the progress bar is to track and display how much progress a student has made in an e-learning activity. Based on the proposed engagement framework, almost all learning activities are expected to have a specific progress bar.
Avatar	When a student uses an avatar, it can make them feel safer as it helps to maintain privacy by hiding their identity and activities from others.
Teams	Teams are associated with specific assignments, and teams that complete their tasks on time are rewarded with badges, points etc.
Content unlocking	Content unlocking refers to moving up a level, where students proceed to a new level when they complete predetermined requirements.
Timer	The timer is used for quizzes and examinations and counts down the hours, minutes and seconds until the time limit that has been set for the completion of quizzes and examinations has been reached.

The learning materials activity refers to all the course-related documents submitted to and distributed by teachers in the e-learning system. The assessment activity involves using all of the exercises created by teachers for students' practical work and examinations and quizzes that are administered to test the students' level of knowledge acquisition, which are delivered through the e-learning platform [17]. The assignment activity allows students to submit assignments directly into the e-learning system [40]. Finally, the discussion activity in an e-learning platform enables students to engage in higher-order thinking and active learning and to have a social presence outside of the classroom environment [41].

Several engagement factors have been studied in the literature on gamification [17, 42, 43]. Behavioural, emotional, and cognitive factors were selected for inclusion in the proposed engagement framework. The main reason for choosing these three factors is that they are all widely used in the gamification of e-learning systems to investigate whether there is improvement in student engagement. There are also specific indicators that need to be used when studying factors [17, 43]. The factors and their corresponding indicators are shown in Table IV.

TABLE IV. ENGAGEMENT FACTORS AND THEIR CORRESPONDING INDICATORS

Factor	Indicators
Behavioural	Participation, collaboration, persistence, independent learning
Emotional	Interest, boredom, enjoyment, fun, curious
Cognitive	Deep understanding, critical thinking skills, competition, problem-solving

The proposed engagement framework is shown in Fig. 1. The framework consists of three elements or gamification components, learning activities, and student engagement factors. The game elements that influence certain learning activities are grouped and mapped to those activities.

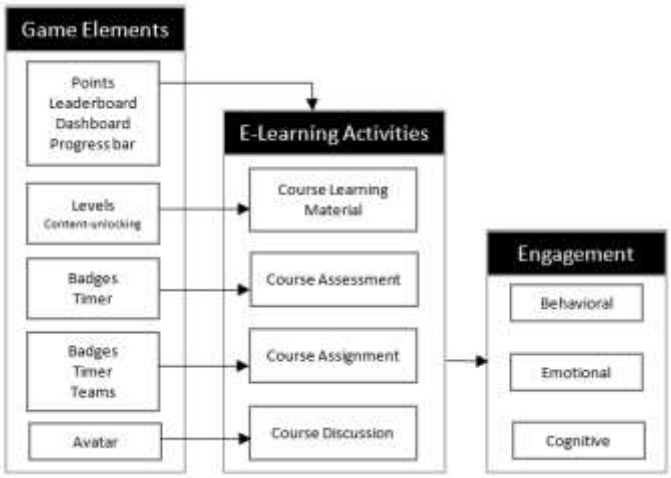


Fig. 1. Proposed Engagement Framework for the Gamification of e-learning Systems.

VI. EVALUATION OF THE FRAMEWORK

The proposed framework underwent a process of validation and evaluation to determine whether it could solve the issue of the lack of a standard for the gamification of e-learning systems. Generally, an assessment of a framework involves two activities: 1) Conducting a formative evaluation to assess the framework in design and development stage to find ways of improving it. 2) Conducting a summative evaluation to assess the framework after a prototype has been developed. This study uses the former type.

The evaluation was based on the views of experts. Gathering such views is necessary because it enables the researcher to draw upon knowledge that could not have been obtained otherwise through an analysis of the literature. Several evaluation methods have been used by previous studies that have proposed artefacts for the information systems. For this study, it was considered that the semi-structured interview method was the best option for the evaluation, which is in line with several other studies [44].

As the framework's design was based on an analysis of previous studies, it might be considered as self-reflection. So, to remove subjectivity from the process, a semi-structured interview was conducted. The experts who participated in the evaluation were selected based on their experience in teaching and the use of e-learning systems. Experts with a minimum of

10 years of teaching experience volunteered to participate in the evaluation. At the time of the interview, they were practising lecturers in higher education. The interview questionnaire was administered to the experts using Google Forms.

The game elements and learning activities in the framework and the engagement indicators selected to measure student engagement were evaluated by the experts for their suitability and completeness. The questions related to these aspects were as follows:

Q1. Do you agree that the customisation of the game elements for specific e-learning activities in the engagement framework will increase student engagement when using the e-learning platform?

Q2. Are there any missing game elements that you think would be important for gamifying the e-learning activities platform to increase student engagement? If yes, please state the game elements.

Q3. Are there any missing e-learning activities you think would be necessary for e-learning course platforms in higher education? If yes, please state the e-learning course activities.

Based on the literature, three engagement factors and their associated indicators for measuring the level of student engagement were included in the proposed framework:

- Behavioural: participation, persistence, collaboration, independent learning.
- Emotional: interest, boredom, enjoyment, fun, curious.
- Cognitive: deep understanding, competition, critical thinking skills, problem-solving.

The question that was posed to the experts in respect of these indicators was as follows:

Q4. Are these indicators adequate to measure student engagement in the e-learning activities platform?

Based on their feedback, overall, the evaluation of the experts was positive. Both experts (EX1 and EX2) unanimously agreed on the suitability of the game elements that were mapped to each of the learning activities in the framework and that these game elements would improve students' engagement in e-learning systems. EX1 also commented that customisation allows students to collaborate more friendly. Regarding the game elements' completeness, both experts indicated that the identified elements were enough to engage students in e-learning systems. Also, commenting on the completeness of the e-learning activities, both experts agreed that there were no missing activities and enough for the e-learning environment. Lastly, the experts agreed that the engagement indicators provided in the framework were adequate for measuring student engagement. Furthermore, EX1 commented that the indicators satisfied the Saudi Arabia national qualification framework.

The framework was also evaluated for its limitations, usefulness, ease of use, and clarity for utilisation in the gamification of e-learning systems. The questions that were posed in this regard were as follows:

Q5. From an expert evaluation perspective, are there any limitations to the framework? If any, please list the restrictions.

Q6. On a scale of 1 (worst) to 5 (best), how would you rate the framework's usefulness?

Q7. On a scale of 1 (worst) to 5 (best), how would you rate the ease of use of the framework?

Q8. On a scale of 1 (worst) to 5 (best), how would you rate the clarity of the framework?

Regarding Q5, which asked about the framework's limitations, the experts agreed that there were none. Likewise, to optimise the evaluation result of the framework, the output of the analysis of the usefulness, easiness of use, and clarity is shown in Table V.

TABLE V. RESULTS OF THE EVALUATION

Factor	Usefulness	Ease of use	Clarity
Positive	80%	83%	90%
Negative	20%	17%	10%

VII. DISCUSSION

The findings of the evaluation of the framework indicate that the experts are satisfied with the components of the framework and the elements and learning activities provided. The findings suggest that the framework is easy and clear to follow by e-learning developers when gamifying e-learning systems. The study showed that the gamified e-learning system using the framework is enough to improve student engagement. The result is in line with numerous studies, as reviewed in the related work, showing improvements in student engagement through gamification.

Moreover, this study reveals that engagement factors proposed in the framework, as commented by the experts, can be used to investigate whether a gamified e-learning system developed with the help of the framework can improve student engagement.

VIII. CONCLUSION

This paper proposed an engagement framework that can be used to develop gamified e-learning systems. The framework consists of three components: game elements, learning activities and engagement factors. The framework uses commonly used game elements and learning activities because it may become cumbersome for developers if more elements and learning activities are employed. The expert evaluation of the proposed framework showed that the framework is useful, easy to follow, and provides clear steps for the gamification of e-learning systems. Therefore, this framework is a value add for developers to stand as a reference that can be used for the gamification of e-learning systems.

In the future, an e-learning system will be developed where the selected game element will be added to it as per the directions and recommendations are given in the framework. Consequently, an experiment will be conducted where the gamified e-learning platform is exposed to a group of students to study their level of engagement.

ACKNOWLEDGMENT

This study was funded by the Research Grant (GPK-P&P-2020-005) and SOFTAM, FTSM, Universiti Kebangsaan Malaysia.

REFERENCES

[1] Alfaqiri, A.S., S.F.M. Noor, and N.S. Ashaari, Exploring indicators of engagement: applications for gamification of online training systems. Periodicals of Engineering and Natural Sciences (PEN), 2020. 8(4): p. 2096-2106.

[2] Khaleel, F.L., et al. Gamification-based learning framework for a programming course. in 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI). 2017. IEEE.

[3] Khaleel, F.L., et al., Gamification elements for learning applications. International Journal on Advanced Science, Engineering and Information Technology, 2016. 6(6): p. 868-874.

[4] Ab Rahman, R., S. Ahmad, and U.R. Hashim, A study on gamification for higher education students' engagement towards education 4.0, in Intelligent and Interactive Computing. 2019, Springer. p. 491-502.

[5] Ratna Zuarni Ramli, N.A.U.M.a.N.S.A., Microorganisms: Integrating Augmented Reality and Gamification in a Learning Tool. International Journal of Advanced Computer Science and Applications (IJACSA), 2021. Volume 12(Issue 6).

[6] Bouchrika, I., et al., Exploring the impact of gamification on student engagement and involvement with e-learning systems. Interactive Learning Environments, 2019: p. 1-14.

[7] Poondej, C. and T. Lerdpornkulrat, Gamification in e-learning. Interactive Technology and Smart Education, 2019.

[8] Alshammari, M.T., Evaluation of Gamification in E-Learning Systems for Elementary School Students. TEM Journal, 2020. 9(2): p. 806-813.

[9] Alsubhi, M.A., N.S. Ashaari, and T.S.M.T. Wook. The Challenge of Increasing Student Engagement in E-Learning Platforms. in 2019 International Conference on Electrical Engineering and Informatics (ICEEI). 2019. IEEE.

[10] Alfaqiri, A.S., S.F.M. Noor, and N.S. Ashaari. Employees' Engagement Issues in Online Training Applications. in 2019 International Conference on Electrical Engineering and Informatics (ICEEI). 2019. IEEE.

[11] Klock, A.C.T., et al. Gamification in e-learning systems: A conceptual model to engage students and its application in an adaptive e-learning system. in International Conference on Learning and Collaboration Technologies. 2015. Springer.

[12] Malas, R.I. and T. Hamtini, A gamified e-learning design model to promote and improve learning. Int Rev Comput Softw (IRECOS), 2016. 11: p. 8-19.

[13] De-Marcos, L., et al., Social network analysis of a gamified e-learning course: Small-world phenomenon and network metrics as predictors of academic performance. Computers in Human Behavior, 2016. 60: p. 312-321.

[14] Vanduhe, V., et al. Students' Evidential Increase in Learning Using Gamified Learning Environment. in Proceedings of the Future Technologies Conference. 2018. Springer.

[15] Özhan, Ş.Ç. and S.A. Kocadere, The effects of flow, emotional engagement, and motivation on success in a gamified online learning environment. Journal of Educational Computing Research, 2020. 57(8): p. 2006-2031.

[16] Ding, L., E. Er, and M. Orey, An exploratory study of student engagement in gamified online discussions. Computers & Education, 2018. 120: p. 213-226.

[17] Zainuddin, Z., et al., The role of gamified e-quizzes on student learning and engagement: An interactive gamification solution for a formative assessment system. Computers & Education, 2020. 145: p. 103729.

[18] Wongso, O., Y. Rosmansyah, and Y. Bandung. Gamification framework model, based on social engagement in e-learning 2.0. in 2014 2nd International Conference on Technology, Informatics, Management, Engineering & Environment. 2014. IEEE.

[19] Strmečki, D., A. Bernik, and D. Radošević, Gamification in e-Learning: introducing gamified design elements into e-learning systems. Journal of Computer Science, 2015. 11(12): p. 1108-1117.

[20] Katsigiannakis, E. and C. Karagiannidis, Gamification and game mechanics-based e-learning: a moodle implementation and its effect on user engagement, in Research on e-Learning and ICT in Education. 2017, Springer. p. 147-159.

[21] Mackavey, C. and S. Cron, Innovative strategies: Increased engagement and synthesis in online advanced practice nursing education. Nurse education today, 2019. 76: p. 85-88.

[22] Bervell, B. and V. Arkorful, LMS-enabled blended learning utilization in distance tertiary education: establishing the relationships among facilitating conditions, voluntariness of use and use behaviour. International Journal of Educational Technology in Higher Education, 2020. 17(1): p. 1-16.

[23] Elfeky, A.I.M., T.S.Y. Masadeh, and M.Y.H. Elbyaly, Advance organizers in flipped classroom via e-learning management system and the promotion of integrated science process skills. Thinking Skills and Creativity, 2020. 35: p. 100622.

[24] Khaleel, F.L., et al. The study of gamification application architecture for programming language course. in Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication. 2015.

[25] Monteiro, S., et al., Reducing attrition and dropout in e-learning: the development of a course design model. 2016.

[26] Tan, M. and P. Shao, Prediction of student dropout in e-Learning program through the use of machine learning method. International Journal of Emerging Technologies in Learning, 2015. 10(1).

[27] Suppan, L., et al., Effect of an E-learning module on personal protective equipment proficiency among prehospital personnel: web-based randomized controlled trial. Journal of medical Internet research, 2020. 22(8): p. e21265.

[28] Monteiro, S.J.d.A., Course design in e-learning and the relationship with attrition and dropout: A systematic review. 2017.

[29] Monteiro, S., et al., A systematic review of design factors to prevent attrition and dropout in e-Learning courses. 2017.

[30] Khaleel, F.L., et al. Methodology for developing gamification-based learning programming language framework. in 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI). 2017. IEEE.

[31] Alsubhi, M.A., N. Sahari, and T.T. Wook, A conceptual engagement framework for gamified e-learning platform activities. International Journal of Emerging Technologies in Learning (iJET), 2020. 15(22): p. 4-23.

[32] Dray, B.J., et al., Developing an instrument to assess student readiness for online learning: A validation study. Distance Education, 2011. 32(1): p. 29-47.

[33] McGonigal, J., Reality is broken: Why games make us better and how they can change the world. 2011: Penguin.

[34] Xi, N. and J. Hamari. The relationship between gamification, brand engagement and brand equity. in Proceedings of the 52nd Hawaii International Conference on System Sciences. 2019.

[35] Reeves, B. and J.L. Read, Total engagement: How games and virtual worlds are changing the way people work and businesses compete. 2009: Harvard Business Press.

[36] Hamari, J., J. Koivisto, and H. Sarsa. Does gamification work? a literature review of empirical studies on gamification. in 2014 47th Hawaii international conference on system sciences. 2014. Ieee.

[37] Khaleel, F.L., et al. Smart Application Criterion based on Motivation of Obese People. in 2019 International Conference on Electrical Engineering and Informatics (ICEEI). 2019. IEEE.

[38] Liu, C., Gamification in Academia Practice–What Motivate Users Most.

[39] Azmi, M.A. and D. Singh, Schoolcube: gamification for learning management system through microsoft sharepoint. International Journal of Computer Games Technology, 2015. 2015.

[40] Armier Jr, D.D., C.E. Shepherd, and S. Skrabut, using game elements to increase student engagement in course assignments. College Teaching, 2016. 64(2): p. 64-72.

[41] Ding, L., C. Kim, and M. Orey, Studies of student engagement in gamified online discussions. Computers & Education, 2017. 115: p. 126-142.

[42] Osipov, I.V., et al., Study of gamification effectiveness in online e-learning systems. International Journal of advanced computer science and applications, 2015. 6(2): p. 71-77.

[43] Ibanez, M.-B., A. Di-Serio, and C. Delgado-Kloos, Gamification for engaging computer science students in learning activities: A case study. IEEE Transactions on learning technologies, 2014. 7(3): p. 291-301.

[44] Adams, R., The advanced data acquisition model (ADAM): a process model for digital forensic practice. 2012, Murdoch University.

Application of Deep Learning in Satellite Image-based Land Cover Mapping in Africa

Challenges, Emerging Solutions and Prospects: A Review

Nzurumike Obianuju Lynda[1], Ezeomedo Innocent C[2], Nwojo Agwu Nnanna[3], Ali Ahmad Aminu[4]
Department of Computer Science, Nile University of Nigeria, Abuja, Nigeria[1, 3, 4]
Department of Environmental Management, Chukwuemeka Odumegwu Ojukwu University, Uli, Anambra, Nigeria[2]

***Abstract*—Deep Learning Networks (DLN), in particular, Convolutional Neural Networks (CNN) has achieved state-of-the-art results in various computer vision tasks including automatic land cover classification from satellite images. However, despite its remarkable performance and broad use in developed countries, using this advanced machine learning algorithm has remained a huge challenge in developing continents such as Africa. This is because the necessary tools, techniques, and technical skills needed to utilize DL networks are very scarce or expensive. Recently, new approaches to satellite image-based land cover classification with DL have yielded significant breakthroughs, offering novel opportunities for its further development and application. This can be taken advantage of in low resources continents such as Africa. This paper aims to review some of these notable challenges to the application of DL for satellite image-based classification tasks in developing continents. Then, review the emerging solutions as well as the prospects of their use. Harnessing the power of satellite data and deep learning for land cover mapping will help many of the developing continents make informed policies and decisions to address some of its most pressing challenges including urban and regional planning, environmental protection and management, agricultural development, forest management and disaster and risks mitigation.**

Keywords*—Deep learning; satellite image classification; land cover mapping; Africa

I. INTRODUCTION

Africa is the world's second largest and second most populous continent in the world with an estimate population of one billion, three hundred (1,300,000,000) people. She has a total land area of approximately thirty million, three hundred and sixty-five thousand square km (30,365,000). Common land cover/use classes or physical features which describe the usage of the land areas include water bodies, forest, agricultural land, barren land, built-up/settlements, etc.

The rate of population growth, socio-economic activities, urbanization, and other environmental forces has led to the exploitation, degradation and subsequent altering of these land features [1]. All these described above have led to the constant changing features in the land cover as illustrated by K. Kalra et al in[2] showing land cover change of Abuja-Nigeria from 1987, 2002, and 2017 as shown in Fig. 1.

These changes have contributed to the enormous challenges it faces today, including food scarcity, degradation of habitats, outbreaks of epidemics, environmental hazards, and consequently, climate change and global warming. The impact of these challenges is complex and has an adverse effect on the people, the economy, and the environment causing a lot of concerns to both individuals and the governments, locally and internationally.

To mitigate the effects of these enormous challenges, it is critical to map the earth surface to collect information on the status of the different classes of bio-physical cover of the earth surface and its changes over time, otherwise known as land cover mapping. The data obtained also provides valuable information for developing sustainable policies or strategies to mitigate the impacts as well as meet the increasing demands for basic human needs and welfare. Land cover mapping is regarded as one of the most important application in remote sensing.

Globally, scientists spend a greater percentage of their time and money on fieldwork mapping the land covers. In Africa this fieldwork typically involves traveling to distant and sometimes isolated locations for long period of time. It also requires braving adverse weather conditions, inaccessible terrains, security risks and performing physically demanding tasks. With the growing demand for real-time land information, it is becoming increasingly difficult to physically visit every location. While carrying out field surveys is still fundamental, the negative costs of field surveys, slow speed of generation of geo-information, objectivity of human interpretation of field observations and inconsistent land cover maps are some of the factors limiting the total reliance on field surveys.

Recently images of the earth surface acquired with highly-accurate remote sensing technology, operated by the government or commercial businesses around the world, have been seen to contain detailed information of the earth's land cover. Its proliferation over the past few decades has given a radically improved understanding of our planet's landforms, vegetation, and resources. Today, it is considered one of the most important data sources for earth observation because of its ability to cover very large area. It has the potential to provide more accurate, reliable and faster land cover information. Additionally, it reduces the manual effort required to conduct field surveys.

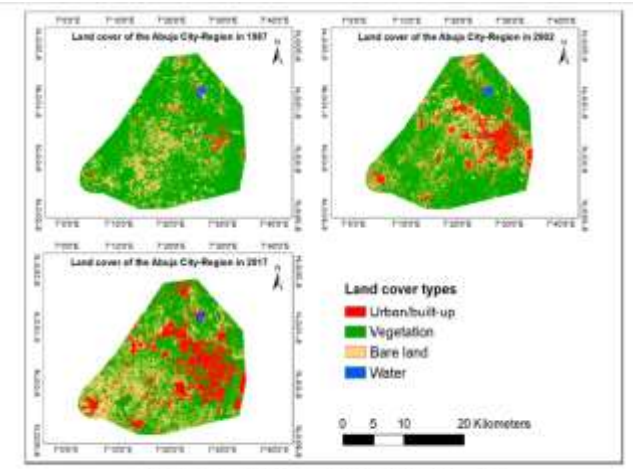


Figure 4. Land cover of Abuja city-region in 1987, 2002, and 2017.

Fig. 1. Land Cover of Abuja-Nigeria in 1987, 2002, and 2017 [2].

The first satellite, Landsat 1, initially named Earth Resources Technology Satellite (ERTS) was launched in the United States on July 23, 1972. Subsequently, hundreds of earth observation satellites with varying resolutions including Landsat 2-8, Goggle Map, Quick bird, SPOT IKONOS, worldview, big map have been launched to collect massive amounts of remote sensing images. Nevertheless, unlocking the rich information in raw satellite images for land cover mapping, requires accurate, timely and reliable analysis.

The advent of Machine Learning (ML), a subset of Artificial Intelligence (AL) has brought a lot of hope to computer vision applications. It is concerned with algorithms and techniques that allow computers to analyze data, learn from the data and make prediction on new data using computational and statistical methods. Several papers have been published introducing the basic concepts of ML to the remote sensing and earth science community. However, earlier experiences with automated classification in Mali, Senegal and Niger produced disappointing results [2]. Important land use and land cover types, such as agriculture in the Sahel, could not be uniquely differentiated from other types based on their spectral reflectance properties. Automated methods of image classification are based on spectral image data and are often plagued by problems of misclassification. Spectral reflectance of land surfaces is measured by remote sensors may be quantitative but not absolute which makes them not necessarily unique. This homogeneity often leads high variation within various land covers and subsequently poses a huge challenge in mapping and analyzing land cover types based solely on their spectral properties.

As the revolution and growth of ML continued to bridge the gap between the capabilities of humans and machines, it gave birth to Deep learning, the fastest growing field of machine learning. Compared to traditional ML techniques, it has impressive learning capability, does not require expertise or domain knowledge and is time-saving. Its potentials in remote sensing tasks have been ascertained by many researchers in developed countries for solving problems in the domain of geological mapping, land cover mapping, land use planning, geological image classification, infrastructural development, mineral resources exploration, etc. Despite its broad use in developed countries, performing land cover classification with deep learning has been very challenging in developing countries. Recently, new approaches have yielded significant breakthroughs, offering novel opportunities for its further research and development. This can be taken advantage of in low resources environments such as Africa; however, the shortage of scientific papers discussing such emerging approaches or concepts in an easy and commonly understood way remains one major obstacle to its application.

Inspired by the above, the main contributions of this paper are to:

1) Provide a brief overview of a typical DL model.

2) Provide a brief overview of satellite image classification with CNN.

3) Discuss the challenges it poses to developing countries.

4) The emerging solutions for satellite image classification in low resource environments are further discussed.

5) Finally, a discussion on the potential of DL technique in land cover mapping in Nigeria is provided.

II. Overview

A. Overview of Satellite Image-land Cover Classification with CNN

This section gives an overview of the processes involved in classification of satellite images using CNN.

Land cover classification from satellite images is the process of assigning a land cover class or theme to each pixel in a satellite image based on its features otherwise known as spectral information. To achieve this task with deep learning, a DL model usually a convolutional neural network model is designed. A set of pre-labeled training data is fed into the CNN model. This model attempts to learn the visual features present in the training images associated with each label, and classify unlabeled or unseen images accordingly. The steps below describe the processes from the input image to classification result:

1) Design the architecture for the CNN model: The first step in land cover classification with CNN is to design the CNN Architecture. There are two ways by which this can be achieved: you can either design a CNN model from scratch (if you have a large dataset and good computational resources) or repurpose/redesign an already existing model using transfer learning technique. In the first case you choose the type of layers and the way they will be arranged and connected to each other. While in the latter case, you choose an existing architecture known as pre-trained model, already trained on large datasets and continue your design from it. Designing a CNN architecture from scratch can be challenging and time-consuming. This is because it requires expert knowledge and effort due to the large number of architectural design choices [3].

2) Hyperparameter selection and optimization: At this step, configuration variables that determine the network architecture or how the network is trained are selected,

initialized and optimized. These variables are known as parameters and hyperparameters and they play a big role in the design time and prediction accuracy of the classification task. Unlike a model's parameter which are learnt or estimated from data during training, the hyperparameters must be selected, and optimized before and during the training process, respectively. Usually, an empirical process that involves a lot of trial and error is used to optimize these hyperparameters. However, it is time-consuming requires expertise from the domain and, in some cases, where the number of hyperparameters in a CNN is so large, it is difficult to optimize manually.

Although a variety of algorithms including grid search, random search, Bayesian Optimization, Gradient-based Optimization, Evolutionary Optimization have been proposed and used for optimizing the values of these hyperparameters, the development of efficient hyperparameter optimization algorithm still remains a huge research area.

Common hyperparameters, their functions and common values are shown in Table I.

3) Select dataset: An important step in landcover classification with CNN is to source for your dataset. These datasets comprise of a collection of N images, each labeled with one of K distinct classes. These datasets are usually divided into three: training set, validation set and Test set commonly in the ratio of 80% 10% 10% respectively. These data can be pre-processed (resizing, rescaling, etc.) To train a high performing CNN model requires large quantity of dataset as well as equal distribution of the known classes in the dataset (balanced dataset).

4) Model training: The training phase is when the network "learns" from the data it will be fed with. CNN models are trained using an optimizer (optimization algorithm) through a back-propagation process. The weights and biases in the network are changed to minimize a cost function. The errors between the network output and the ground-truth value are calculated by a predefined loss function. The larger the loss function the farther we are from the correct answer. An optimizer updates the model's parameter based on the output of loss function with the goal of reducing the loss function as much as possible. The errors are back propagated based on the partial derivatives after which each weight and the corresponding error term are adjusted. The iterative process is illustrated by the steps below:

- Input a batch of Data.
- Forward propagate it through the neural network.
- Get the loss (i.e., the Actual value or the predicted value).
- Back Propagate to calculate the gradient.
- Update the weights parameter using the gradient.
- The above process is performed in an iterative way.
- Until the network has learned or converged.

TABLE I. HYPERPARAMETERS, FUNCTIONS AND COMMON VALUES

S/N	Hyper Parameters	Functions	Common values for classification tasks
1	Optimizer	Controls the learning rate by adjusting it throughout the training.	RMSprop gradient descent, Stochastic gradient descent, Mini-batch gradient descent. ADAM
2	Batch Size	The number of training examples in one forward/backward pass	32, 64
3	Epoch	Indicates the number of times the entire training dataset the machine learning algorithm has completed.	10, 30
4	Loss Function	It measures the absolute difference between our prediction and the actual value. It is use to evaluate how well your model is working.	Categorical _cross entropy Mean squared error
5	Learning rate	Determines how fast the optimal weights for the model are calculated	0.01, 0.001, 0.0001

The forward and backward process is illustrated in Fig. 2.

This is done using an optimization strategy such as gradient descent. The gradient descent algorithm simply measures the change in all weights with regards to change in error. This is illustrated in the formula 1.

$$w^{+} = w - \eta \nabla c \quad (1)$$

where $[w^{+}]$ is the new weight, w is the current weight, η denotes the learning rate, a parameter that determines how much an updating step influences the current value of the weights, i.e., how much the model learns in each step. ∇c is the gradient of the cost function. Gradient can be thought of as a slope of a function. The higher the gradient, the steeper the slope and the faster a model can learn. If the slope is zero, the model stops learning. In calculating the error of the model during the optimization process, a loss function must be chosen.

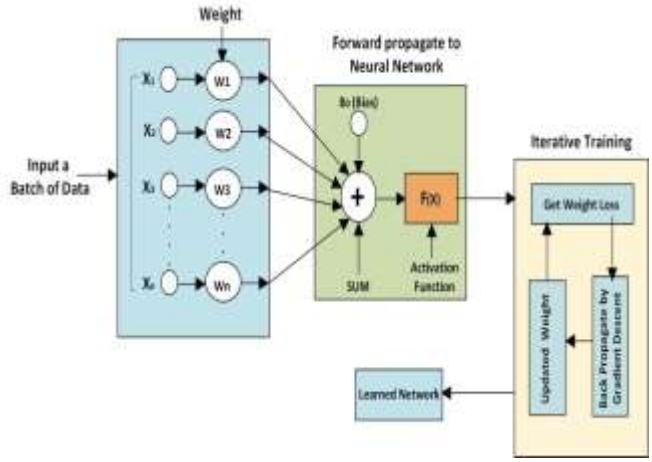


Fig. 2. FModel Training Iterative Process.

Training is commonly done using the Python programming language with either Goggle TensorFlow or Pytorch. Both are extended by a variety of APIs, cloud computing platforms, and model repositories.

5) Model evaluation: At this step, the model accuracy is evaluated. Confusion matrix, accuracy, precision, recall and F1 are commonly used metrics to quantitatively evaluate the performance of a model. These metrics are adopted as a classification performance indicator by the research community. The validation and testing datasets are used to evaluate the model. While the validation dataset is used to give an estimate of a model's performance while fine tuning the model's hyper parameters, the test set is used to provide an unbiased evaluation of the final model fit on the training dataset. A well-performing trained neural network is expected to correctly predict the labels of unseen images.

III. Challenges, Emerging Solutions and Potentials of Satellite Image Classification with CNN in Nigeria

A. Challenges

Although deep learning has been around for decades, its use among developing countries is still relatively new. This section discusses some of the challenges to using deep learning for satellite image classification tasks in scare resource environments:

1) Lack of training data: Despite deep learning's powerful feature extraction capability, in practice it is difficult to train CNN models with small quantity of datasets[4]. This is because a deep neural network has so many layers with many nodes in each layer, which results in exponentially many more parameters to tune. Without enough data, we cannot learn from these parameters efficiently. Unlike the huge number of everyday training images such as clothes, furniture, cars, animals etc. used in popular deep learning models, satellite image data is expensive to obtain. Therefore, there are no sufficient labeled satellite image patches to train a very deep network. The lack of the right quantity of training data in remote sensing domain is a significant obstacle to utilizing the full power of deep learning models especially in developing countries[5].

2) Lack of technical expertise: Technical expertise is an important factor in DL for satellite image classification research, development and application. Taking full advantage of this emerging technology, requires considerable technical expertise. This is because satellite image-based land cover mapping with DL lies in the intersection of distinct disciplines such as information technology, remote sensing, mathematics, image processing, programming etc., which means that, a fairly high skill level is required for this area.

Unfortunately, finding the right professionals with the appropriate skill combination is a huge task in Africa [6]. Furthermore, this domain is very dynamic, with new technologies emerging daily, making it difficult for professionals to remain abreast of latest trends. This challenge can put this technique out of the reach of many researchers and operational experts that wish to use this emerging technology in land cover mapping.

3) Image quality (spatial and spectral resolution): Two fundamental characteristics of a satellite image are its spatial and spectral resolution. They are important factors that play a role in determining the performance of a satellite image-based classification task using DL. The Spatial resolution refers to the smallest size an object can be depicted in an image. It is usually presented as a single value representing the length of one side of a square. While spectral resolution refers to the number and dimension of distinct wavelength intervals (bands) of an electromagnetic radiation a sensor is capable of measuring.

The higher these resolutions mean the more detail it can provide and the higher the cost of obtaining them. Purchasing the images is very expensive and unaffordable in most developing countries. Most researchers in African have made use of freely accessible satellite images from earth observation programs such as European space Agency's (ESA) and the United States National Aeronautics and space Administration's (NASA)[7], such images usually have limited spatial and spectral resolution and as such affects the degree of accuracy. Lack of access to high-resolution images is a major limiting factor to DL use in land cover classification.

In addition, some researchers in Africa use images from different sources with varying resolution. Despite its usefulness, the extraction of information from such multi-resolution sources using deep learning technique is still a grey area in the research world.

4) Lack of image pre-processing techniques: Satellite data often contain cloud cover, noise and other distortions from imaging systems, sensors, and observing conditions. Therefore, further pre-processing is required to deal with the defects. Image processing is a method to perform some operations on an image, in order to improve the image data (features) by suppressing unwanted distortions and/or enhancement of some important image features to extract some useful information from it. Despite its high potential value to DL accuracy, the automation of data preprocessing has been mostly overlooked by the machine learning community [8]. The dearth of methodological knowledge in this area continues to be a limiting factor to DL usage especially in developing continents.

5) Lack of data repositories: In most developed countries several remote-sensing image datasets were introduced by different groups to enable machine-learning based research development and application for satellite image classification [9]. These dataset repositories include Merced (UCM), PatternNet, NWPU-RESISC45, Aerial Image Dataset (AID), EuroSAT and most recently BigEarthNet. Labeling of satellite images is very expensive since it requires trained professionals, so optimally leveraging existing data repository is essential. Unfortunately, such data repositories are not available in Africa and this has hindered the research, development and application of DL in satellite image classification tasks.

The heterogeneous appearance of satellite images, variation in geography and difference in spatial details of different images limits the use of such data repositories in developing countries.

6) Computational resources limitation: A typical DL application requires huge computational resources owing to (i) the large amount of multiply and accumulate (MAC) operations as well as memory access operations it executes (ii) the huge number of parameters (weights) to learn (iii) repeated model optimization to select the best hyperparameters for optimal performance.

In the last decade, the non-availability of computers with such high processing capability posed a huge challenge to the development and application of deep learning. The birth of Graphic processing units (GPUs) has given hope to machine learning enthusiasts. This is as a result of their highly parallel structure for distributed computational processing, large memory bandwidth to accommodate large dataset and other GPU resources which allows them to handle the processing of large amount of data faster and more efficient. However, due to the cost of GPU's, this powerful supercomputer is widely out of reach of most DL researchers in developing countries.

The widely available CPUs cannot be sufficiently relied on for training a deep learning model. Consequently, non-availability of computational resources remains a huge challenge in the development and application of deep learning in developing continents.

7) Political and economic constraints: Currently, Nigeria lack access to emerging tools and techniques for satellite image-based classification tasks with DL due to budgetary constraints, licensing issues or bandwidth limitations. Although significant investments in geo-information technologies have been made in the past by multilateral and bilateral aid projects from developed countries, national governments in Africa have generally not supported RS applications development [6]. Most universities and government organizations in Africa are poorly funded and thereby do not have the capacity to utilize emerging technology in remote sensing [10].

B. Emerging Solutions

A number of notable attempts have been made in the past to solve some of the challenges mentioned above. This section discusses some of these emerging solutions to satellite image classification using Deep learning.

1) Augmentation: The classification performance of DL models highly relies on the training procedures and the quantity of diverse training data. Data Augmentation is a very powerful method that offers the opportunity to solve the problem of inadequate dataset to train accurate and robust classifiers. It is used to artificially expand the size of a training dataset. It creates variations of the images using a range of operations. The augmented data will represent a more comprehensive set of possible data points which helps to improve the ability of the model to generalize what they have learned to new images and prevent overfitting [11]. Basic augmentation approaches which can make a model invariant to changes in size, translation, occlusion, viewpoint, illumination etc. include:

a) Reformation of original images by basic operations such as cropping, stretching, rotation of the image, image rescaling, horizontal and vertical flips, etc.

b) Altering the intensities of the RGB channels of raw data proposed by Krizhevsky et al. [12].

c) Random Erasing introduced by Zhong et al [13]. This technique is specifically aimed at proffering solution to the problem of occlusion which is a limiting factor to the generalization ability of CNN's. An occlusion occurs when part of an image is closed up or blocked off.

d) Generative Adversarial Networks (GANs), proposed by C. Bowles et al. [14], which offers a way to unlock additional information from a dataset by generating synthetic samples with the appearance of real images.

Most of these operations are useful in remote sensing because they do not increase the spectral or topological information in the satellite images which is important for consistent classification result. In most research papers, the experimental results with augmentation operations outperformed those from the same deep model architecture training on the original dataset. Recently there has been extensive use of data augmentation to improve CNN task performance [15].

2) Transfer learning (TL): Transfer learning is one of the most emerging design methodologies for also solving the challenge of limited training data. It involves training a CNN model on a base dataset for a specific task and then using those learned features/ model parameters as the initial weights in a new classification task. Instead of starting the learning process from scratch you start from patterns that have been learnt while solving a different task. It relies on the fact that features learned in the lower layers of a CNN, like edges or curves or color may be general enough to be useful for other classification tasks. By transfer learning we are able to take advantage of the expensive resources (expertise, training data and computational power) that were used to acquire it.

To mitigate the problem of limited labelled training data in remote sensing, most researchers have used transfer learning to leverage on either satellite image data repositories such as UC Merced (UCM) [16], EuroSAT [7] and more recently BIGEARTHNET [17] or using natural image data repositories such as ImageNet or CIFAR-10. For instance, R. P. De Lima et al [18] successfully used transfer learning to address a suite of geologic interpretation tasks. M. Xie et al. [4] used a sequence of transfer learning steps to design a novel ML approach using satellite image.

3) Improvement in architectural design: As seen in Fig. 1, a typical CNN architecture is formed by the stacking of multiple and non-linear processing neurons in layers. Different variations of CNN architectures have been proposed in an attempt to solve some of the challenges faced in using DL in less resource environments. This includes:

a) Improving the layer structure of CNN's to accommodate the limited training data [19].

b) Encoding rotational equivalence in the network structure [20]. Rotation Equivariance is the ability of a network to generalize feature detection in different locations, which is key to the generalization ability of CNN's.

c) Capsule network (CapsNet) [21] and light convolutional neural network (LCNN) [22], were designed to suite low computational resources and the small number of training samples.

d) Other notable techniques incorporated in CNN architectures include regularization, batch normalization, addition of shortcut connections between layers etc.

4) Improvement in image quality: To solve the problem of non-availability of high-resolution images, use of images with different resolution, poor-quality images due to varying distortions. Researchers have successfully used different techniques such as:

a) Super resolution (SR) techniques- These techniques are used to enhance or produce a high-resolution image from one or two low resolution images using algorithmic means. Several SR techniques have been proposed and successfully used.

b) Multi-source data fusion- These techniques combine data from multiple sources to produce a high-quality visible representation of the data to improve model training effort [23].

Improving image quality using various techniques remains an active area of research.

5) Computational resources optimization: To solve the challenge of computational resources which is a limiting factor to the use of DL, a number of notable solutions has been proposed and successfully used. These techniques/approaches can be beneficial to researchers in developing countries.

a) Hyperparameter optimization: it has become popular to use different optimization strategies to design network architectures that are computational efficient to train [1].

b) Algorithmic Design: reducing computation resources consumption through CNN model designs [24].

c) Cloud-powered DL: This is another active research area [24]. It allows deep learning training and evaluation on cloud-based platforms, thereby limiting the need for high-end computational resources.

6) Open-source developmental tools: The emergence of open-source developmental tools and libraries such as Pytorch, Caffe, Keras, TensorFlow, Theano for developing and training models are also speeding up progresses in deep learning [24] and will be also helpful in developing deep learning models. These tools and libraries allow the re-use of existing state of the art models thereby taking advantage of the expensive resources (expertise, training data and computational power) used in their development. Other usage of these open-source developmental tools includes:

a) Ease of configuring the learning process

b) Ease of plotting accuracy graphs which provides useful insights about the training of the model.

c) The capability to register callbacks when training a deep learning model. (This function allows saving model weights when the model's performance on tracked metrics improves or view a model while it's training).

GitHub [25], a web-based collaborative platform for software developers is another interesting development which can aid the application of DL in satellite image classification tasks in less resource areas. It is a free to use platform which allows developers to collaborate on a project. African researches can take advantage of the platform to learn through shared resources in the platform. It also provides cloud resources, programming tools, etc.

C. Potentials of Land Cover Mapping Applications

The benefits of using intelligent systems such as deep learning to automatically classify satellite images are becoming more evident in land cover mapping applications. These applications will aid the management of long-term challenges faced in Africa. In this section, we present an overview of some land cover mapping applications.

1) Land use planning: Land use planning is the process of allocating and reallocating land resources for different purposes such as infrastructure distribution, recreational and industrial use, transportation routes etc. This is usually done by the government to ensure efficient usage of the land resources for orderly development. Although humans have been modifying land to obtain food and other essentials for thousands of years, the current frequency and intensities of LULC changes in African are far greater than ever in history, driving unprecedented changes in ecosystems and environmental processes at local, regional and global scales [26]. The use of deep learning techniques for automatic land cover classification is very critical for studies involving extensive land mass such as found in most part of Africa, Other benefits include its cost and time saving as opposed to other machine learning techniques. Additionally, the digital information is also easy to retrieve, update, edit and store.

2) Agricultural development: Despite the conscientious effort by the government to increase the gross domestic product (GDP) in Africa, the population of malnourished people is still on the increase. This menace can be salvaged by expanding food production through agriculture using satellite images and deep learning technique. Ways through which the above can be achieved include:

a) Classifying agricultural land types

b) Predicting crop yield

c) Classifying crop disease types

d) Monitoring crop growth condition

e) Identifying and mapping weeds

3) Population estimation: Population estimation involves obtaining a reliable estimate of the number of persons in a given area at a particular time. Accurate mapping of population distribution is essential for policy-making, urban planning, admiration and risk management in hazardous areas [27]. Although census (a count of the population) has been used in the past to collect this important information in Africa, the high cost of carrying out the exercise has made it difficult to sustain. For instance, in Nigeria, the last census was conducted in 2006. The automatic classification of land cover using satellite images and DL techniques will provide a cost-effective alternative for estimating the population. It will also allow for large coverage mapping. As seen in [28], the CNNs model's estimate gave comparable results to the census county-level population projections. The population can be estimated through mapping the number of settlement areas, the area of urban land, the area occupied by different land uses or even directly from spectral and textural information available at the pixel scale[27].

4) Geoscience: In recent years deep learning networks has become an increasingly important interdisciplinary tool that has advanced several fields such as healthcare, image processing, speech recognition etc. however, its adaptation in geoscience has been relatively low[29]. Mapping large land areas with satellite data and deep learning technique can be applied in various geoscience tasks including:

a) Explore and prepare maps quickly to help evaluate the geo-potential of any specific area.

b) Study the general physical characteristics of rocks (lithology).

c) Geological and geo-structural mapping

d) Mineral exploration

e) Borehole drilling

f) Geo- hazard monitoring

5) Disaster management: According to a world bank report "Developing countries suffer more than 95 percent of all deaths caused by Natural disaster and losses (as a percentage of GDP) are 20 times greater. Africa has experienced its fair share as a result of their high population densities, poor infrastructure, unstable landforms and severe weather condition. These Disasters range from natural to man-made such as drought, floods, earthquakes, desertification, climate change, etc.

Satellite images cover wide range of areas and provides massive amount of land cover information. Analysis of these images using deep learning technique is imperative for effective mitigation and management of these risks.

Its uses in disaster management include:

a) Mapping fire, flood and hazard prone areas for disaster monitoring and mitigation.

b) Weather forecasting.

c) Disaster response planning.

d) Impact/ Damage assessment.

IV. Conclusion

Africa is confronted with serious developmental challenges arising from unplanned and unguided use of its land resources. It is therefore critical to provide accurate and up-to-date information of its land surface. The benefits of using intelligent systems such as deep learning to map massive large land area is becoming more evident. This information will aid the management of the long-term challenges faced in Africa. Furthermore, it will provide a better cost effective and time management solution than the use of visual interpretation or other machine learning techniques (unsupervised, supervised and object-based) currently obtainable in Nigeria today. This work aims to deepen the understanding of application of CNN for satellite image- based land cover mapping in developing continents, and encourage African researchers/scientist to leverage these new digital technologies to drive large-scale transformation and competitiveness in earth observation application.

References

[1] S. M. Herrmann, M. Brandt, K. Rasmussen, and R. Fensholt, "Accelerating land cover change in West Africa over four decades as population pressure increased," Commun. Earth Environ., vol. 1, no. 1, pp. 1–10, 2020.

[2] E. C. Enoguanbhor, F. Gollnow, J. O. Nielsen, T. Lakes, and B. B. Walker, "Land Cover Change in the Abuja City-Region , Nigeria : Integrating GIS and Remotely Sensed Data to Support Land Use Planning," 2019.

[3] S. Albelwi and A. Mahmood, "A framework for designing the architectures of deep Convolutional Neural Networks," Entropy, vol. 19, no. 6, 2017.

[4] R. P. De Lima and K. Marfurt, "Convolutional Neural Network for Remote ‐ Sensing Scene Classification : Transfer Learning Analysis," 2020.

[5] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping," 2014.

[6] M. E. Brown, K. Lance, E. Khamala, B. Siwela, A. Adoum, and M. Brown, "Review of Remote Sensing Needs and Applications in Africa Prepared by : Contributors :," no. October, 2015.

[7] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," Int. Geosci. Remote Sens. Symp., vol. 2018-July, pp. 204–207, 2018.

[8] T. N. Minh, M. Sinn, H. T. Lam, and M. Wistuba, "Automated Image Data Preprocessing with Deep Reinforcement Learning," pp. 1–9, 2018.

[9] L. Cover, L. Use, and D. Mining, "Benchmarking Deep Learning Frameworks for the Classification of Very High Resolution Satellite Multispectral Data," vol. III, no. July, pp. 83–88, 2016.

[10] T. Woldai, "The status of Earth Observation (EO) & Geo-Information Sciences in Africa–trends and challenges," Geo-Spatial Inf. Sci., vol. 23, no. 1, pp. 107–123, 2020.

[11] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," J. Big Data, 2019.

[12] B. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," 2012.

[13] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation."

[14] C. Bowles et al., "GAN Augmentation : Augmenting Training Data using Generative Adversarial Networks."

[15] L. Taylor and G. Nitschke, "Improving Deep Learning using Generic Data Augmentation," no. October, 2017.

[16] G. J. Scott, R. A. Marcum, C. H. Davis, and T. W. Nivin, "Fusion of Deep Convolutional Neural Networks for Land Cover Classification of High-Resolution Imagery," IEEE Geosci. Remote Sens. Lett., vol. 14, no. 9, pp. 1638–1642, 2017.

[17] G. Sumbul, M. Charfuelan, V. Markl, and D. Gmbh, "Bigearthnet : A Large-Scale Benchmark Archive for Remote Sensing Image Understanding Technische Universit ¨," pp. 2–5, 2018.

[18] R. P. De Lima, A. Bonar, D. D. Coronado, K. Marfurt, and C. Nicholson, "Deep convolutional neural networks as a geological image classification tool," no. Dl.

[19] L. Alzubaidi et al., Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, vol. 8, no. 1. Springer International Publishing, 2021.

[20] Y. Zhou et al., "Remote sensing scene classification based on rotation-invariant feature learning and joint decision making," 2019.

[21] W. Zhang, P. Tang, and L. Zhao, "Remote Sensing Image Scene Classification Using CNN-CapsNet," Remote Sens., vol. 11, no. 5, p. 494, 2019.

[22] H. Song, Y. Kim, and Y. Kim, "A Patch-Based Light Convolutional Neural Network for Land-Cover Mapping Using Landsat-8 Images," pp. 1–19, 2019.

[23] J. Zhang, "Multi-source remote sensing data fusion : status and trends," vol. 9832, 2010.

[24] C. Chen et al., "Deep Learning on Computational-Resource-Limited Platforms: A Survey," Mob. Inf. Syst., vol. 2020, 2020.

[25] A. C. Martin, "An introduction to Git and GitHub," pp. 1–25, 2018.

[26] I. Y. Dadson, "Land Use and Land Cover Change Analysis along the Coastal Regions of Cape Coast and Sekondi," Ghana J. Geogr., vol. 8, no. 2, pp. 108–126, 2016.

[27] S. Mossoux, M. Kervyn, H. Soulé, and F. Canters, "Mapping population distribution from high resolution remotely sensed imagery in a data poor setting," Remote Sens., vol. 10, no. 9, 2018.

[28] C. Robinson and F. Hohman, "Satellite Imagery," vol. 1996, no. 1, pp. 1–16.

[29] Y. Of, M. Learning, I. N. Geoscience, and I. N. Review, "Years of Machine Learning in Geoscience In Review," 2020.

Semi-supervised Deep Learning for Stress Prediction: A Review and Novel Solutions

Mazin Alshamrani
The Custodian of the Two Holy Mosques Institute for Hajj and Umrah Research (HURI)
Umm Al Qura University, Makkah, Saudi Arabia

***Abstract*—This research introduces a novel self-supervised deep learning model for stress detection using an intelligent solution that detects the stress state using the physiological parameters. The first part of this research represents a concise review of different intelligent techniques for processing physiological data and the emotional states of humans. Also, for all covered methods, special attention is made to semi-supervised learning algorithms. In the second part of the paper, a novel semi-supervised deep learning model for predicting the stress state is proposed. It is the first attempt of using contrastive learning for the stress prediction tasks. The model is based on utilizing generative and contrastive features specially tailored for treating time-series data. A widely popular multimodal WESAD (Wearable Stress and Affect Detection) data set is exploited for experimental purposes. It consists of physiological and motion data recorded from the wrist and chest-worn devices. To provide an intelligent solution that will be widely applicable, only the wrist data recorded from smartwatches is exploited during the model's training. The proposed model in this research is tested on a single subject's data and predicts the stress and non-stress events. Keeping in mind that the initial data was unbalanced with only 11% of the stress data, data augmentation techniques are applied within the model to provide additional reliable training information. The model shows significant potential in clustering stress conditions, and it presents accuracy in the range with other state-of-the-art solutions. The most significant benefits of using this model are its prediction capabilities when dealing with unlabeled data and performances when undersized data cannot be processed optimally by traditional intelligent methods.**

Keywords*—*Deep learning; semi-supervised learning; contrastive learning; physiological data; stress prediction

I. Introduction and Literature Review

The coronavirus pandemic has directly affected the health of millions of people and living conditions all over the planet. To reduce the impact of the infection, worldwide governments temporarily closed schools, malls, restaurants, sports and art facilities, and all other public and private institutions in which a large number of people are generally present. Like a domino effect, such situations have affected the loss of jobs of millions of people, almost the closure of some professions, the closure of small and large companies, and finally, significant disruptions in the world economic system. In addition to the negative effects that this situation is causing, so far, not much attention is paid to the health problem that will affect a much larger number of people than the virus itself did: stress. To contribute to the diagnosis of stress and the detection of stress in humans, this research addresses the development of an intelligent method for the early detection of stress and the timely prevention of more serious health problems.

Until now, stress and emotions, in general, have been examined in numerous researches. In [1] an emotion recognition system was presented that is applicable when limited data resources are available. The system identified emotions with a 65% success rate and with the confidence of 57%. Further, in [2] the system that discovers classroom emotions and moods of students was introduced. Wristband sensors from Empatica E4 and smart-phones were used to detect all emotions using physical activities, event tags data, and various physiological parameters. The results were exploited for finding associations and correlations between students' data and extracting meaningful insights. In another research [3], a deep learning approach for the classification of emotions was presented. This approach was based on processing data acquired from three sensor modalities: locations into the global model, environmental and on-body. It was also proven that deep learning algorithms can be very effective in classifying human emotions, especially when many sensors are utilized. The average accuracy of the proposed model was 73%. Next, the importance of emotions in user-modeling and multimodal computer interaction was presented in [4]. Additionally, in this paper, the results of different supervised learning algorithms for categorizing physiological signals from the autonomous nervous system were shown. The multimodal system for recognizing users' emotions and generating responses to recognized emotions was presented in [5]. The experiment is thoroughly explained, and the mapping principles of physiological signals to certain emotions were introduced. The utilization of positive and negative emotional electroencephalogram (EEG) signals was described in [6] to research emotions. The support vector machine algorithm was exploited for data analysis, and an accuracy of 58.3% was achieved. Besides using EEG signals, an important role in emotion recognition is the electrocardiogram (ECG) signal. In [7] wearable ECG device was used to follow four kinds of emotional states recorded while involved participants watched prepared movie clips. Features from different analysis domains (time, frequency, static analysis) were sensed and recorded, and the most relevant features for evaluating human's emotions were highlighted. An interesting approach to establishing emotional connections between SMART TVs and the audience was presented in [8]. EEG signals of involving participants were recorded and analyzed, and three emotional states were registered: relaxation, neutral, and horror. These signals were classified by using the support vector machine with an accuracy of 92% for all subjects.

Keeping in mind that the paper's novel research is based on using smartwatch data to detect emotions, it is beneficial to highlight the paper of [9]. The paper examined if smartwatches or wrist bands can be useful in collecting valuable information to recognize emotions. The primary limitations, potential problems, and crucial research steps in this domain were successfully found. Further, in [10], an automatic emotion recognition system was proposed. The system is based on using a wearable wristband, human emotions were evoked artificially, and multimodal physiological signals were collected by exploiting three different sensors. Finally, support vector machines were used once again for classifying emotions, achieving an accuracy of 76%. The usage of wearable sensors was also verified in [11], where they were used within Ambient Intelligence systems to provide affect-based adaptations. Wearable sensors were also utilized in [12], where a real-time mobile biosystem "iAware" was proposed. The system was efficient in depicting five basic emotional states and emotional feedback information was provided to users. A smartwatch application that integrates heart rate, motion, and light data to sense mental health was used in [13], and the PRISM-Passive platform was proposed. Both supervised and unsupervised learning algorithms were applied, and it is proven that smartwatch data could be useful in evaluating and predicting mental health. Finally, the usage of the pervasive wearable devices within the "emotional IoT" concept for recognizing emotions was presented in [14]. This paper presented the end-to-end real-time solution based on smartwatch and smartphone devices that showed great applicability potential in consumers' everyday lives.

In the following papers, the research cover techniques to estimate physiological signals, their processing, and signal quality improvements. In [15] psychophysiological signal quality estimators were proposed that were utilized to affect recognition systems. Further, in [16] findings in the domain of estimations of affective states of users' optimal experiences were presented. Estimated signals through end-to-end intelligent architecture possessed 67.5% accuracy in recognizing different affective states, including stress. The difference in emotion recognition accuracy between laboratory and wearable sensors was examined in [17]. The results showed a similar level of accuracy between the two approaches, which implies that wearable sensors' usage is reliable enough for serious considerations and accurate collection of physiological information of a user outside of a laboratory. Another research [18] covered a framework for signal processing pertaining to clarifying patterns of humans' physiological changes. An urban environment was taken as a scientific background, and the framework included signal unification, filtering, quantification, and the usage of techniques for data labeling. Finally, one interesting research of transformation of emotional signals is presented in [19], where the biosensing prototype for transforming emotions into music was proposed. Four emotional states were covered within the research (neutral, anger, sadness, and happiness). The appropriate EEG signals were recorded, and Audiolize Emotion was used to transform collected data into audio files. In the cases when it is difficult to assign labels to training input data consistently, Multiple Instance Learning from [20] allowed the training of classifiers from not precisely defined labeled data. A potential guideline for increasing the accuracy of labels was presented in [21]. However, in the cases when the labels cannot be completely provided, a self-supervised approach from [22] can be applied. The approach was designed in a way to learn valuable representations from unlabeled sensor inputs as blood volume pulse, electroencephalography, accelerometer, etc. The proposed methodology showed performances in the range with fully-supervised networks and improved generalization capabilities in semi-supervised settings.

As one more interesting topic for proposing the novel research in this paper, deep learning techniques for processing physiological and emotional parameters should also be presented. In [23] deep learning techniques for real-time stress and affect detection was examined. New models based on Multiple Instance Learning were proposed and applied, showing the performances 10% better in terms of accuracy. Further, the hyperparameter optimization framework of long short-term memory networks in the context of emotion classification was presented in [24]. It was shown that the framework provided an improved recognition rate accuracy of more than 10% compared to other state-of-the-art optimization methods. One more classification model built with deep neural networks was presented in [25]. A fully convolutional network was proposed and achieved performances were in the range or better than other state-of-the-art time series classification algorithms. A convolutional neural network architecture was also used in [26] to classify the biosignals, achieving the precision across all the classes equal to 97.65%. Finally, in [27, 28] novel techniques for optimizing network architectures to improve their processing power were presented. Specifically, in [27] rectifying neurons as improved models of biological neurons were presented. The structures based on these neurons are suitable for sparse data, they do not require unsupervised pre-trainings, and deep rectifier networks were very efficient in environments where there was a lack of labeled data. Additionally, classification models can be improved by making a normalization process an integral part of a model architecture [28]. This method performed the normalization process for each training batch, provided the same accuracy with 10-15 times fewer training iterations than some traditional network structures. Two more successful approaches of utilizing deep learning techniques for prediction purposes were given in [29, 30].

In this novel research, a novel self-supervised learning (SSL) algorithm is proposed and utilized for processing a widely popular WESAD data set. The goal is to accurately detect stress by using smartwatch sensor data. To describe the research methodology and proposed framework, the mentioned dataset must first be introduced properly. In general, WESAD is a multimodal dataset for wearable stress and affect detection [31]. The set includes information recorded both from chest and wrist measurements of sensors. It was based on three different affective states: neutral, amusement, and stress. Fifteen subjects participated in the experiment, 12 males and three females. The average age of participants was 27.5. The stress condition occurred as a response from public speaking exercises, and no other types of stress environments were included. Linear Discriminant Analysis (LDA) model was used

in [31] as a stress classification model, which achieved 93% accuracy. The next important paper is [32], in which the WESAD data were classified into four classes: neutral, amusement, and stress as in the previous paper, and meditation was an additional class. In comparison to [31], [32] only used the wrist sensor data for classification purposes. A machine learning model was trained for each subject separately (logistic regression, decision tree, and random forest models). The best performances were achieved using random forest models: accuracy between 88% and 99% depending on the examined subject. Whether it was possible to perform stress detection using only a smartwatch sensor data from the WESAD data set was examined in [33]. Three different models were used: LDA, Quadratic Discriminant Analysis (QDA), and Random Forest (RF). The best performances were achieved with LDA in combination with the next sensors' data: heart rate (HR), blood volume pulse (BVP), and skin temperature (ST). The next paper [34] relied on using deep learning techniques for processing the WESAD data. The primary model processed inputs of different sampling rates by utilizing four different sub-models as classifiers that individually process per one different sampling rate. The final model was based on the RF algorithm and generated the final classifications following the fusion mechanism in [35]. Finally, the research in [36] used a self-supervised methodology and deep learning for processing only the ECG signal from four data sets (including WESAD). The methodology was developed using data augmentation techniques, including stacked convolutional network layers and a final "SoftMax" dense classification layer. By examining previously presented WESAD research papers, a major issue that significantly influences the dataset's classification results was identified: a lack of data diversity. An attempt to solve this problem will be made in the research by introducing a novel SSL methodology. In the next section, the research background is presented, and fundamental SSL concepts are introduced.

II. Research Background

For this research application purposes, the future model's output labels are two physiological states: "stress" and "neutral". The labels represent the outputs of a model for corresponding input data points. To determine when these two labels occur, it is required to know whether a subject of examination is stressed or not stressed. This can only be accurately determined in an experimental environment by making long-term observations by expert knowledge from the field. If a supervised learning algorithm is selected for processing the data, it is required to provide output labels for all the input data. For the situations when complex and time-consuming experimental procedures should be performed (as in this case of the stress prediction), much effort should be made and costs covered to collect the complete database of labels. One possible way to optimize this process is to use an SSL approach for the training process and reduce the need for the number of prepared labels. Unlike supervised learning that relies on using pre-prepared input/output pairs of data, SSL techniques create their output labels from available input information. These techniques reduce the dependency on labels by constructing meaningful and invariant representations that capture the original data's high-level information. SSL techniques are based on two essential concepts: the pretext task and making appropriate representations. The key to the pretext task is to use information about the input data to construct pretext labels. On the other hand, a representation is a way to simplify the data while keeping relevant information. For sensor data that is processed within this research, a useful representation preserves emotional and physical state information and discards redundant information like noise. Additionally, the previous SSL applications show that representations with significant invariance are often more robust and provide better quality than other types of representations.

A review of SSL methods for treating images, text files, and graph data was presented in [37]. The comparisons between supervised and unsupervised learning algorithms were made, and three main categories of SSL were explained: Generative, Contrastive, and Adversarial SSL. One efficient approach of contrastive learning in the form of a simple framework called SimCLR was proposed in [38]. The approach was based on utilizing the data augmentation techniques and using two different networks within the architecture: the first one to create representations of different inputs and the second one for comparison of representations and preserving important information. Examined research showed that data augmentation could be a key tool to build accurate SSL models. Augmentation techniques apply input data transformations to create new relevant samples from the existing ones and increase the size of data sets when needed. One example of using data augmentation for creating the models with significant accuracy was presented in [39]. In this paper, a popular contrastive method called Dimensionality Reduction by Learning an Invariant Mapping (DrLIM) was proposed. The method is efficient in creating relevant features out of complex data, and it was commonly applied in previous years for solving a variety of practical problems. Further, the effectiveness of data augmentation techniques can be observed within [36], where six different augmentation techniques were applied, and the accuracy of the SSL algorithm on the WESAD data was more than 95%.

In this research, two related SSL approaches within a novel algorithm will be used: the first one based on temporal classification and the second one on contrastive learning. It can be considered that the temporal classification is a pre-step toward contrastive learning. Generally speaking, temporal classification [40] is an SSL methodology specially tailored for time series data. The methodology's basic principle is that the data features vary slowly compared to the sampling time of recorded measurements. The methodology does not require labels because the classification is performed using time intervals (seconds, minutes) generated via a random data preparation procedure. Another advantage of temporal classification is that it transforms complex unsupervised features into simpler classifying segments.

This scientific approach represents one of the first implementation attempts of contrastive learning to stress classification. Contrastive learning is an SSL technique that provides a model's training without requirements for output labels [41], and it is based on learning a similarity metric between data samples. Using a similar technique, SimCLR from [38] reached state-of-the-art accuracy and even matched

some supervised models' performances. In [42], the representation learning method with contrastive predictive coding that applies to different data modalities was presented. The method's predictive coding component relies on training the model to predict the representation in time instance t+1 by using the history of the specific representation until time t. In other words, the model must understand what activity the subject is currently doing to generate future predictions accurately. The efficient methodology from [42] was successfully used to improve a speech recognition algorithm based on SSL representations in [43]. The methodology was utilized on high-frequency sensor data and represented an excellent example for the novel research in this work from the perspective of the existence of similar research environments in both cases. The research from [43] was based on using an encoder network that produces representations that are further mixed by the context network to create a context vector. The vector was finally used to predict the next representation. Another way of solving speech recognition tasks was presented in [44] when an unsupervised learning algorithm was used. It is another proof that these complex kinds of tasks can be successfully solved without supervised learning algorithms. Significant results within [43] and other related papers represented the motivation for the authors of this new research to implement contrastive learning on the WESAD data. However, besides all the advantages of contrastive learning that have been proven experimentally through introduced papers, its implementation remains a challenge because of the potential difficulties with creating the pairs, evaluating the model performance, and implementing and evaluating a proper loss function [45]. In the following sections of this paper, the research attempts to prevent all these difficulties and proposes a novel intelligent SSL solution.

III. Data Preparation and Research Methodology

The initial step of almost every intelligent approach is data exploration and pre-processing. As previously explained, in this research, the WESAD data set was exploited. It was created and maintained by the University of California Irvine and stored within their open-source machine learning repository. WESAD is the multimodal dataset that consists of motion and physiological data recorded from the chest and wrist-worn devices. For this novel research, only the wrist data coming from smartwatch sensors were used. Complete data were collected by recording vital parameters of 15 involved subjects during the study (labeled with S1 to S15, accordingly).

Besides measured parameters, three affective states were also registered during the experiment: neutral, stress, and amusement. The primary deficiencies of the WESAD dataset that should be mentioned are the lack of examined subjects (only 15 participated in the experiment) and a single type of evaluated stress activity. Collecting new stress labels would be expensive from the perspectives of the required time for new laboratory experiments, and the costs of assigning new participants required the recreation of the WESAD experiment and increase the database. Application of SSL techniques can solve this problem and give optimal performances from already available data and provide maximum possible accuracy in stress prediction. Besides described approaches of classifying emotions by recording physiological parameters, an interesting language-independent acoustic emotion classification was presented in [46].

Initial data preparation work in this novel research was based on putting all the sensors on the same timeline (700Hz) and merging all the subject data. The data set was split into train, validation, and test set by following standard machine learning procedures. Further, the data exploration process was performed on the training data composed of subjects S2 to S15. Within 40 million rows within the training data, only 11% (around 4 hours) was collected from the stress state. Such a small percentage of the stress information made this set imbalanced and presented a real challenge to make an accurate system for detection and prediction of it. One possible approach for treating imbalanced data was given in [47], where an intelligent algorithm based on utilizing a genetic algorithm was proposed. Finally, 3% of data was considered invalid.

Keeping in mind the main research task to detect and predict a subject's stress status, it was essential to determine which sensors were the most correlated with stress. After performing a basic correlation analysis, it was concluded that acceleration and electrodermal activities were the most correlated with stress. The next effort was made in data pre-processing, where the outliers were removed and the signals denoised. For the outlier removal process, each sensor was assigned to a valid range of values. The values outside of the specified ranges were deleted and replaced by the closest valid values. Finally, for dealing with the noise components, a low-pass filter was utilized to remove undesirable frequencies. For each sensor value, a cutoff frequency (the highest frequency that is meaningful for a specific sensor) was specified. A Butterworth low-pass filter of the second order was then used with the corresponding cutoff to process the signal.

After a brief introduction of the WESAD data and finishing basic pre-processing tasks, the next research phase was to utilize a novel SSL methodology to a small subset of the data and progressively increased the size of exploited information and the complexity of processing. Considering that WESAD was a representative of time series data, in [48] how an intelligent approach was utilized for treating and classifying time series data was researched. Following the progressive experimental approach, the research task was to train the model on a single subject of data, which corresponded to 1 hour and 30-minute readings from sensors. The features used for processing purposes were wrist acceleration, blood volume pulse, and wrist temperature measurements.

SSL techniques were already successfully applied to processing the WESAD data in [22, 36]. However, the novelty of the new research that will make the novel approach unique was that it was the first attempt to use contrastive learning for the stress prediction tasks. Further, the approach was wholly based on using commercial smartwatch data and making it available for a broad audience. In this research, two previously established SSL methods were implemented: the temporal classification approach from [40] and the contrastive learning from [38]. As in [40], it was important to mention that the features of interest for this novel research also varied slowly (every few minutes) compared to the sampling rates of including sensors (a few milliseconds). The research data was

split into one-minute segments, and the model was trained to classify these segments by their natural belongings. The model was based on three complementary techniques: contrastive learning, the utilization of slow-moving features, and data augmentation techniques. These techniques and the overall algorithm of the model are presented in Fig. 1.

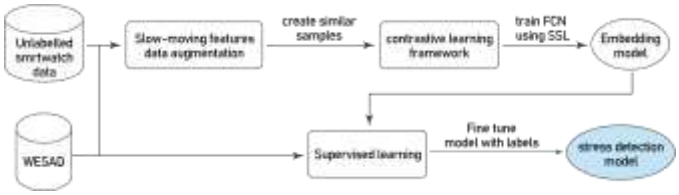


Fig. 1. The Algorithm of the New Model.

IV. Model Design

To set up the temporal classification part of the algorithm, the initial process was divided into three steps: generating one-minute segments, associating a corresponding label to each window, and training the deep learning model to predict the segment from offered windows. A similar methodology was applied in [43], where segments of 12,5 seconds were formed, and a linear classifier was used on top of the representations to predict which stimuli were experienced by a subject. One of the methodology's observed drawbacks was the requirement of creating the segments strictly before their usage within machine learning and deep learning models. It was not possible to add new data later if it was generated during an online learning process. Another deficiency was in terms of data size restrictions and the applied number of segments to prevent losing the speed of a model. The number of segments matched the output of the final layer of the model, implying that the number of segments was directly proportional to the model's size.

To evaluate the initial setup from Fig. 1, a dataset consisting of a single subject's data was processed. A simple neural network with a single hidden layer and 20 neurons were utilized as the supervised learning model (Fig. 2). The network was based on SoftMax activation functions, and its purpose was to classify the labeled data. Finally, the model was trained using stochastic gradient descent with the Adam optimizer, while the cross-entropy was used as the loss function.

Next, Fully Convolutional Network from [25] is implemented as the embedding model from Fig. 1. The network algorithm is adjusted to this specific research case. Three layers within the network are proposed, and the graphical representation of its structure is presented in Fig. 3.

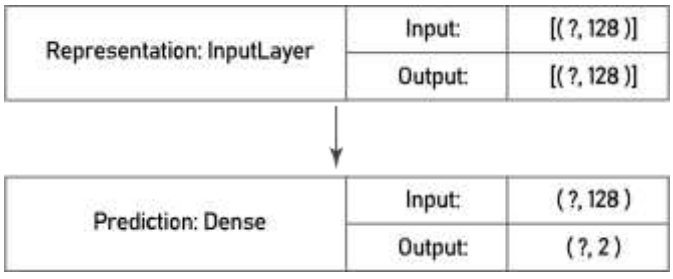


Fig. 2. The Supervised Learning Model Algorithm.

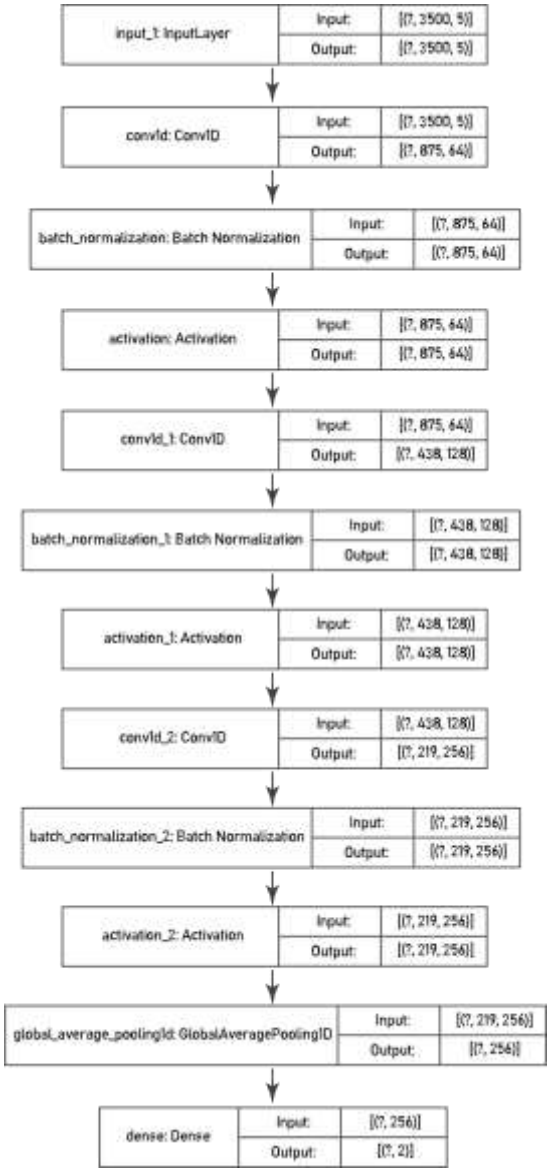


Fig. 3. Fully Convolutional Network (FCN) Architecture.

The contrastive method's key was to create a local classification objective that does not depend on the whole dataset. The task is to classify pairs as similar and different following their temporal closeness. To create similar pairs, the following data preparation algorithm was used. First, the data was split into 5 seconds windows of five 700 Hz sensors. The splitting phase resulted in 17500 elements for each input time point. A pair was created for each window by associating it randomly with another window that occurred less than a minute before. Each such pair was called a positive pair because it was considered temporally close enough to be similar. Further, ten pairs were merged to form a batch of 20 pairs of windows. If two windows belonged to the same positive pair, they were considered similar; otherwise, they were dissimilar. Initiated labels were local because they were created within each batch of data separately.

Once the labels were created, a loss function from [38] was used to estimate how well the model classified the pairs.

TensorFlow 2, as one of the most famous Python libraries, was used for this purpose. The loss function called the Noise Contrastive Estimation loss (NCE loss), and was presented in the following form:

$$L_i = -\log \frac{\exp(z_l(\mathrm{i}) z_r(\mathrm{i})^T)}{\sum_{i \neq j} \exp(z_l(\mathrm{i}) z_r(\mathrm{i})^T)} \quad (1)$$

Where z_l and z_r are the representations obtained from the model of the left and right samples of each pair. To accurately measure the loss, it was essential to evaluate the loss on small subsets of the data individually. When each small sample was selected, the binary cross-entropy was computed using only elements from this subset. Because of the small size of any subset, the cross-entropy computation was fast. To learn from the whole dataset, all subsets were processed one by one. Finally, to get the final loss, the losses from each pair were averaged. The model was trained to utilize this loss function within the gradient descent algorithm on the pairs of previously created batches. Additionally, Adam optimizer was used to update the weights at each pass.

V. Results

The goal of this case study was to recognize stress and neutral states from the input representation data on a single subject's data. The model was trained to learn two-dimensional representations to validate the ability to learn relevant features. The training time of a single model on the machine with Nvidia K80 GPU was 10 minutes approximately. The features that were generated after the training process are presented in Fig. 4. The train set (left) and the test set (right) for Subject 15 is shown with the separation between the different reported activities. For the test set for Subject 15 the stress and meditation results are shown only because the data is split by time into train and test where the first 70% of this subject data goes to train and the rest goes to test. For this subject, the final 30% of data only have these two activities which is a very interesting result considering only a very limited amount of data that been used for this training. Therefore, by using unlabeled data, the model is efficiently capable of clustering different activities as shown.

It is observable from the previous figures that the model efficiently learned to cluster different activities. It learned to separate activities without knowing them in advance, which is the temporal classification paradigm's success. If Fig. 4 is presented from the perspective of only stress and no stress activities, Fig. 5 can be generated as well. It can be concluded from the figure that the model is efficient in separating stress from no stress data. It showed 85% accuracy demonstrating that generated representations can capture most of the relevant information in some simpler cases.

In the final part of this research, the performances of the proposed model were compared to five similar researches by other authors. Table I present information about the advantages and limitations of developed models, as well as exploited types of data during the training processes. Finally, the results and achieved accuracies are shown as the evaluation measures for all the models. Final examination showed that the proposed model provides the accuracy in the range with other state-of-the-art solutions, and an advantage in the term of processing unlabeled data and augmenting existing data. Through this conducted research and the previous one cited within this paper, it can be concluded that the contrastive methods could be very efficient in processing large data sets. It is even possible to parallelize the computations to train such sets, to provide online training, and add additional data within the training process. All these topics will be examined in detail in the future work of the authors.

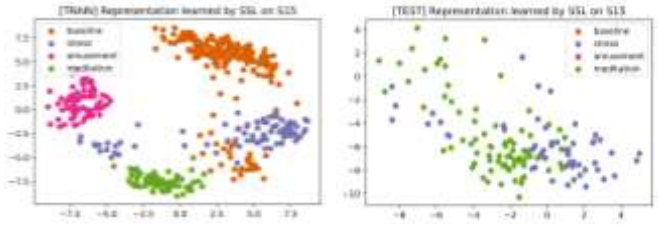


Fig. 4. Data Representations Learned by Contrastive Learning by Examining Subject 15.

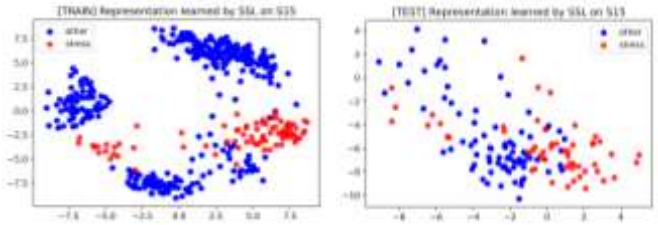


Fig. 5. Representations of Stress and no Stress Features of S15.

TABLE I. Comparison of Six different Approaches for Stress Detection

Reference number	Method	Advantages	Limitations	Exploited data	Results, accuracy
[31]	Feature-based ML approach	Speed, interpretability, low compute power	Required expert knowledge of the features	All the sensors	93% (stress detection)
[32]	ML approach, per subject	Same as [1] + tailored to a single subject	Same as [1] + need to be retrained for each subject	All the sensors	88-99% (per subject)
[33]	ML approach	Same as [1] + applicable to commercial smartwatches	Same as [1]	Smartwatch compatible data: wrist only without EDA	87% (stress detection, balanced accuracy)
[34]	DL approach	High-accuracy, no expert knowledge	High compute, complex model	All sensors	85% (3 classes)
[36]	DL, self-supervised learning	Very high accuracy, no expert knowledge	High compute, complex method	ECG only + additional data sets	97% (4 classes)
New research	DL, self-supervised learning	Can leverage unlabeled data	More complex than supervised learning	WESAD smartwatch data (all wrist data without EDA as in [3])	85% (stress detection)

VI. Conclusion

In this paper, the SSL concept of deep learning was presented and analyzed, and a novel SSL solution was proposed. As a popular case study nowadays, emotional states and stress detection were selected as test cases for this research. In the first section of the paper, a review of popular scientific papers dealing with intelligent techniques for processing emotions was made. It was proven that deep learning and machine learning approaches can threaten emotion data effectively and produce desirable results in the form of a prediction, label detection, classification, or clusterization. This paper's contribution was the utilization of only wrist sensor data (from smartwatches) in the processing phase, without the requirement for any additional data that should be collected by using any intrusive method. State-of-the-art research papers concerning smartwatch sensor data applications were also provided, highlighting the smart approaches for treating such data. Special attention was made to deep learning techniques in the field of emotion recognition and solving similar tasks. It was shown that different supervised and unsupervised learning techniques could be effectively applied for processing physiological data and providing valuable insights. Finally, the WESAD data set, as a base for the case study in this paper, was presented, and the most important research papers were introduced and described. In the second section, a literature review concerning SSL techniques was provided, and the main features were exposed. Special attention was made to the introduction of generative and contrastive SSL algorithms, as they represented the basis for the future model. Additionally, temporal classification as a pre-step toward contrastive learning was also highlighted, as it was an efficient methodology specially tailored for time series data, as was the WESAD data set. In the third section, the nature and features of the data and applied pre-processing techniques were described. All outliers from the data were removed, and important correlations between the features were discovered. A small subset of optimized data (measurement information for a single subject) was finally used to train and test the proposed model in section IV. The model was based on using slow-moving features and data augmentation techniques to increase available data and create similar samples. Then, the contrastive learning framework was used to train the developed network by using the SSL approach. Finally, the embedding model outputs were used within the supervised learning algorithm to provide fine-tuning of the proposed stress detection model.

The novelty of the proposed solution was in utilizing pairs of samples instead of state-of-the-art models that processed a single sample at a time. The SSL algorithm's potential was also shown by the developed ability to cluster human activities without knowing their specifics. The model was able to recognize features very efficiently and abstract concepts, such as meditation, stress, and amusement using only the raw sensor data. Finally, generated representations of the model were evaluated in two ways: the first one using the accuracy metric on pairs of batches and the second one utilizing a shallow supervised model on the top of the representations. In the next steps and future research work in this domain, new functionalities will be added and the updated algorithm utilized on the complete WESAD data set. Novel research should answer if SSL models are capable of processing a large quantity of data and providing accurate stress predictions when multiple subjects are treated at once.

References

[1] Pollreisz, D., and Taheri, N. (2017) A simple algorithm for emotion recognition, using physiological signals of a smart watch. 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju Island, Korea, 11-15 July, 2353– 2356, IEEE, New York, USA.

[2] Dao, M., Nguyen, D., Tien, D., and Kasemet, A. (2018) Healthyclassroom-a proof-of-concept study for discovering students' daily moods and classroom emotions to enhance a learning-teaching process using heterogeneous sensors. 7th International Conference on Pattern Recognition Applications and Methods, 16 - 18 Jan, 2018, Funchal, Madeira, Portugal. ISBN 978-989-758-276-9.

[3] Kanjo, E., Younis, E., and Ang, C. (2019) Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. Information Fusion. 2019, 49, 46–56.

[4] Lisetti, C., and Nasoz, F. (2006) Categorizing autonomic nervous system (ANS) emotional signals using bio-sensors for HRI within the MAUI paradigm. 15th IEEE International Symposium on Robot and Human Interactive Communication. 20116, 277–284.

[5] Lisetti, C., and Nasoz, F. (2004) Using non-invasive wearable computers to recognize human emotions from physiological signals. EURASIP journal on applied signal processing. 2004, 1672–1687.

[6] Nie, Y., Wu, Y., Yang, Z., Sun, G., Yang, Y., and Hong, X. (2017) Emotional evaluation based on svm. 2nd International Conference on Automation, Mechanical Control and Computational Engineering.

[7] Guo, H., Huang, Y., Chien, J., and Shieh, J. (2015) Short-term analysis of heart rate variability for emotion recognition via a wearable ecg device. International Conference on Intelligent Informatics and Biomedical Sciences. 262–265.

[8] Jalilifard, A., da Silva, A.G., and Islam, K. (2017) Brain-tv connection: Toward establishing emotional connection with smart tvs. IEEE Region 10 Humanitarian Technology Conference. 2017, 726–729.

[9] Saganowski, S., Dutkowiak, A., Dziadek, A., Dziezyc, M., Komoszyńska, J., Michalska, W., Polak, A., Ujma, M., and Kazienko, P. (2019) Emotion recognition using wearables: A systematic literature review - Work in progress. IEEE International Conference on Pervasive Computing and Communications Workshops. 2019, 1-6.

[10] Zhao, B., Wang, Z., Yu, Z., and Guo, B. (2018) Emotion sense: Emotion recognition based on wearable wristband. IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation. 2018, 346–355.

[11] Nalepa, G., Kutt, K., Gizycka, B., Jemioło, P., and Bobek, S. (2019) Analysis and use of the emotional context with wearable devices for games and intelligent assistants. Sensors. 2019, 1-24.

[12] Albraikan, A., Hafidh, B., and El Saddik, A. (2018) iaware: A real-time emotional biofeedback system based on physiological signals. IEEE Access. 2018, 6, 78 780–78 789.

[13] Kamdar, M.R., and Wu, M.J. (2016) Prism: a data-driven platform for monitoring mental health. Biocomputing 2016: The Pacific Symposium.- World Scientific. 2016, 333–344.

[14] Setiawan, F., Khowaja, S.A., Prabono, A.G., Yahya, B.N., and Lee, S. (2018) A framework for real time emotion recognition based on human ansusing pervasive device. IEEE 42nd Annual Computer Software and Applications Conference. 2018, 1, 805–806.

[15] Gupta, R., Khomami Abadi, M., Cárdenes Cabré, J.A., Morreale, F., Falk, T.H., and Sebe, N. (2016) A quality adaptive multimodal affect recognition system for user-centric multimedia indexing. ACM 2016 on international conference on multimedia retrieval. 2016, 317–320.

[16] Maier, M., Marouane, C., and Elsner, D. (2019) Deepflow: Detecting optimal user experience from physiological data using deep neural networks. 18th International Conference on Autonomous Agents and Multi Agent Systems. International Foundation for Autonomous Agents and Multiagent Systems. 2019, 2108–2110.

[17] Ragot, M., Martin, N., Em, S., Pallamin, N., and Diverrez, J.M. (2017) Emotion recognition using physiological signals: laboratory vs. wearable sensors. International Conference on Applied Human Factors and Ergonomics.Springer. 2017, 15–22.

[18] Ojha, V., Griego, D., Kuliga, S., Bielik, M., Buš, P., Schaeben, C., Treyer, L., Standfest, M., Schneider, S., Koenig, R., Donath, D., and Schmitt, G. (2019) Machine Learning Approaches to Understand the Influence of Urban Environments on Human's Physiological Response. Information Sciences. 474, 154–169.

[19] Lu, X., Liu, X., and Bergqvist, E. (2019) It sounds like she is sad: Introducing a biosensing prototype that transforms emotions into real-time music and facilitates social interaction. CHI Conference on Human Factors in Computing Systems. 2019, 1-6.

[20] Babenko, B. (2008) Multiple Instance Learning: Algorithms and Applications. Technical Report, San Diego, USA.

[21] Schmidt, P., Reiss, A., Dürichen, R., and Van Laerhoven, K. (2018) Labelling affective states in the wild: Practical guidelines and lessons learned. ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. 2018, 654–659.

[22] Saeed, A., Salim, F., Ozcelebi, T., and Lukkien, J. (2020) Federated Self-Supervised Learning of Multi-Sensor Representations for Embedded Intelligence. IEEE Internet of Things Journal. Early access, 2020, 1-11.

[23] Ragav, A. (2019) Scalable Deep Learning for Stress and Affect Detection on Resource-Constrained Devices. 18th IEEE International Conference On Machine Learning And Applications. 2019, 1585–1592.

[24] Nakisa, B., Rastgoo, M.N., Rakotonirainy, A., Maire, F., and Chandran, V. (2018) Long short term memory hyperparameter optimization for a neural network based emotion recognition framework. IEEE Access. 6, 49 325–49 338.

[25] Wang, Z., Yan, W., and Oates, T. (2017) Time series classification from scratch with deep neural networks: A strong baseline. International Joint Conference on Neural Networks. 2017, 1578-1585.

[26] Chakraborty, S. (2019) A Multichannel Convolutional Neural Network Architecture for the Detection of the State of Mind Using Physiological Signals from Wearable Devices. Journal of Healthcare Engineering. 2019, 1–17.

[27] Glorot, X., Bordes, A., and Bengio, Y. (2010) Deep Sparse Rectifier Neural Networks. Journal of Machine Learning Research. 15, 315-323.

[28] Ioffe, S., and Szegedy, C. (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 32nd International Conference on Machine Learning. 37, 448-456.

[29] Taplak, H., Erkaya, S., Yildirim, Ş. (2014) The Use of Neural Network Predictors for Analyzing the Elevator Vibrations. Arabian Journal for Science and Engineering. 39, 1157–1170.

[30] Elkatatny, S. (2019) A Self-Adaptive Artificial Neural Network Technique to Predict Total Organic Carbon (TOC) Based on Well Logs. Arabian Journal for Science and Engineering. 44, 6127–6137.

[31] Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., and Van Laerhoven, K. (2018) Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. 20th ACM International Conference on Multimodal Interaction. 400-408.

[32] Indikawati, F., and Winiari, S. (2020) Stress Detection from Multimodal Wearable Sensor Data. IOP Conference Series: Materials Science and Engineering. 1-6.

[33] Siirtola, P. (2019) Continuous stress detection using the sensors of commercial smartwatch. ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 ACM International Symposium. 1198-1201.

[34] Lin, J., Pan, S., Lee, C., and Oviatt, S. (2019) An Explainable Deep Fusion Network for Affect Recognition Using Physiological Signals. 28th ACM International Conference on Information and Knowledge Management. 2069-2072.

[35] Ruta, D., and Gabrys, B. (2000) An Overview of Classifier Fusion Methods. Computing and Information Systems. 7, 1-10.

[36] Sarkar, P., Etemad, A. (2020) Self-supervised ECG Representation Learning for Emotion Recognition. IEEE Transactions on Affective Computing. Early access, 1-13.

[37] Liu, X., Zhang, F., Hou, Z., Wang, Z., Mian, L., Zhang, J., and Tang J. (2006) Self-supervised Learning: Generative or Contrastive. arXiv:2006.08218. 1-23

[38] Chen, T. (2020) A Simple Framework for Contrastive Learning of Visual Representations. 37th International Conference on Machine Learning. 119, 1597-1607.

[39] Hadsell, R., Chopra, S., and Lecun, Y. (2006) Dimensionality Reduction by Learning an Invariant Mapping. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1735-1742.

[40] Hyvarinen, A., and Morioka, H. (2016) Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. 30th International Conference on Neural Information Processing Systems. 3772–3780.

[41] Kreuk, F., Keshet, J., and Adi, Y. (2010) Self-Supervised Contrastive Learning for Unsupervised Phoneme Segmentation. Electrical Engineering and Systems Science : Audio and Speech Processing. Early access, 1-5.

[42] Oord, A., Li, Y., and Vinyals, O. (2018) Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv:1807.03748.

[43] Gutmann, M., and Hyvärinen, A. (2010) Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. Journal of Machine Learning Research - Proceedings Track. 9, 297-304.

[44] Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019) wav2vec: Unsupervised Pre-Training for Speech Recognition. Interspeech 2019, 3465-3469.

[45] Alshamrani, M. (2021) IoT and artificial intelligence implementations for remote healthcare monitoring systems: A survey, Journal of King Saud University - Computer and Information Sciences, 2021, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2021.06.005.

[46] Singh, R., Puri, H., and Aggarwal, N. (2020) An Efficient Language-Independent Acoustic Emotion Classification System. Arabian Journal for Science and Engineering. 45, 3111–3121.

[47] Jiang, K., Lu, J., and Xia, K. (2016) A Novel Algorithm for Imbalance Data Classification Based on Genetic Algorithm Improved SMOTE. Arabian Journal for Science and Engineering. 41, 3255–3266.

[48] Dempster, A., Petitjean, F., and Webb, G. (2020) ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. Data Mining and Knowledge Discovery. 1-42.

Customer Segmentation and Profiling for Life Insurance using K-Modes Clustering and Decision Tree Classifier

Shuzlina Abdul-Rahman, Nurin Faiqah Kamal Arifin, Mastura Hanafiah, Sofianita Mutalib
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam, Malaysia

***Abstract*—Customer segmentation and profiling has become an important marketing strategy in most businesses as a preparation for better customer services as well as enhancing customer relationship management. This study presents the segmentation and classification technique for insurance industry via data mining approaches: K-Modes Clustering and Decision Tree Classifier. Data from an insurance company were gathered. Decision Tree Algorithm was applied for customer profile classification comparing two methods which are Entropy and Gini. K-Modes Clustering segmentized the customers into three prominent groups which are "Potential High-Value Customers", "Low Value Customers" and "Disinterested Customers". Decision Tree with Gini model with 10-fold cross validation was found as the best fit model with average accuracy of 81.30%. This segmentation would help marketing team of insurance company to strategize their marketing plans based on different group of customers by formulating different approaches to maximize customer values. Customers can receive customization of insurance plans which satisfy their necessity as well as better assistance or services from insurance companies.**

***Keywords*—Customer segmentation; customer profiling; decision tree; insurance domain; k-modes clustering**

I. Introduction

Insurance industry has been in the global market for decades and it is a critical contributor to a country's long term economic growth. Life insurers improve their policyholders' quality of life by pooling the risk of mortality, morbidity, and longevity among a wide number of people and returning the benefits of this pooling in the form of guaranteed payments [1]. In insurance industry, maintaining current customers is a challenge. Customer retention is more important than acquisition of new customers. It is said that 20% of the customers contribute more to the revenue of the company than the rest, as according to Pareto principle [2]. Despite the belief that clients are important for insurance organizations in gaining income and enhance their profitability, acquiring and retaining clients are serious issues faced by insurance firms [3]. It is not easy to obtain and influence new clients because when compared to the current clients, generally, new clients purchase 10% fewer than them, fewer involvement in the purchasing procedure as well as association with the seller [4]. Additionally, acquisition of new clients is more expensive compared to the maintenance of existing clients of the company [5]–[7]. Besides that, the likelihood of effectively selling a good or service to existing active clients is approximately 60-70 percent, while the likelihood is just 5-20 percent for potential clients, which made a greater likelihood of success in selling a good or service to existing clients compared to the potential ones [8]. It is also worthy to note that different clients contribute different amount of revenue to insurance companies, and so it is vital to handle clients based on their profitability due to uneven revenue generated by them [9].

Insurance companies are growing in numbers and the diversity of services offered, in which the clients have full control of their decisions [7]. It is thus important to have a good customer relationship management to retain the existing customers. To achieve that, insurance companies need to identify their target markets by segmenting the customers into groups. This allows them to choose whichever services that match their needs from any service providers. Customer segmentation helps business people to customize marketing plans, identify trends, plan product development, advertising campaigns and deliver relevant products, as well as personalizing messages of individuals for better communication with the intended groups [10]. Consumer sectioning is a great instrument in separating the consumers into various groups and perform analysis on their traits [3], and thus organizations are able to focus on clients in distinct features and determine the most valuable clients by sectioning the clients [9]. Clustering methods have been employed in many studies to segmentize customers [3], [9], [11]–[14], while classification via Decision Tree has also been widely used in past studies [15]–[17].

The following are the contributions of this paper:

- This research uses K-Modes Clustering and Decision Tree Classifier for customer segmentation and profiling for insurance domain.
- Marketing team of insurance company will be able to strategize their marketing based on different group of customers by formulating different strategies to maximize customer values.
- Customers can receive customization of insurance plans which satisfy their necessity as well as better assistance or services from insurance companies.

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Selangor, Malaysia.

The remaining of this paper is structured as follows: Section II discusses the related works on clustering and classification methods, while Section III describes the study's methodology. Section IV highlights the results and Section V provides the discussion and finally Section VI concludes the paper with future works.

II. Literature Review

A. Data Mining and Machine Learning

Investigation of unseen data and recognition of designs as well as affiliations that have valuable usages can be performed by information mining methods [18]. Organizations are able to pull out beneficial information from the data and obtain comprehension of their clients as well as their necessity through this information by implementing data-mining methods [7]. Data mining which is also part of knowledge discovery in database (KDD) involves the following process [19]: data selection, pre-processing, transformation, performing data mining algorithm, and data interpretation and evaluation. Data mining techniques like regression, classification, clustering, forecasting, association and visualization are also part of the classification framework in customer relationship management (CRM) [20].

While data mining extracting information from the vast amount of data, machine learning discovers algorithms that allows the machines to learn by itself without human intervention. Some examples of machine learning algorithms are Neural Networks, Decision Trees, Naïve Bayes, and Logistic Regression. K-Means, initiated by Mc. Queen in 1967 is the most popular and relevant clustering model [21]. Predictive classification models have been used to study customer purchasing behavior in past researches [13], [22], [23] in which classification models like K-Means and Decision Tree were commonly employed. This study explores these two models for segmenting and classifying customers.

B. Customer Segmentation via Clustering Methods

Past research shows that K-Means Clustering method has been widely used. K-Means Clustering was used to segment bank's customers whereby customers were grouped into five categories: potential growth customers, general customers, intermediate customers, senior customers and VIP customers [24]. Meanwhile, K-Means Clustering algorithm was also applied to segmentize private banking customers and the results showed three clusters named 'Core Value Customers', 'Financial Products Oriented Customers' and 'Deposit Oriented Customers' [25].

Khalili-Damghani et al. [9] employed K-Means Clustering for insurance customers segmentation. The results were three clusters labelled as 'profitable customer', 'potential profitable customers' and 'disinterested customers. Fuzzy C-Means clustering was used to cluster life insurance customers [26]. The results explained that two was the optimal number of clusters for the study which denoted as "investment" and "life security". In [3], a comparison of k- prototypes was conducted which combined K-means (for numerical element) and K-modes (for categorical element) algorithms, improved k-prototypes and SBAC (Similarity-Based Agglomerative Clustering (SBAC) for customer segmentation in auto insurance case study. The results showed that SBAC algorithm is more effective in clustering auto insurance customers with higher silhouette index value.

In another study by Qadadeh & Abdallah [27], K-Means Clustering and Self-Organizing Map (SOM) techniques were used to cluster insurance customers. The comparison was made between K-Means Clustering and the combination of SOM with K-Means Clustering which resulted in a better overall performance of the combined method with six clusters of customers. Further studies on SOM had been applied on imbalanced dataset for clustering categorical data in which Kohonen SOM (KSOM) algorithm was improved by focusing on the distance calculation amongst objects [35][36]. Another study on K-Means for clustering was done in [28] whereby K-Means algorithm was used to analyze the network traffic trend and type of traffic in campus network. The result showed that it was beneficial for managing or shaping the bandwidth usage and strengthens the security policy of the network.

K-Modes are generally the extended version of K-Means algorithm. The dissimilarity measure applied in K-Means algorithm is the reason that K-Means is unable to cluster categorical variables [29]. K-Modes clustering algorithm is introduced by Huang [30] by presenting a new measurement of dissimilarity to cluster categorical attributes [31]. While maintaining its proficiency, K-Modes clustering model eliminates the numeric data restriction. K-Modes removed the constraints imposed by K-Means through some adjustments including the usage of simple matching dissimilarity measure or hamming distance for categorical attributes and the replacement of means of cluster to the modes of cluster. The frequency-based approach is utilized by this model in updating the modes during clustering procedure to decrease the cost function which is estimated by calculating the standardized sum of within sum errors.

C. Rules Extration using Decision Tree Classifier

Clustering and classification techniques complement each other and are proved to perform well in segmenting customers. Clustering methods which are good at handling data without any labels have a setback of not being able to predict new and unknown data. On the other hand, classification methods are able to perform prediction to a set of unknown data but need to be trained by a set of labelled data. Decision Tree works in a way to guarantee the similarity of the sub-groups by splitting data points into two or more sub-categories [17]. A feature is represented by each node of the tree, and a value or a range of values for the feature that represents the node is portrayed by each edge aroused in a node [15]. The final output of the classification, known as class label, is stored in a leaf node. The comparatively straightforward process of Decision Tree makes it easy to understand and interpret, and the process that addresses a number of data intricacy that usually presents in the real data makes the method popular [32]. Hypotheses on each feature's own influence in the classification procedure are produced with the help of the decision rules uncovered on the pathways [15].

There are several applications that implement Decision Tree as classifiers. Clustering analysis using K-Medoid Clustering was performed on family farmers in Brazil, and

used Decision Tree aside from Support Vector Machine, Neural Network (Multilayer Perceptron) to identifying character that distinguish between those identified clusters [15]. Classifications using Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbour, and Naïve Bayes were performed to predict the churning of credit card holders [16]. In a study by Ganjali and & Teimourpour [33] on life insurance customers, K-Means Clustering was used to group the customers based on their lifetime value. The researchers also performed association rules to the most valuable customer group as well as classification to predict position of new customers in each cluster. In [34], Decision Tree was used in job profiling analytics to select the most significant skillsets for each job position intelligently. It produced accuracy of 63.5% when used together with Capacity Utilization Rate. Decision Tree approach was used for classification process and the results showed that the model achieved 61.3% accuracy, 38.97 % classification error, 0.012% Kappa and 0.024% Correlation criteria.

Based on the previous research, K-Means Clustering and Decision Tree Classifier have been proven as the most popular clustering technique to group and classify customers across industries. Nevertheless, there is very limited study that perform and model categorical data which is proposed in this study. K-Means is only suitable for numerical data, whereas K-Modes is the extension of K-Means algorithm which can handle categorical data.

III. Methodology

This section presents the methodology of the study. Fig. 1 illustrates the four main steps which are detailed in the next subsections: 1) Data; 2) Variable Selection; 3) Model Development, and 4) Model Evaluation.

A. Data Preparation

The data used in this study was obtained from one of the life insurances companies in Malaysia. The total number of data was 37,181 records and it consisted of daily new business customers information including their demographic details and their policy information ranging from January 2018 until December 2019. Prior to conducting analysis, the data underwent a pre-processing phase including handling missing values, imputing outliers as well as transforming the variables. Missing values were imputed accordingly with blanks, while detected outliers were transformed. Data transformations methods include discretizing numerical variables via quantile-based approach, re-grouping of data in certain variables as well as changing data types. Discretization results in either conversion of some variables into categorical data, re-labelled to avoid redundancy or merged accordingly. Table A1 in Appendix shows the pre-processed variable description.

B. Variable Selection

Variable selection involves selecting attributes that provides meaningful insights towards targeting the right customers. Based on the data used, several variables were removed as they did not have impact in the analytics including 'Occupation Group', 'Distribution Channel', 'Insured (Self/Others)', 'Occupation Group', 'Payment Frequency', 'Premium Status', 'Race PO' and 'Sum Assured'. Additionally, business expert has suggested including some of the information regarding the policy purchased by the customers including duration of policy issuance, annual net premium, premium payment method, product type and policy status. Table I shows the final attributes selection.

C. Model Development

This study developed customer segmentation model using K-Modes and Decision Tree Classifier. Python language was used to perform the modelling for K-Modes Clustering and Decision Tree Classifiers. The first model, K-Modes was implemented with cost function in getting the minimize distance for the intra cluster distance. The number of clusters was set into k = 2, 3, 4 and 5. Then, the output for clustering is compared and evaluated in determining the best number of clusters. The optimal K value for K-Modes was determined by using Elbow Method. Fig. 2 shows that the elbow shape is detected when number of clusters suggested was 3 using cost function value and the precise value of the cost function is given in Table II. Therefore, K-Modes clustering with K=3 was chosen as the best number of cluster.

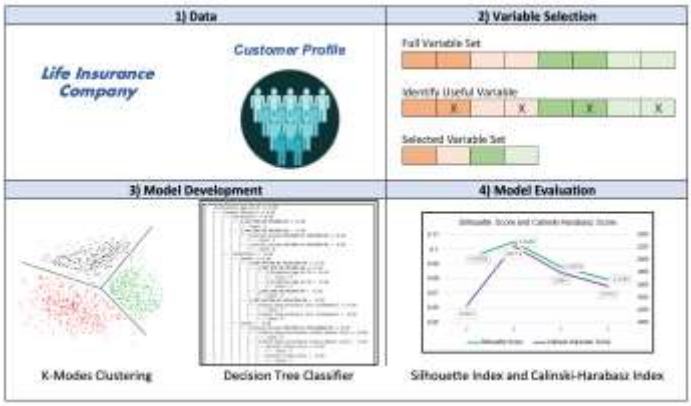


Fig. 1. Workflow Representation of the Methodology used in this Study.

TABLE I. Final Selected Attributes for Analysis

Index	Attribute	Attribute Type
1.	Annual Income PO	Ordinal
2.	Annual Net Premium (ANP)	Ordinal
3.	Client Type	Binary
4.	Duration of Issuance (Days)	Ordinal
5.	Gender PO	Binary
6.	Inception Age PO	Ordinal
7.	Location PO	Nominal
8.	Marital Status PO	Nominal
9.	Occupation Risk Class	Ordinal
10.	Payment Method	Nominal
11.	Active Policy (Y/N)	Binary
12.	Product Type	Nominal

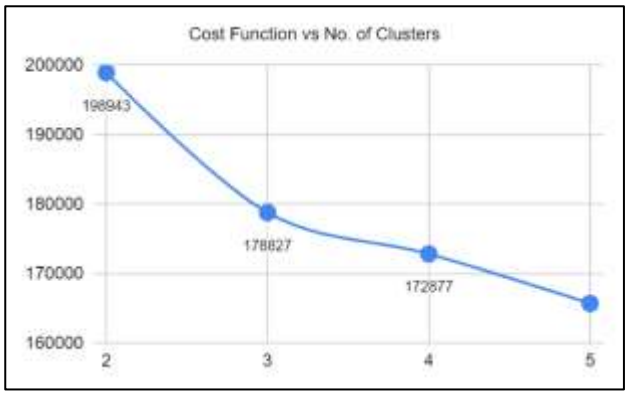


Fig. 2. Cost Function Plot.

The second model is a Decision Tree Classifier that was implemented with a built-in function of Python's Scikit Learn package. This function applies optimized CART (Classification and Regression Trees) in which it can perform well for binary classification as well as multi-class classification. The classifier was tuned by using the criterion function. This study experimented with two criteria of Decision Tree, 'Gini' and 'Entropy' with k-fold cross validation approach to achieve the best fit model. The number of labels was determined by the best number of clusters by K-modes; and in this study, 3 labels was defined for the classification task with rule extraction. The experiments were based on different number of clusters with evaluation of the cost function, thus the development time of clusters did not give any significant value, with 0.001 - 0.005 differences.

D. Model Evaluation

The clustering validation was done by using Silhouette Index score. The expectation of a good clustering is the shorter distance between each point in a cluster, the farther distance between clusters and a balanced proportion of data points among clusters. The equation of Silhouette Index is as shown in (1) [9]:

$$Silhouette = \frac{b(i)-a(i)}{\max\{a(i);b(i)\}} \quad (1)$$

where *a(i)* is the non-similarity between one object and other objects in the same cluster, and *b(i)* is the non-similarity between one object and other objects in the closest cluster. Evaluation using Calinski-Harabasz Index, CH was also performed to justify the performance of the clusters based on the formula shown in [9]:

$$\mathrm{CH}(q) = \frac{\frac{trace(B_q)}{(q-1)}}{\frac{trace(W_q)}{(n-q)}} \quad (2)$$

where *n* is number of records, *q* is number of clusters, W_q is intra-cluster scatter matrix, B_q is inter-cluster scatter matrix. The highest score portrays the best number of clusters for the dataset.

Meanwhile for the classification, the evaluation was performed using K-fold cross validation whereby the datasets were randomly divided into K equally sized subsets. The models were trained, and tested K times and the results were determined during each phase. The final accuracy of the models was measured as the average of all accuracies obtained in every iteration made. The formula for accuracy is as shown in (3) [15]:

$$\text{Accuracy } \% = \frac{c}{n} \times 100\% \quad (3)$$

where *c* is number of test samples classified correctly and *n* is total number of test samples.

IV. Results

A. Clustering Analysis

The results of distribution of each attribute in each cluster with K=3 is shown in Table A2 in Appendix. Based on the cluster analysis, it is shown that 51% (19,047) of the total observation falls under Category 0. This group has the highest percentage of young working customers with a relatively low annual income, and they aged in the range of 26 to 34 years old (37.7%) and earn MYR27,000.01 until MYR42,000.00 yearly (35.1%). The distribution of gender shows more than half of the customers are female and are married. Top residential location for customers in this group is Central Malaysia with 25.8, and slightly more than half of the customers in this group opt to pay low annual net premium which is in the range of MYR0 – MYR 1,800. Majority of the customers are new customers (88.9%) who purchased policies for the first time between of year 2018 and 2019. In addition, more than half of the customers belong to Class 1 of occupational risk which means that their occupations are having the least risk of exposure towards hazardous elements. For payment method, most of the customers use Auto Debit (80.5%) to pay their premium. The highest percentage of the policies' issuance days goes to '0 – 186 days' category (34.4%). Lastly, more than half of the customers purchase Ordinary Life (Endowment) products (68.7%) and their policies remain active (72.0%) at the end of year 2019. Hence, Cluster 0 is named as Low-Value Customers.

Cluster 1 makes up 27% (10,081) of the whole dataset. Customers are in young group aged between 10 to 25 years old (62%) and almost half of the customers have a low annual income from MYR0.00 – MYR27,000.00. Males are higher customers (68.8%) compared to female. Majority of the customers in this cluster are single (81.0%) and most of them reside in Northern Malaysia. In terms of the occupational risk class, more than of the customers belong to Class 1. Almost all customer in this group are new customers (90.6%). The annual net premium paid by the customers in this group are mostly between MYR1,800.01 – MYR 2,400.00 with 40.1% and they also prefer Auto Debit (75.7%) as the method to pay their premium. The distribution has the highest percentage for '187

– 334 days' duration policy (33.1%) and more than half of the customers purchase Investment-Linked (Whole Life) products. Lastly, the distribution of policy status is almost balance for this cluster with slightly higher percentage of inactive policies (55.3%). This means that this group of customers has higher chances to turn their policies inactive. Cluster 1 is labelled as Disinterested Customers.

On the other hand, Cluster 2 makes up of 22% (8,053) of the total observation. Majority of the customers in this cluster are older customers with more stable earnings since they have a high annual income. 49.3% of them age in the range of 42 – 76 years old and 61.8% of them have annual income in the range of MYR67,000.01 - MYR1,400,000.00. Moreover, more than half of the customers are male and 80% of the customers are married. Central Malaysia has the highest percentage for this cluster with 49.2%. This cluster also has the highest percentage of customers who belong to Class 1 hence their occupation is not very risky. Besides that, this cluster has a slightly higher percentage of existing customers (58.6%) compared to new customers.

Aligned with the range of annual income, this group of customers has the highest percentage of annual net premium in the range of MYR3,380.01 - MYR369,200.00 which is the highest category of annual net premium in this dataset with 51.5%. This cluster also has the highest percentage of those who issued their policy between 335 to 543 days with 35.5%. This group of customers prefers Credit Card the most with 58.4% as the medium to pay their premium to the insurer. For product type, the customers mainly purchase Investment-Linked (Whole Life) products with 79.5%. Finally, majority of the policies purchased are still active as of December 31st, 2019 with 84.3%. Cluster 2 is called as Potential High-Value Customers.

The cluster performance can be measured by evaluating intra-cluster performance and hence, we implemented Silhouette Index and Calinski-Harabasz Index. The results are shown in Table II. The cost function values are also included in the table for analysis purpose. Based on Table II, for Silhouette and Calinski-Harabasz Indexes, the scores need to be the highest to have the best cluster performance. In this study, it is shown that both scores are the highest when the number of clusters used are 3, as shown in Fig. 3. For the cost function, the value is decreasing when we add a greater number of clusters. However, the largest difference of the cost value is when the number of clusters is changed from 2 to 3 compared to the change from 3 to 4 clusters and 4 to 5 clusters which results in the elbow shape. Therefore, it is justified that K-Modes with 3 clusters has the best performance for this study.

B. Classification Analysis

The purpose of performing classification is to predict the characteristics of each class label by extracting the rules developed by Decision Tree. There was a total of 43 attributes including the target variable. We implemented K-Fold Cross Validation where it divides the dataset into K-folds and they have roughly the same size of samples. In this study, we performed experiments on both 'Gini' and 'Entropy' criteria for Decision Tree Classifier and the number of folds selected are 2, 5 and 10. We also set the maximum number of leaf nodes to 50 to ease the validation of the decision rules as this parameter enable the model to grow a tree in best-first decisions. Table III shows the outputs selected at random for all the experiments.

The performance evaluation of Decision Tree classification is done by comparing the accuracy of the models in each experiment. Since the experiments are implemented based on the k-folds cross validation method, the average accuracy for each model is compared. Referring to Table IV and Fig. 4, it is shown that the accuracy of the models increases as the larger value of K is used. It can be concluded that, Decision Tree classifier with Gini criterion and 10-fold cross validation is the best fit model for this dataset as it has the highest average accuracy compared to other models with 81.30%.

TABLE II. INTRA-CLUSTER PERFORMANCE EVALUATION

No. of Cluster	Cost Function	Silhouette Score	Calinski-Harabasz Score
2	198943.0	0.0935	1259.5
3	**178827.0**	**0.1049**	**2217.9**
4	172877.0	0.0878	1789.4
5	165769.0	0.0794	1574.6

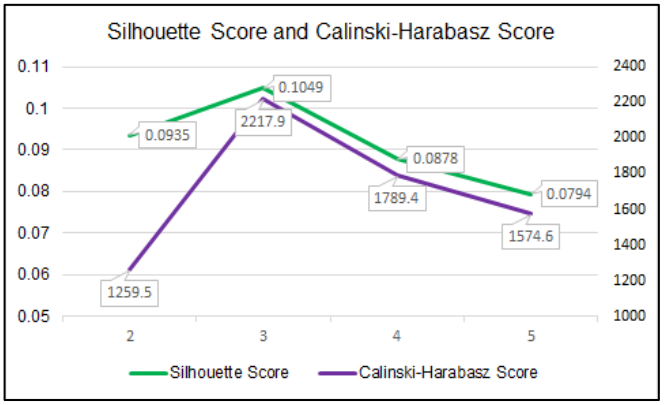


Fig. 3. Silhouette and Calinski-Harabasz Index for each cluster.

TABLE III. RULES OUTPUT GENERATED (RANDOM) FROM EXPERIMENTS

Exp.	Criteria	K	Rule 1	Rule 2	Rule 3
1	Entropy	2	If Marital Status is NOT Single, Payment Method is NOT Auto Debit, Annual Income is NOT MYR67000.01-MYR1400000.00 and Product Type is NOT Investment-Linked (Whole Life), hence, Cluster 0.	If Marital Status is NOT Single, Payment Method is NOT Auto Debit, Annual Income is MYR67000.01-MYR1400000.00 and Client Type is Existing Client, hence, Cluster 2.	If Marital Status is Single, Inception Age is NOT 10 – 25 Years Old, Policy is NOT Active, Product Type is NOT Investment-Linked (Whole Life) and Gender is Male, hence, Cluster 1.
2	Gini		If Inception Age is NOT 10 – 25 Years Old, Payment Method is NOT Auto Debit, Client Type is Existing Client, ANP is NOT MYR0.00-MYR1800.00, hence, Cluster 2.	If Inception Age is NOT 10 – 25 Years Old, Payment Method is NOT Auto Debit, Client Type is New Client, Gender is Male, Annual Income is NOT MYR67000.01-M1400000.00, Product Type is Investment-Linked (Whole Life) and Marital Status is Single, hence, Cluster 1.	If Inception Age is 10 – 25 Years Old, Gender is Female, Product Type is NOT Investment-Linked (Whole Life) and Annual Income is NOT MYR0.00 – MYR27000.00, hence, Cluster 0.
3	Entropy	5	If Marital Status is NOT Single, Payment Method is NOT Auto Debit, Annual Income is NOT MYR67000.01 - MYR1400000.00, Client Type is Existing Client and ANP is NOT MYR0.00-MYR1800.00, hence, Cluster 2.	If Marital Status is Single, Inception Age is NOT 10-25 Years Old, Product Type is NOT Investment-Linked (Whole Life) and Gender is Female, hence, Cluster 0.	If Marital Status is Single, Inception Age is 10-25 Years Old, Gender is Female, Product Type is Investment-Linked (Whole Life) and Annual Income is MYR0.00-MYR27000.00, hence, Cluster 1.
4	Gini		If Inception Age is NOT 10–25 Years Old, Payment Method is Auto Debit, Product Type is NOT Investment-Linked (Whole Life) and Gender is Female, hence, Cluster 0.	If Inception Age is NOT 10-25 Years Old, Payment Method is Auto Debit, Product Type is Investment-Linked (Whole Life), Gender is Female, ANP is MYR3380.01-MYR369200.00 and Client Type is Existing Client, hence, Cluster 2.	If Inception Age is 10-25 Years Old, Gender is Female, Product Type is Investment-Linked (Whole Life) and Annual Income is MYR0.00-MYR27000.00, hence, Cluster 1.
5	Entropy	10	If Marital Status is NOT Single, Payment Method is Auto Debit, Product Type is NOT Investment-Linked (Whole Life), Gender is Female and Client Type is New Client, hence, Cluster 0.	If Marital Status is NOT Single, Payment Method is Auto Debit, Product Type is Investment-Linked (Whole Life), Annual Income is NOT MYR67000.01-MYR1400000.00, Gender is Male, Policy is Active, Client Type is New Client and Inception Age is 43-76 Years Old, hence, Cluster 2.	If Marital Status is Single, Inception Age is 10-25 Years Old, Gender is Male, Product Type is Investment-Linked (Whole Life), Client Type is New Client, hence, Cluster 1.
6	Gini		If Inception Age is NOT 10-25 Years Old, Payment Method is Auto Debit, Product Type is NOT Investment-Linked (Whole Life), Gender is Male, Marital Status is Married and Client Type is Existing Client, hence, Cluster 2.	If Inception Age is 10-25 Years Old, Gender is Male, Issuance Duration is NOT 0-186 Days and Marital Status is Married, hence, Cluster 1.	If Inception Age is 10-25 Years Old, Gender is Male, Issuance Duration is 0-186 Days, Product Type is NOT Investment-Linked (Whole Life) and Policy is Active, hence, Cluster 0.

TABLE IV. MODEL EVALUATION RESULTS (K-FOLD CROSS VALIDATION)

K-Fold Cross Validation \ Criterion	**Entropy**	**Gini**
2-Fold Cross Validation	76.03%	77.81%
5-Fold Cross Validation	78.49%	80.19%
10-Fold Cross Validation	79.29%	**81.30%**

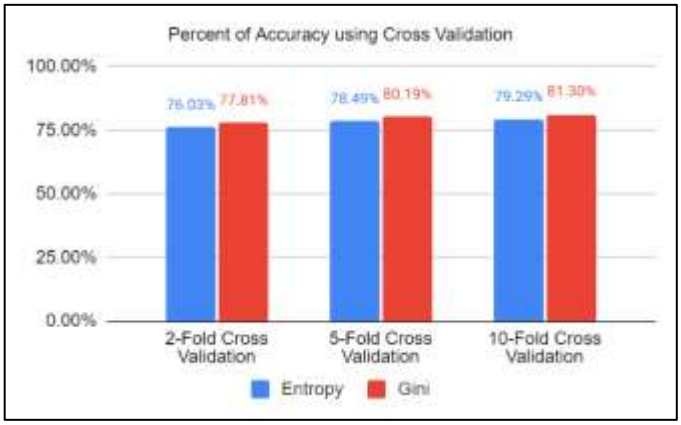


Fig. 4. The Accuracy of Classifiers for different K Fold.

V. DISCUSSION

Customer segmentation analysis is crucial for insurance companies to identify who are the profitable customers, how many percentages of them from the total population, to find more clients with similar profiles and how to manage less profitable clients [37]. Based on the results from this study, it can be concluded that customer segmentation can be achieved by using data mining techniques. Both clustering and classification methods are complementing the outcome of customer segmentation.

Once the customer segmentation is identified, there are some strategies that may be incorporated to cater each group of customers. For those customers who fall under 'Potential High-Value Customers', insurance company may want to focus more on this group to make the customers stay loyal to the company for a long period of time by providing good services to them. Since this group of customers contribute a lot to the company, the insurer may want to find more opportunities to sell more products to them based on their needs [9]. Insurer also needs to keep in touch with the customers from time to time to update their condition and keep on track of their well-being so that they feel comfortable with the company [9][38]. The insurance companies could build trusts relationship with the customers and this would increase cross-selling and up-selling [38]. Being

aware on customers' triggering events such as having a baby or buying a house could increase the product sales even more significantly, as according to the research done by [38]. Furthermore, insurer may provide a better customer experience by providing such a strategic and tactical focus based on the five key organizational process i.e., making strategic choices, creating value for customers, customer acquisition, customer retention, service quality and loyalty or rewards program, which can be achieved with a good CRM tool [39].

Nevertheless, the company must have strategies for those in group of 'Low-Value Customers' and 'Disinterested Customers'. For 'Low-Value Customers', although they are not the main contributors to the company, they are still the customers who are willing to take a chance in trusting the insurance company. Insurance companies may want to adopt customer centric approach to provide a superior customer experience [40]. This includes providing better customer services towards more customization and personalization by providing appropriate channels for communication to keep the customers informed, demanding and connected.

On the other hand, upon detecting the customers who fall into 'Disinterested Customers' group, insurance companies may be able to discuss and advise them to keep their policies in the event of customers cancelling the policies. In a study done by [39], it is evident that customers demand more on service quality, interaction management, contact programs, retention management, service strategy, customer satisfaction and customer loyalty, and this also could be achieved with a systematic CRM tool in place.

VI. Conclusion and Future Work

This study has presented the work on customer segmentation and profiling for insurance industry by using K-Modes clustering and Decision Tree Classifiers. The grouping of customers is made by analyzing the similarity of their characteristics and hence, able to determine the target customers. It is highly recommended for life insurance companies to segmentize their customers to enable them to offer suitable products or services in accordance with the needs of customers.

Future researchers may consider using a larger dataset with longer time periods to perform customer segmentation to have a more accurate result. If the data is too large, they may consider performing dimensional reduction technique such as Principle Component Analysis (PCA) to handle the data by transforming them into useful components. It is also suggested that future studies should use transactional details of the customers to monitor their behaviors and include more product categories such as Credit Life Insurance products as well as all rider products purchased by customers.

Further future work could also include result comparison with other classification models such as Random Forest, Naïve Bayes or even Artificial Neural Network (ANN). Also, computational complexity analysis could also be studied to analyze the learning efficiency and performance while implementing customer segmentation. The proposed approach in this study can also be applied in other industries like retail, hospitals, food chains, bookstores and so forth.

Acknowledgment

The authors would like to thank the Faculty of Computer and Mathematical Sciences, and the Research Management Centre (RMC), Universiti Teknologi MARA, Malaysia, for the support throughout this research.

References

[1] Cummins, M. Cragg, and B. Zhou, "The Social and Economic Contributions of the Life Insurance Industry," no. September, p. 39, 2018, [Online]. Available: https://brattlefiles.blob.core.windows.net/files/14446_life_insurance_industry_white_paper_final_2018.pdf.

[2] R. Srivastava, "Identification of customer clusters using RFM model: a case of diverse purchaser classification," Int. J. Information, Bus. Manag., vol. 9, no. 4, pp. 201–208, 2017.

[3] K. Zhuang, S. Wu, and X. Gao, "Auto insurance business analytics approach for customer segmentation using multiple mixed-type data clustering algorithms," Teh. Vjesn., vol. 25, no. 6, pp. 1783–1791, 2018.

[4] E. S. Levy, "Repeat Business is Online Retail" s Core Customer Metric," Thursday June 5th, eNewsletter. Core Cust. Metr., 2008.

[5] J. Griffin and M. W. Lowenstein, Customer winback: How to recapture lost customers--And keep them loyal. John Wiley & Sons, 2002.

[6] M. H. Hosseini, O. Mohammad MahmoudiMaymand, and M. Ahmadinejad, "Predicting the Bank Customer Switching Based on Data Mining Technique," Spectr. A J. Multidiscip. Res., vol. 2, no. 10, pp. 637–2278, 2013.

[7] F. Abdi, K. Khalili-Damghani, and S. Abolmakarem, "Solving customer insurance coverage sales plan problem using a multi-stage data mining approach," Kybernetes, 2018.

[8] M. Tarokh and K. Sharifian, "Applications of data mining in improving customer communication management," Iran. Ind. Manag. Stud. Q., vol. 6, no. 17, pp. 153–181, 2010.

[9] K. Khalili-Damghani, F. Abdi, and S. Abolmakarem, "Insurance customer segmentation using clustering approach," Int. J. Knowl. Eng. Data Min., vol. 4, no. 1, pp. 18–39, 2016.

[10] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "RFM ranking – An effective approach to customer segmentation," J. King Saud Univ. - Comput. Inf. Sci., 2018, doi: 10.1016/j.jksuci.2018.09.004.

[11] A. Ansari and A. Riasi, "Customer Clustering Using a Combination of Fuzzy C-Means and Genetic Algorithms," Int. J. Bus. Manag., vol. 11, no. 7, p. 59, 2016, doi: 10.5539/ijbm.v11n7p59.

[12] F. H. Bin Yusoff and N. L. A. B. Rosman, "A Case Study of Customers' Payment Behaviour Analytics on Paying Electricity with RFM Analysis and K-Means," in International Conference on Soft Computing in Data Science, 2019, pp. 40–55.

[13] N. M. N. Mathivanan, N. A. M. Ghani, and R. M. Janor, "Analysis of K-Means Clustering Algorithm: A Case Study Using Large Scale E-Commerce Products," in 2019 IEEE Conference on Big Data and Analytics (ICBDA), 2019, pp. 1–4.

[14] Y. Lu, A. Ioannou, I. Tussyadiah, and S. Li, "Segmenting travelers based on responses to nudging for information disclosure," e-Review Tour. Res., vol. 17, no. 3, pp. 394–406, 2019.

[15] C. Maione, D. R. Nelson, and R. M. Barbosa, "Research on social data by means of cluster analysis," Appl. Comput. Informatics, vol. 15, no. 2, pp. 153–162, 2019, doi: 10.1016/j.aci.2018.02.003.

[16] R. Rajamohamed and J. Manokaran, "Improved credit card churn prediction based on rough clustering and supervised learning techniques," Cluster Comput., vol. 21, no. 1, pp. 65–77, 2018.

[17] J. Asare-Frempong and M. Jayabalan, "Predicting customer response to bank direct telemarketing campaign," 2017 Int. Conf. Eng. Technol. Technopreneurship, ICE2T 2017, vol. 2017-Janua, no. December, pp. 1–4, 2017, doi: 10.1109/ICE2T.2017.8215961.

[18] Y.-H. Liang, "Integration of data mining technologies to analyze customer value for the automotive maintenance industry," Expert Syst. Appl., vol. 37, no. 12, pp. 7489–7496, 2010.

[19] K. Umamaheswari and S. Janakiraman, "Role of data mining in insurance industry," Int J Adv Comput Technol, vol. 3, pp. 961–966, 2014.

[20] E. W. T. Ngai, L. Xiu, and D. C. K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," Expert Syst. Appl., vol. 36, no. 2, pp. 2592–2602, 2009.

[21] R. A. Soeini and K. V. Rodpysh, "Applying data mining to insurance customer churn management," Int. Proc. Comput. Sci. Inf. Technol., vol. 30, pp. 82–92, 2012.

[22] N. Isa, N. S. M. Yusof, and M. A. Ramlan, "The implementation of data mining techniques for sales analysis using daily sales data," Int. J. Adv. Trends Comput. Sci. Eng., vol. 8, no. 1.5 Special Issue, pp. 74–80, 2019, doi: 10.30534/ijatcse/2019/1681.52019.

[23] Y. B. Wah, N. H. Ismail, and S. Fong, "Predicting car purchase intent using data mining approach," in 2011 eighth international conference on fuzzy systems and knowledge discovery (FSKD), 2011, vol. 3, pp. 1994–1999.

[24] D. Dong, J. Zhang, and J. Ye, "Research on Customer Segmentation Method of Commercial Bank Based on Data Mining," no. Icidel, pp. 62–65, 2017, doi: 10.25236/icidel.2017.016.

[25] X. Yang, J. Chen, P. Hao, and Y. J. Wang, "Application of clustering for customer segmentation in private banking," in Seventh International Conference on Digital Image Processing (ICDIP 2015), 2015, vol. 9631, p. 96311Z.

[26] G. Jandaghi and Z. Moradpour, "Segmentation of Life Insurance Customers Based on their Profile Using Fuzzy Clustering," Int. Lett. Soc. Humanist. Sci., vol. 61, no. October 2015, pp. 17–24, 2015, doi: 10.18052/www.scipress.com/ilshs.61.17.

[27] W. Qadadeh and S. Abdallah, "Customers Segmentation in the Insurance Company (TIC) Dataset," Procedia Comput. Sci., vol. 144, pp. 277–290, 2018, doi: 10.1016/j.procs.2018.10.529.

[28] M. A. M. Ariffin, R. Ishak, S. A. Ahmad, and Z. Kasiran, "Network traffic profiling using data mining technique in campus environment," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 1.3 Special Issue, pp. 422–428, 2020, doi: 10.30534/ijatcse/2020/6691.32020.

[29] S. S. Khan and S. Kant, "Computation of Initial Modes for K-modes Clustering Algorithm using Evidence Accumulation,", 2007, pp. 2784–2789.

[30] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD), 1997, pp. 21–34.

[31] S. A. Sajidha, S. P. Chodnekar, and K. Desikan, "Initial seed selection for K-modes clustering–a distance and density based approach," J. King Saud Univ. Inf. Sci., 2018.

[32] H. A.Elsalamony, "Bank Direct Marketing Analysis of Data Mining Techniques," Int. J. Comput. Appl., vol. 85, no. 7, pp. 12–22, 2014, doi: 10.5120/14852-3218.

[33] M. Ganjali and B. Teimourpour, "Identify Valuable Customers of Taavon Insurance in Field of Life Insurance with Data Mining Approach Keyword :," vol. 4, 2016.

[34] E. A. Kamaru Zaman, A. F. Ahmad Kamal, A. Mohamed, A. Ahmad, and R. A. Z. Raja Mohd Zamri, "Staff employment platform (StEP) using job profiling analytics," Commun. Comput. Inf. Sci., vol. 937, no. September 2020, pp. 387–401, 2019, doi: 10.1007/978-981-13-3441-2_30.

[35] A. Ahmad, R. Yusoff, M. N. Ismail, and N. R. Rosli,, "Clustering the imbalanced datasets using modified Kohonen self-organizing map (KSOM)," 2017 Computing Conference, London, 2017, pp. 751-755. doi: 10.1109/SAI.2017.8252180.

[36] A. Ahmad, and R. Yusof, R, "A Modified Kohonen Self-Organizing Map (KSOM) Clustering for Four Categorical Data," Jurnal Teknologi, 2016, vol. 72 no. 1, pp. 1–6.

[37] C. Matiş, L. Ilieş, "Customer Relationship Management in the Insurance Industry," Procedia Economics and Finance, vol. 15, 2014, pp. 1138-1145.

[38] J. Godsall, A. Jain, K. Javanmardian, F. Nauck, S. Ray, and S. Yang, "Unlocking the next horizon of growth in the life insurance industry," McKinsey&Company, Sep 2017, https://www.mckinsey.com /~/media/McKinsey/Industries/Financial%20Services/Our%20Insights/ Unlocking%20the%20next%20horizon%20of%20growth%20in%20the %20life%20insurance%20industry/Unlocking-the-next-horizon-of-growth-in-the-life-insurance-industry.pdf

[39] E. A. Kumar, "Customer Relationship Management (CRM) Practices in Life Insurance Industry," Shanlax International Journal of Commerce 5 (4),2017, pp. 77-84.

[40] Y. Michaux, "Four ways insurance companies are improving their customer experience," IBM, 18 May 2021, https://www.ibm.com /blogs/services/2021/05/18/four-ways-insurance-companies-are-improving-their-customer-experience/

APPENDIX

TABLE AI PRE-PROCESSED VARIABLES DESCRIPTION

Attribute	Attribute Type	Description and Frequency of Data
Active Policy (Y/N)	Binary	Indicator whether the policy is still active 0: No (N) (32.8%) 1: Yes (Y) (67.2%)
Annual Income PO	Ordinal	Annual Income of Policy Owner (Customer) 0: MYR0.00-MYR27000.00 (25.0%) 1: MYR27000.01-MYR42000.00 (25.5%) 2: MYR42000.01-MYR67000.00 (24.6%) 3: MYR67000.01-MYR1400000.00 (25.0%)
Annual Net Premium (ANP)	Ordinal	Annual Net Premium (ANP) amount 0: MYR0.00-MYR1800.00 (38.0%) 1: MYR1800.01-MYR2400.00 (26.7%) 2: MYR2400.01-MYR3380.00 (10.3%) 3: MYR3380.01-MYR369200.00 (25.0%)
Client Type	Binary	Whether customer is an existing or new customerrecorded in the 2 years 0: Existing (20.9%) 1: New (79.1%)
Distribution Channel	Binary	Channel of which the policy is sold to the customers 0: Agency (39.4%) 1: Bancassurance (60.6%)

Duration of Issuance (Days)	Ordinal	Duration of days since the issuance date of thepolicy 0: 0-186 days (25.1%) 1: 187-334 days (25.0%) 2: 335-543 days (24.9%) 3: 544-729 days (25.0%)
Gender PO	Binary	Gender of Policy Owner (Customer) 0: Female (F) (50.7%) 1: Male (M) (49.3%)
Inception Age PO	Ordinal	Age of Policy Owner (Customer) at point ofpurchase of the insurance policy 0: 10-25 years old (25.3%) 1: 26-34 years old (27.6%) 2: 35-42 years old (23.0%) 3: 43-76 years old (24.1%)
Insured (Self/Others)	Binary	Whom the insurance policy covered 0: Others (22.5%) 1: Self (77.5%)
Location PO	Nominal	Customer's residential (in region) 0: Central Malaysia (28.3%) 1: East Coast Malaysia (5.4%) 2: East Malaysia (19.6%) 3: Northern Malaysia (26.4%) 4: Other Country (0.2%) 5: Southern Malaysia (20.1%)
Marital Status PO	Nominal	Marital Status of Policy Owner (Customer) 0: Divorced (1.3%) 1: Married (57.4%)2: Single (40.5%)3: Widowed (0.8%)
Occupation Group	Nominal	Occupation group of the customer 0: Housewife (1.6%) 1: Retiree (0.3%) 2: Self-Employed (1.1%) 3: Student (3.1%) 4: Unemployed (0.1%) 5: Worker (93.8%)
Occupation Risk Class	Ordinal	Risk class of the customer's occupation 0: Class 1 (Least hazardous) (73.4%) 1: Class 2 (Less hazardous) (14.3%) 2: Class 3 (Moderate hazardous) (5.3%) 3: Class 4 (Most hazardous) (7.0%)
Payment Frequency	Nominal	Frequency of the premium paid 0: Single which represented by 0 (0.9%) 1: Annually which represented by 1 (12.4%) 2: Semi-annually which represented by 2 (1.5%) 3: Quarterly which represented by 4 (3.2%) 4: Monthly which represented by 12 (82.0%)
Payment Method	Nominal	Premium's payment method 0: Cash / Cheque (8.0%) 1: Auto Debit (67.0%) 2: No Billing / Single (0.9%) 3: Credit Card (23.8%) 4: Advance Premium Payment (0.3%)
Premium Status	Nominal	Status of premium paying 0: Cancelled (5.5%) 1: Deceased) (0.0%) 2: Premium Holiday (4.1%) 3: Lapsed (22.6%) 4: Premium Paying (62.3%) 5: Single Premium (0.9%) 6: Surrender (4.5%) 7: Terminated (0.0%)

Product Type	Nominal	Type and sub-type of the product purchased bycustomers 0: Investment-Linked (Whole Life) (47.5%) 1: Ordinary Life (Endowment) (48.4%) 2: Ordinary Life (Hospital & Surgical) (3.0%)3: Ordinary Life (Whole Life) (1.2%)
Race PO	Nominal	Policy Owner's (Customer) Race 0: Chinese (52.5%) 1: Indian (21.2%) 2: Malay (18.6%) 3: Others (7.7%)
Sum Assured	Ordinal	Policy's sum assured amount. 0: MYR150.00-MYR13500.00 (25.1%) 1: MYR13500.01-MYR30000.00 (26.6%) 2: MYR30000.01-MYR100000.00 (34.8%) 3: MYR100000.01-MYR1800000.00 (13.5%)

TABLE AII DISTRIBUTION OF ATTRIBUTES IN CLUSTERS

Attribute	Cluster 0 (51%)	Cluster 1 (27%)	Cluster 2 (22%)
Active Policy (Y/N)	Inactive = 28.0% **Active = 72.0%**	**Inactive = 55.3%** Active = 44.7%	Inactive = 15.7% **Active = 84.3%**
Annual Income PO	RM0.00-MYR27000.00 =21.0% **RM27000.01-MYR42000.00 = 35.1%** RM42000.01-MYR67000.00 = 27.0% MYR67000.01-RM1400000.00 = 16.9%	**RM0.00-MYR27000.00 =47.9%** RM27000.01-MYR42000.00 = 19.9% RM42000.01-MYR67000.00 = 21.4% MYR67000.01-RM1400000.00 = 10.8%	RM0.00-MYR27000.00 =5.9% RM27000.01-MYR42000.00 = 9.5% RM42000.01-MYR67000.00 = 22.8% **RM67000.01- MYR1400000.00 = 61.8%**
AnnualNet Premium (ANP)	**RM0.00-MYR1800.00 =50.6%** RM1800.01-MYR2400.00 = 22.9% RM2400.01-MYR3380.00 =6.5% RM3380.01-MYR369200.00 = 20.0%	RM0.00-MYR1800.00 =35.9% **RM1800.01-MYR2400.00 = 40.1%** RM2400.01-MYR3380.00 =10.8% RM3380.01-MYR369200.00 = 13.2%	RM0.00-MYR1800.00 =11.0% RM1800.01-MYR2400.00 = 18.9% RM2400.01-MYR3380.00 =18.6% **RM3380.01-MYR369200.00 = 51.5%**
Client Type	Existing = 11.1% **New = 88.9%**	Existing = 9.4% **New = 90.6%**	**Existing = 58.6%** New = 41.4%
Duration of Issuance(Days)	**0-186 days = 34.4%** 187-334 days = 20.5% 335-543 days = 19.7% 544-729 days = 25.4%	0-186 days = 12.0% **187-334 days = 33.1%** 335-543 days = 26.3% 544-729 days = 28.6%	0-186 days = 19.7% 187-334 days = 25.5% **335-543 days= 35.5%** 544-729 days = 19.3%
Gender PO	**Female = 67.8%** Male = 32.2%	Female = 31.2% **Male = 68.8%**	Female = 34.6% **Male = 65.4%**
InceptionAge PO	10-25 years old = 15.1% **26-34 years old= 37.7%** 35-42 years old = 26.0% 43-76 years old = 21.3%	**10-25 years old = 62.3%** 26-34 years old = 16.5% 35-42 years old = 12.1% 43-76 years old = 9.1%	10-25 years old = 3.4% 26-34 years old = 17.8% 35-42 years old = 29.5% **43-76 years old = 49.3%**
LocationPO	**Central Malaysia = 25.8%**East Coast Malaysia = 6.2%East Malaysia = 23.7% Northern Malaysia = 23.7%Other Country = 0.0% Southern Malaysia = 20.6%	Central Malaysia = 16.4% East Coast Malaysia = 5.7%East Malaysia = 21.6% **Northern Malaysia = 36.1%** Other Country = 0.1% Southern Malaysia = 20.0%	**Central Malaysia = 49.2%**East Coast Malaysia = 3.4%East Malaysia = 7.3% Northern Malaysia = 20.4%Other Country = 0.6% Southern Malaysia = 19.1%
Marital Status PO	Single = 30.1% **Married = 67.9%** Divorced = 1.2% Widowed = 0.8%	**Single = 81.0%** Married = 17.5% Divorced = 0.9% Widowed = 0.6%	Single = 14.4% **Married = 82.7%** Divorced = 2.0% Widowed = 0.9%
OccupationRisk Class	**1 = 74.2%** 2 = 13.2% 3 = 4.5% 4 = 8.1%	**1 = 68.1%** 2 = 17.3% 3 = 7.2% 4 = 7.4%	**1 = 78.2%** 2 = 13.2% 3 = 4.8% 4 = 3.8%
PaymentMethod	C = 6.2% **D = 80.5%** N = 1.0%	C = 5.3% **D = 75.7%** N = 0.5%	C = 15.5% D = 24.2% N = 1.4%

Attribute	Cluster 0 (51%)	Cluster 1 (27%)	Cluster 2 (22%)
	R = 12.0% Y = 0.3%	R = 18.3% Y = 0.1%	**R = 58.4%** Y = 0.6%
ProductType	Investment-Linked (WholeLife) = 26.7% **Ordinary Life (Endowment) = 68.7%** Ordinary Life (Hospital &Surgical) = 3.5% Ordinary Life (Whole Life) = 1.1%	**Investment-Linked (WholeLife) = 61.2%** Ordinary Life (Endowment) = 36.4% Ordinary Life (Hospital &Surgical) = 1.4% Ordinary Life (Whole Life) = 1.0%	**Investment-Linked (WholeLife) = 79.5%** Ordinary Life (Endowment) = 15.3% Ordinary Life (Hospital &Surgical) = 3.7% Ordinary Life (Whole Life) = 1.5%

TABLE AIII LIST OF ABBREVIATION

ANN	Artificial Neural Network
ANP	Annual Net Premium
CART	Classification and Regression Trees
CRM	Customer Relationship Management
KDD	Knowledge Discovery in Databases
KSOM	Kohonen Self-Organizing Map
MYR	Malaysian Ringgit
PCA	Principle Component Analysis
PO	Policy Owner
SBAC	Similarity-Based Agglomerative Clustering
SOM	Self-Organizing Map

Building a Standard Model of an Information System for Working with Documents on Scientific and Educational Activities

Serikbayeva Sandugash[1], Tussupov Jamalbek[2]
Department of Information Systems, L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan

Yerzhanova Akbota[4]
S. Seifullin Kazakh Agro Technical University
Nur-Sultan, Kazakhstan

Sambetbayeva Madina[3]
L.N. Gumilyov Eurasian National University, Institute of Information and Computational Technologies CSMES RK
Nur-Sultan, Almaty, Kazakhstan

Abduvalova Ainur[5]
NLC "Taraz Regional University Named after M.KH. Dulaty", Taraz
Kazakhstan

***Abstract*—To increase the effectiveness of research, it is necessary to have access to systematic information resources of scientific work. Therefore, in any field of science, it begins with research, the search for scientific information, but with the growing number of scientific articles, books, monographs, patents, the search for information becomes more and more difficult. Creating a unified information system that allows scientists to quickly get acquainted with the results of other scientific research and prevent their duplication. The article discusses the technological techniques of distributed information systems that provide scientific and educational activities. The main tasks of creating a model of a distributed information system that supports scientific and educational activities, the functional capabilities of the model, the concept of metadata and the requirements for the metadata profile are described. The task, subject area, subjects, objects, the main functionality of the information system are defined, a list of the main types of information resources is provided. The paper analyzes the functional requirements for such systems. The paper describes a technological approach to creating a standard model of an information system to support scientific and educational activities organized in the form of an electronic library for working with documents on scientific heritage. The article describes the architecture of the information system and the principles of integration with the digital depository, the rules for the presentation and transformation of metadata.**

***Keywords*—Scientific and educational activities; distributed information systems; electronic library; metadata; model; search; interoperability; document; ontology; Z39. 50; SRU/SRW; Apache Solr**

I. Introduction

The rapid development of information technologies and means of data transmission has led to qualitative changes in the solution of one of the most important tasks facing humanity – the preservation of information for the purpose of its transmission. Until now, the main form of storing and distributing information was printed publications, and the main means of accessing information were libraries. New information technologies have provided new opportunities for solving the problems of creating repositories of information resources, their organization, means and ways of accessing them by users. In a generalized form, such approaches have become known as "Electronic Libraries (EL)". The information service on printed media has been replaced by the provision of users based on the electronic presentation of a wide variety of information, replicated in unlimited quantities and quickly accessible over global computer networks, regardless of the time of access to it and the location of users. The digital library should be understood more broadly as an environment that brings together "collections, services, and people to support the full life cycle of creating, distributing, using, and preserving data, information, and knowledge". The main tasks of the electronic library are the integration of information resources and effective navigation in them. The integration of information resources is understood as their integration in order to use (with the help of convenient and unified user interfaces – preferably one) different information while preserving its properties, presentation features, and user manipulation capabilities. However, the pooling of resources does not have to be physically performed. The main thing is that it should provide the user with the perception of the available information as a single information space. In particular, electronic libraries should provide work with heterogeneous databases or database systems, providing the user with the effectiveness of information searches, regardless of the features of the specific information systems to which access is made [1].

Distributed information systems to support scientific and educational activities operate with various types of information. These can be publications, electronic documents, electronic collections, ontological descriptions, data sets, logical descriptions, etc. These resources, which are in demand by different groups of researchers, may not be available due to problems with their search and identification. Semantic links between information resources increase their value and provide additional opportunities for information search and identification.

Data integrated into an open semantic space is a collection of knowledge about a certain subject area. At the same time, the use of resources is accompanied by problems of determining the rules for obtaining up-to-date copies of resources and their transformation according to the specifics of a particular research environment. These problems are caused by the lack of adequate mechanisms to distribute the high-demand knowledge-intensive resources, as well as the heterogeneity of ways to represent and store resources in the form of files (binary, text, XML markup), spreadsheets, databases, electronic documents, catalogs, etc. Most often, these problems are solved in each individual case individually.

Special systems are needed to provide access to such resources. Therefore, providing universal ways to work with distributed and heterogeneous data, where it is not known in advance what types of objects the end user will have to work with, and unifying the presentation of this data, is the main task when integrating information resources in distributed information systems. At the same time, the idea of using Z39.50 technologies for resource integration looks very attractive, since to date, the Z39.50 standard is the only standard that regulates universal network access to databases based on an abstract data model [2].

The main requirement for information systems designed to support scientific and educational activities is interoperability.

The interoperability of any information system, including an electronic library, is understood as the degree of its ability to interact with other information systems, including with a person [3]. But if, when interacting with the latter (as with an information system), the main burden on ensuring mutual understanding falls on a person who is able to process even very poorly organized information, then special technological approaches and general agreements are required to ensure effective interaction between information systems. Ensuring the interoperability of systems is impossible without strict compliance with the relevant international standards and recommendations. At the same time, the standards must comply with:

- data access protocols and interfaces;
- search languages and interfaces;
- data representation schemes and formats;
- interfaces for visualizing the same type of data;
- rules for encoding information;
- data access control rules.

In [4], the basic profile of the information system standards for supporting scientific research, organized in the form of an electronic library, was defined.

The metadata profile refers to the adaptation of the existing metadata schema to the needs of a specific task being solved by the information system 1. Based on the analysis of the existing metadata formats intended for working with publications, documents and other information resources, it can be concluded that the most suitable format for research work with materials on scientific heritage is GOST 7.19-2001 (MEKOF). Compared to other commonly used metadata formats (formats of the MARC family), this format has the most complete classification system for document types and other information resources and a fairly large set of reference dictionaries necessary for describing and identifying information resources.

In this paper, we consider a technological approach to creating a standard model of an information system designed to support scientific research. The developed model of the information system for working with materials related to the scientific heritage should solve the problems of long-term storage of information, organization of abstract search by attributes, organization of collection and exchange of metadata and information between remote repositories of information resources.

In the information space, events, facts, and any other entities of the real world exist only in the form of documents [4]. As a result, the document is the main element of the system under consideration.

As you know, the information system (IS) and some parts of information systems are defined ambiguously in the scientific educational literature. Differences in definitions are usually caused by differences in the subject areas of application of the described IR and different approaches to the description of IR.

Let's consider a number of different information systems:

An information system is a set of information contained in databases and information technologies and technical means that ensure its processing [2].

An information system is an information processing system that works together with organizational resources, such as people, technical means, and financial resources that provide and distribute information [3].

An information system is a conceptual scheme, an information base, and an information processor that together make up a formal system for storing and manipulating information [4].

Information system: A system designed for storing, processing, searching, distributing, transmitting, and presenting information [5].

An information system is a material system that organizes, stores, and transforms information. This is a system in which the main subject and product of labor is information [6].

An information system is an applied software subsystem focused on the collection, storage, search and processing of textual and / or factual information [7].

Information system – a system designed for storing, processing, searching, distributing, transmitting and providing information [8].

A modern information system is a set of information technologies aimed at supporting the life cycle of information and including three main components of the process: data processing, management, information management and knowledge management [9].

Among Russian scientists in the field of computer science, the broadest definition of IR is given by M. R. Kogalovsky, in his opinion, an information system is a complex that includes computing and communication equipment, software, linguistic tools and information resources, as well as system personnel and provides support for a dynamic information model of some part of the real world to meet the information needs of users [10].

The part of the real world that is modeled by an information system is called its domain.

The dynamic model is understood as the variability of the model over time. This is a "live", working model, which displays the changes occurring in the subject area. Such a system must have a memory that allows it to store not only information about the current state of the subject area, but also, in some cases, the background.

Since the domain model supported by the information system materializes in the form of information resources organized in the necessary way, it is called an information model.

The above definition of M. R. Kogalovsky covers information systems of all types, in particular, fact-based systems that are based on database technologies and operate with structured data, text search systems that operate with documents in natural languages, the global hypermedia information system Web, etc.

In the textbook Information Systems, M. M. Telemtayeva considers IP as a complex and large system.

In relation to complex systems, it is based on the postulate of Academician A. I. Berg: "to create a model of a complex system, it is usually necessary to use more than two theories, more than two languages for describing the system, due to the qualitative difference in the internal nature of the system elements among themselves and the presence of different approaches to modeling objects of different nature"[11].

In relation to large systems, it is based on the definition given by V. I. Chernetsky: "a large system (BS) is a system that is a set of interconnected controlled subsystems united by a common control system, the characteristic feature of which is the presence of separate parts. Moreover, for each part, it is possible to determine: the purpose of functioning, subordinate to the general purpose of the entire system; the participation of people, machines and the natural environment in the system; the existence of internal material, energy and information connections between the parts of the system; as well as the presence of external links of the system under consideration with others" [12].

An information system is a complete set of parts that interact with each other and with the external environment of the IR. Parts of the IR are: database, information, information technologies, technical means, information products, models of internal and external user environments [13].

A system is a complete set of ways and / or means of ensuring the interaction of the internal environment of the elements (parts) of the system with the external environment of the system. The external environment of the system is usually structured from the point of view of the system in the form of sources of resources and consumers of the products of the system's activities;

Consistency is the integrity of an element (parts) of a system in relation to a given system. The element (parts) of the system is intended for activities in the interests of the system.

IR is a complete set of ways and means to ensure the interaction of the internal environment of the IR parts with the user of the IR-part of the external environment of the system that consumes the products of the IR activity.

An information system is an interconnected set of information, technical, software, mathematical, organizational, legal, ergonomic, linguistic, technological and other means, as well as personnel, designed to collect, process, store and issue information and make management decisions.

The information system consists of objects – elementary units of documents, and documents-information units. Many documents containing factual information, having the same physical structure and logical, informative purpose, form collections. Collections are characterized by their descriptions and descriptions of the structure of the documents that make up it [14].

A collection is a common form of organization of information resources that is determined by its parameters (style, attributes) and the structure of its documents and is a systematized collection of documents that are united by some criterion of belonging, for example, by content, purpose, access method, etc., provided with a meta description (metadata) in accordance with standards and data schemes. The document is characterized by its parameters (style, attributes) and the structure of the objects that it consists of. An object is defined by the type of data (according to the selected data schema) that it contains, and the description of the object's properties and methods [15].

Due to the fact that the information in the IR displays some entities of the real world (physical objects: objects, processes, phenomena, persons, publications, documents, algorithms, programs, files, facts, key terms, etc.), it is necessary to consider the IR as a set of information objects-data sets that represent (describe) Note that the development of the EB model should use ontological descriptions and conceptual models that summarize the accumulated experience in the field of creating and using EB [20]. A good overview of the existing conceptual models of EB is given in [22].

The ontological model of the IRIS EB is based on the conceptual models of the RM OAIS EB [23] and DELOS DLRM [24].

According to the DELOS conceptual model, an Information Resource (IR) is an abstract concept expressed by instances of one of its specializations. In particular, instances of the concept of IR are instances of an information object of any type (for example, documents, databases, collections, functions, etc.). Each resource in accordance with the DELOS model:

- has an identifier;
- organized according to the resource description. A resource can be complex and structured, because it, in turn, can consist of smaller resources and have connections to other resources;
- can be regulated by the functions that control its life cycle;
- expressed in terms of an information object;
- must be described with metadata, and can also be described or supplemented with additional metadata and annotations.

The implementation of the Electronic Library Management System is based on the meta-model, based on the fact that each information resource is characterized by a set of attributes inherent in it, and methods that characterize its properties and relationships with other resources. An effective means of describing information objects is metadata – data that is an integral part of an information object and describes a real object or group of objects.

Each information object in the IR consists of:

- Information content of the object (primary information object: for example, an image, full text, etc.) - an object that can be used independently;
- metadata object-an object whose main purpose is to provide information about the IR (usually about the primary information object);
- annotation object – an object whose main purpose is to annotate an IR or part of it. Examples of such annotations include notes, structured comments, and links. Annotation objects help to interpret the IR, contain either support, detailed explanations, or information about how the IR can be used.

II. Document Concept

One of the most important manifestations of human behavior is communication, i.e. communication with other people through certain signs or symbols. Initially, information about the surrounding world was transmitted by a person using gestures, facial expressions, shouting, touching, etc. - the simplest means of visual, auditory, tactile communication. The emergence of meaningful speech and language marked, according to a number of scientists, the emergence of the first information technology in the history of human society [16].

Meanwhile, with the development of man, the need to transmit information not only in space, but also in time, i.e., in the storage of information, increased. However, the simplest means of communication and information transmission were imperfect. The same human speech is heard only at a short distance and only at the moment of its utterance. It was difficult to retain the necessary information, since the knowledge was not yet separated from the subject who possessed it. It is no accident that at that time the role of a kind of knowledge banks and channels of their transmission was played by the elderly, i.e., the most experienced members of society.

The separation of information from the subject and the first attempts to consolidate it were associated with the use of signaling. To transmit information in space, signaling was used by smoke, fire fires, the sounds of trumpets, drumming, a branch or arrow placed in a certain way, etc. Objects were also used, which were given a symbolic meaning. The example of the symbolic message of the ancient Scythians to the Persians given by the ancient Greek historian Herodotus became a textbook. This message consisted of a bird, a frog, a mouse, and a bunch of arrows, and meant: "If you Persians do not learn to fly like birds, to jump through swamps like frogs, to hide in holes like mice, you will be showered with our arrows as soon as you enter the Scythian land."

Later, symbolic signaling was replaced by conditional signaling, in which objects were used as conditional signs by prior agreement of people about what a certain object would mean. As a result, there were systems of mnemonic signs for counting with the help of objects, as well as more complex "nodular writing": among the ancient Incas, in Ancient China, among the Mongols. Probably, this kind of "letter" was also available to the Slavs. It is no accident that the Russian language has preserved the expression "tie a knot for memory", i.e., to keep some information in memory. Tags (plaques) with notches were also used as conventional signs - in trade, financial, and creditor operations. Among the Slavs, such tags were called "noses", since they were usually carried with them, fixing any information with the help of various notches, notches. Hence, the expression “cut on the nose” is, remember it firmly [17].

Grave mounds, burial mounds, crosses, tombstones, property signs (heraldic signs, boundary stones, cattle brand marks, etc.) were used to consolidate and transmit information over time.

Objective ways of communication have been preserved to this day: the presentation of bread and salt as a sign of hospitality, bouquets of flowers and souvenirs as a sign of attention, military insignia, flags of states, traffic lights and semaphores, etc. The appearance of writing marked the transition of humanity to a new information technology. With the help of graphic sign systems, it became possible to separate information from the subject and fix it on some material for the purpose of subsequent transmission in time and space. As a result, there was documented information, i.e. a document.

The concept of "document" is currently the most common in the sciences that study different ways of storing and transmitting knowledge (or information) in society. There are many definitions of a document that have fixed ideas about it.

The term "document" comes from the Latin word "documentum". Documents appeared as an additional (to sound speech) means of communication of people. They were brought to life primarily by the need to capture, fix and transmit a particular message in time and space. Carriers, a material object on which information was recorded,

performed in ancient times mainly the function of evidence. Therefore, the Latin word "documentum" meant a sample, proof, testimony.

This term, in turn, came from the verb "docere" - to teach, to teach. The roots of this word go back to the Indo-European proto-language, where it meant a gesture of outstretched hands associated with receiving or receiving something. On the basis of this word, the words "doceo" were formed - I teach, I teach, "doctor" - a scientist, "doctrine" - teaching, and finally, "documentum" - what teaches an instructive example. In this sense, the word document was used, for example, by Caesar and Cicero.

Later, the word "document" acquired a legal meaning and came to mean "written evidence", "evidence drawn from books, supporting records, official acts". In the sense of a written certificate, the word "document" was used from the Middle Ages to the XIX century.

In the XIX century, a new aspect is highlighted: the importance of the document in management. For the designation of documents at this time, synonymous concepts were used: "business paper", "act", "case". The document was considered to be the information recorded in the form and intended for the implementation of the management process.

In the twentieth century, the term "business paper" is gradually replaced by «service document". In Soviet times, the term "document «was firmly established in normative acts, special literature, and work practice. It is still preserved at the present time.

Along with it, the concept of "act" is also used to refer to documents related to the field of management and law. They include almost all actions of the authorities and public administration, documented.

Over time, the term "historical documents" appears. They are considered chronicles, chronicles, notes, and other written sources that indicate a historical event, person, epoch, etc.

Thus, in the definition of the document, three aspects can be distinguished - legal, managerial, and historical.

The word "document" came to the Russian language in the time of Peter I, as a loan from the German and Polish languages, in the meaning of a written certificate [18]. At the beginning of the XX century, it had two meanings:

1) any paper drawn up in a lawful manner and can serve as proof of rights to something (property, fortune, free residence) or to perform any duties (conditions, contracts, debt obligations);

2) in general, any written evidence.

By the second half of the XIX century, the terms derived from the word "document" appeared in reference publications of some countries of the world: documentation-in the meaning of the preparation and use of documented evidence and authority; documentary - related to the document.

At the end of the XIX century, there is a tendency to narrow the boundaries of the concept of "document": first it was considered as any object that serves to obtain and prove, then - as a written certificate confirming certain legal relations. The concept was used mainly in the legal sense.

Since the beginning of the XX century, a new, broader understanding of the concept of "document" has been introduced into the term system: it was introduced by a well - known Belgian scientist, the founder of documentation - the science of the totality of documents and the field of practical activity-Paul Hautelet-in his treatise on Documentation, defines the concept (term) Document: "a material object containing information, specially designed for its transmission in space and time" [19] - which is interpreted as the main "object" with which any information system operates[20]. Thus, a document is an information object that represents a structured description of a real entity (object, subject, fact or concept), the totality of which makes up the information content of the system. The document presented in electronic form has a certain standard set of attributes and allows for unambiguous identification. A document can describe an article from a journal, the journal itself, a person, a digitized image, experimental data, a program or computational algorithm, a database, a fragment of a database, etc.

P. Otle first used a comprehensive approach to the typological classification of documents, taking into account the content and form of the document, in the "Treatise on Documentation".

The scientist divided the entire set of documents into three main classes:

1) Bibliographic documents, i.e., texts that are traditionally considered works of writing and printing. Among them are brochures, monographs, essays, treatises, manuals, encyclopedias, dictionaries, periodicals and continuing publications (magazines, newspapers, yearbooks, etc.). In addition to these, bibliographic documents included texts of personal origin (letters), official messages and accounting (registration) books (or magazines), as well as signs, slogans, tickets and other travel documents.

2) Other graphic documents, i.e., non-text documents: cartographic, pictorial, musical notation. Among the pictorial ones are: iconographic, containing a printed image (prints, engravings, postcards, etc.); photographs; documents perceived through projection devices (including microcopies). As a special variety, "pictorial monuments are distinguished: inscriptions, coins, medals, seals (stamps).

3) Documents - substitutes for books: discs, phonograms, films, and along with this-radio broadcasting (recording and transmitting sound), television, including telephotography, radio telephotography and television itself.

A special place in this classification series was occupied by" documents of three dimensions": natural (minerals, plants, animals) and artificial, created by man (materials, products, technical objects, as well as medals, models, reliefs). They also include scientific tools, didactic materials, and visual aids. Especially highlighted among them are three-dimensional works of art: works of architecture and sculpture.

It is in this broad sense that the concept of "document" was later used when it comes not only to the collections of libraries, archives, museums, information services, but also to social, in particular, mass communications in general.

We can distinguish the following values of "document", introduced by P. Otle:

1) Any source of information, transmission of human thought, knowledge, regardless of whether it is embodied in a material-fixed form or is a conductor (transmitter) of information in time, can be considered a document. This concept covers material objects-information carriers, as well as, radio, television, and theatrical performances.

2) Documents are material objects with recorded information collected by a person to create any collections. This includes both artificial objects created by man, and natural, technical objects located in the museum.

3) Documents also include material objects created by a person specifically for recording, storing and reproducing information in order to transmit it in space and time, regardless of the method of recording. These are both "written" documents (i.e., with information recorded by writing characters), and visual, phonographic recordings and films (the results of machine recording of images and sound).

The author of the" Treatise... "repeatedly emphasizes the synonymy of the concepts of" document "and" book"; from the context, it can be understood that he considers the former as broader.

Thus, P. Otle entered the world history of documentary studies as the founder of documentation-science and practice. He was the first not only to introduce the basic concept of "document" into scientific use, but also to reveal its broadest meaning. P. Otle made the first attempt at a comprehensive classification of documents by a set of features. Although it had significant drawbacks, the author managed to group the existing variety of information sources that function in social communication. Subsequently, the theoretical thoughts of domestic and foreign specialists moved in the same direction as the thought of the documentarian P. Otle.

The concept of P. Otle considers the document as a carrier of social information. However, in reference publications of that time, there continues to be a narrow meaning of this word: in addition to the legal one, the concept of "historical document" (a fixed certificate of an era, person, etc.) and "accounting document" (serving as the basis for carrying out economic actions-receiving and issuing valuables) is introduced.

Paul Marie Ghislain Otlet is a Belgian writer, entrepreneur, thinker, bibliographer, lawyer, and peace activist. He was one of those who are considered the founding fathers of computer science. He is the author of numerous publications on the problems of book studies, bibliography and documentation. He was an active promoter of international cooperation in the field of book studies.

In 1934, he published his famous book "Trait de documentation", which laid the foundation for the science of documents (in the broad sense), which can now be called "documentation" - the forerunner of modern computer science, the author of the concept of the information universe (electronic, but not digital), as the development of telephone communication and television. They are also the developers of the Universal Decimal Classification (UDC) - one of the most outstanding examples of faceted classification.

P. Otle repeatedly emphasizes the synonymy of the concepts of "document "and" book"; from the context, it can be understood that he considers the former as broader. He was the first not only to introduce the basic concept of "document" into scientific use, but also to reveal its broadest meaning. P. Otle made the first attempt at a comprehensive classification of documents by a set of features. The concept of P. Otle considers the document as a carrier of social information.

Consider a number of other document definitions:

A follower of P. Hautelet, Suzanne Brie (1894-1989), in 1951 published a work entitled "What is documentation?", which begins with the following statement: "A document is a certificate confirming a certain fact", "it is any physical or symbolic sign, preserved or recorded, intended to represent, reconstruct, confirm some physical or individual phenomenon". Suzanne Brieux equated the document with an organized physical observable object. This approach resembles the definition of «material culture". in terms of cultural anthropology and the ideological (more precisely, methodological) approach of "object as a sign" in semiotics. Also in the 1951 Manifesto on the Nature of Documentation, it stated: "A document is evidence in support of a fact ... it is any physical or symbolic sign, stored or recorded, intended to represent, recreate, or demonstrate a physical or conceptual phenomenon. " [21].

According to S. Brie, the wild antelope is not a document. But if it is placed in a zoo cage and studied, it becomes a physical object, a primary document, and all the articles about it are secondary, derived documents. (Here we are talking about the antelope, which is a new, newly discovered species of African antelope).

The definition of a document is proposed by Yu. N. Stolyarov "Document - semantic information created by a person specifically for social communication and recorded in any way on any medium" [22]. Where "fixing information" defines as an essential characteristic of a document: "A document is an object that allows you to get the required information from it" [23]. He also states, "The document status can have any object", "the same object may or may not be a document - it all depends on whether it serves to get information from it (from it, from it) or not." Any objects of reality can provide information, but not all of them become documents. For example, "Madame Brieux's antelope" can become a document if it is recorded in the form of an image (photo, drawing, sculpture, etc.) or if the words "Madame Brieux's antelope" are recorded, and the real antelope, even if it has such an original name, is not a document. That is, "A document is an object that allows you to get the required information from it."

A. V. Sokolov's definition: "A document is a stable material object intended for use in social semantic communication as a completed message". Where the author draws attention to the distinctive features of the document: the presence of semantic content, stable material form, intended for use in communication channels, the completeness of the message. Note that the presence of semantic communication is the essence of any information process. The sign of the stability of existence is a mandatory characteristic of the document as a way of storing information.

III. Models of Distributed Information Systems to Support Scientific and Educational Activities

The rapid development of global information and computing networks leads to a change in the fundamental paradigms of data processing, which can be described as a transition to the support and development of distributed information resources [21]. Therefore, the most important task associated with the technology of working with information is to study ways to integrate distributed data sources.

Integration of information resources is understood as combining them in order to use (using convenient and unified user interfaces – preferably one) different information while preserving its properties, presentation features, and user manipulation capabilities. However, the pooling of resources does not have to be physically performed. The main thing is that it should provide the user with the perception of the available information as a single information space. In particular, electronic libraries should provide work with heterogeneous databases or database systems, providing the user with the effectiveness of information searches, regardless of the features of the specific information systems to which access is made.

An urgent task is to create a model of a distributed information system to support scientific and educational activities:

- to unify the process of sharing the results of scientific research;
- operate with data and documents integrated into an open.
- semantic space;
- provide services for the transformation of heterogeneous resources that implement the means of description, representation, automatic linking of resources, as well as interaction with search and classification mechanisms in accordance with the needs of users.

The model should provide the following functionality:

- publishing resources, including registration, naming, annotation, and format definition procedures;
- analytical processing of resources;
- access to published resources, including dynamic generation functions;
- for automated operation, you need a function for monitoring resources and updating their meta descriptions, functions for notifying users about the appearance of new resources and updating existing ones, and a dispatching function [24].

IV. Metadata Profile in Distributed Information Systems to Support Scientific and Educational Activities

An effective means of describing information objects is metadata, which is an integral part of an information object and describes a real object or group of objects.

An important property of metadata is its specificity with respect to the scope of the described objects (resources). Metadata can characterize entities that relate to both the virtual (information) space and the real world (persons, organizations, events). Metadata can be part of information resources, or it can be stored separately from information resources.

Metadata is necessary for solving the following tasks:

1) providing information about documents, their content, structure, methods of use, etc.;
2) systematization and classification of documents;
3) organization of in-system processing procedures;
4) Support for sharing with external IS.

Standards application profiles are created for a specific group of functional tasks or users. This makes it easier to create systems that work with metadata. A profile can be defined as "one or a combination of several basic standards with the identification of the selected classes, subsets, optional capabilities, and parameters of these basic standards required to perform a specific function" [22].

In the metadata area of resources for publications, the profile should contain a list of mandatory elements present in the resource description, and set dictionaries for describing the values of elements that complement or extend the acceptable set of values defined in the standard. In addition, additional description elements may be suggested.

Thus, the basis for the development of a scientific and educational system consists of standards and international recommendations that form the profile of a scientific and educational system, which is understood as one or a set of several basic normative and technical documents (standards and specifications) aimed at solving a specific task (the implementation of a given function or a group of functions of an application or environment), indicating, if necessary, the selected classes, subsets, options of basic standards necessary to perform a specific function [27]. The most important is the metadata profile of the information circulating in the system.

V. Requirements for Distributed Information Systems for Scientific and Educational Activities

Distributed information resources the development of the organization's information resources leads to the need to create an infrastructure for their integration into a single

information system that provides transparent access to distributed information.

The development of global information and computing networks today leads to a change in the fundamental paradigms of working with information resources. Today, the transition to distributed resources, the creation of infrastructure for and integration into a single information system that provides transparent access to distributed information is relevant.

Therefore, the most important task related to information technology is to research ways to integrate distributed data sources and create scientific groundwork in the field of distributed information systems and databases in order to develop technology that supports the creation and operation of large-scale information infrastructures based on virtual integration. This technology will allow you to create global infrastructures from dozens and hundreds of heterogeneous databases and solve strategic tasks in the field of automation of various forms of distributed activities. A narrower goal is to develop principles and software tools for virtual integration of distributed data sources based on international standards and recommendations for creating large-scale information infrastructures designed to virtualize data access to various DBMS using common rules and policies [23].

The tasks of distributed, as well as conventional, information systems are to store information and provide it to users in a convenient form. As a rule, such systems can be organized on the basis of various technological solutions aimed at implementing a particular distribution paradigm. Based on the main functions of information systems, various aspects of distribution can be considered:

1) Distributed information storage (distributed storage, network storage systems, and network file systems).
2) Distributed DBMS (adding, upgrading, changing data).
3) Distributed information access management and distributed information management.
4) Search for information in distributed sources.
5) Extracting information from distributed sources.
6) Visualization of information from distributed (heterogeneous) sources in unified user interfaces [24].

Distributed information systems represent an increasingly important trend for computer users. Distributed processing is a method for implementing a single logical set of processing functions on multiple physical devices, so that each performs some part of the overall required processing. Distributed processing is often accompanied by the formation of a distributed database. A distributed database exists when data items stored in multiple locations are interconnected, or if a process (program execution) in one location requires access to data stored elsewhere.

Distributed information systems to support scientific and educational activities is designed to collect, classify, analyze text publications of the Kazakhstan segment of electronic mass media for the management of information resources.

The purpose of creating the system: To develop a system for distributed information systems to support scientific and educational activities, to create a program to explore the capabilities of the Apache Solr platform for processing distributed data that uses big data technologies.

A set of the most general functional requirements for the IP support of scientific and educational activities was identified:

1) Collection of information resources.
2) Relevance of documents.
3) The relevance, completeness, and authenticity of the origin of the documents.
4) Use of intelligent services for processing user requests.
5) Knowledge extraction.
6) Support for non-centralized information system architectures.
7) Structuring of the information space.
8) Adaptive presentation of information.
9) Historicity of information.
10) Archive.

In the conditions of working in a distributed environment, the requirements for the IP support of scientific and educational activities are:

- support for the adopted metadata standards for data export and import;
- support for information exchange protocols with other information systems;
- support for the ability to link to internal resources both in user interfaces and at the system level.

System tasks:

1) Collection, storage and selection of unique publications from the Internet space to the system database.
2) Distribution of publications by topic: clustering, classification, definition of thematic combinations, ranking and filtering (by social spheres, regions, industries, etc.)
3) Definition of information occasions.
4) Calculation of the degrees of informative features of the publication, such as: collective use of purchased electronic literature catalogs, databases and bibliographic publications.
5) Definition of information trends.

VI. Architecture of Distributed Information Systems to Support Scientific and Educational Activities

Our generation has a rich scientific heritage that should be preserved.

Delay in this work can lead to irreparable losses associated with a temporary factor: the loss of documents, the departure from the lives of eyewitnesses to the events. One of the most important tasks related to the preservation of scientific heritage is a set of measures aimed at creating specialized information systems (electronic libraries) designed to store information, to organize access and mechanisms for using information [25].

Two necessary requirements that can be imposed on such information systems are obvious. The first requirement is the need to create and provide a system for reliable long-term storage of digital (electronic) documents while preserving all the semantic and functional characteristics of the original documents. The second is to provide a "transparent" search and access to users ' documents, both for review and for analysis and scientific work [23].

In the existing developments of electronic libraries, as a rule, the search and access to information is provided only through visual graphical interfaces. This is good for the human user, but very bad for the application user (for example, for conducting various analytical studies).

To provide search functions outside of graphical interfaces, support for special network services and query languages is required. Ideally, all information systems should support a single search profile and a single query language. The implementation of the abstract search paradigm today exists in the form of several models for organizing search services, for example, the Z39.50 model [5; 6] and the simpler SRW/SRU model [6]. The practical implementation of services such as SRW / SRU provides a significantly new quality of the electronic library – the ability to include its resources in global search engines at a higher level than the level of external indexing of static Web pages by other systems. Other possible search types are related to search by specified templates and to search using ontology. The search involving ontology is more intelligent. Its implementation requires additional information about the domain, including definitions of terms, entities, and relationships. It should be noted that the presentation of this additional information must comply with global agreements and international standards, otherwise the search using dictionaries, thesauruses and ontologies will always be limited to the current system, and interoperability will not be implemented [26].

As part of the tasks set, an information system architecture was developed (Fig. 1) to systematize the resources of the electronic library, a multi-level EC architecture is used, consisting of a data warehouse, a repository, a metadata server, an application server, a dictionary, reference books, as well as a software implementation of the developed architecture, deployed on existing hardware and put into operation.

Based on the formulated requirements, the information system for supporting research on scientific and educational activities should consist of a long-term storage system and an information management system for organizing an abstract search necessary for the analysis and conduct of scientific works. A very important component of the technology of working with scientific heritage is metadata, which contains the information necessary to document the process of storing information resources. This metadata is information about the format, structure, and use of information resources, the history of all operations, including any changes, authenticity, technical history, responsibilities, rights, and so on.

Thus, the information system on scientific heritage should functionally consist of three blocks.

1) Digital Depository 6 (or repository, hereinafter referred to as DD) is an independent system of long-term storage and access to heterogeneous digital objects, which is designed to provide electronic (digital) versions of documents on the scientific heritage (books, scientific articles, reprints, letters, images and other materials presented in electronic form).

2) Reference books are a set of databases containing information about authors and other persons (authoritative files), geographical locations, cities, publishing houses related to a particular scientific school, thematic dictionaries-classifiers, thesauri, descriptions of the subject area of this scientific school and classifiers of documents in accordance with the IECOF.

3) The metadata server should provide metadata management-cataloging of all information resources in accordance with generally accepted international standards. It should run a whole set of application services that should: support abstract search schemes in accordance with the schemes proposed by the Z39.50 protocol and SRW/SRU, support search schemes based on specified templates and using ontology, support fact detection and document identification based on information that is in the directory, as well as provide metadata collection from its own and remote data centers (exchange, synchronization and modification), metadata conversion between existing standards (GOST, MARC, etc. and the corresponding translation of metadata schemas from one format to another.

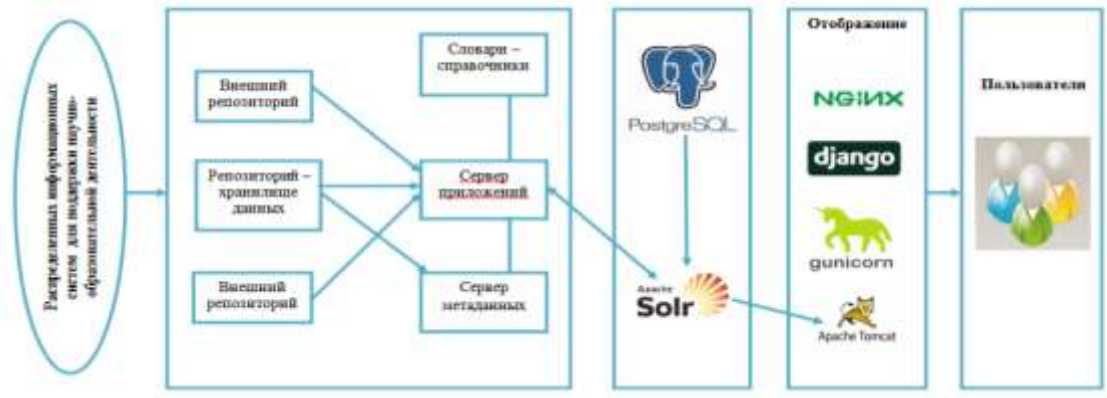


Fig. 1. Architecture of Distributed Information Systems to Support Scientific and Educational Activities.

Accurate categorization of the material using the dictionary of the reference book increases the probability that the search results will find documents relevant to the search expression when organizing a search in one or more electronic libraries [27].

4) PostgreSQL-acts as a persistent storage for structured data. The main types of data stored in this database:

a) News and metadata;

b) Processed data at the level of different basic units of analysis (token/word/phrase/sentence/text), including vectorization, lemmatization results, cleaning, etc.

c) Results of thematic modelling;

Results of news classification by various criteria (tone, politicization, social significance, etc.).

5) Apache Solr is a popular, fast-growing open-source search platform built on Apache Lucene. Solr is highly reliable, scalable, and fault tolerant, providing distributed indexing, replication, and load balancing, automatic disaster recovery, centralized configuration, and more. Solr supports the search and navigation functions of many of the largest Internet sites in the world. Since Solr has distributed search and replication capabilities, Solr is highly scalable. Here are some of the main features that solr provides:

a) Advanced full-text search capabilities

b) Optimized for high volume traffic

c) Open interfaces based on standards-XML, JSON and HTTP

d) Comprehensive administration interfaces.

e) Easy monitoring.

f) High scalability and fault tolerance.

g) Flexible and adaptable with simple configuration.

h) Next to the real-time index.

i) Extensible plugin architecture.

Data Processing:

When developing the architecture for data processing, the following main needs were identified:

1) The ability to parallelize calculations, including on multiple machines;

2) Flexible scheduling of various data processing tasks;

3) The ability to monitor the execution of tasks in real time, including prompt notification of exceptions;

4) Flexibility in the tools and technologies used.

Distributed information systems will contain the following subsystems:

- Subsystem-a repository of digital objects that provides user and administrative WEB - interfaces for accessing digital objects and collections, as well as interfaces for integration with other subsystems based on open international standards.
- A subsystem for managing current research information (SUEB), which includes information about the publications of employees, their participation in conferences and in the implementation of research projects.
- The subsystem will include user and administrative interfaces, as well as interfaces for integration with other subsystems based on open international standards.
- Subsystem for integration of distributed information resources based on Apache Solr technologies.
- Subsystem for access to distributed information resources based on technologies-Nginx, Djang, Apache Tomcat.

These subsystems together should provide:

- identification of information resources;
- identification, authentication and authorization of users;
- metadata management;
- information resource management;
- collecting statistics;
- monitoring the availability of services and resources.

In distributed search, a collection is a logical index on multiple servers. The part of each server that runs the collection is called the core. Thus, in an unallocated search, the core and the collection are the same, since there is only one server.

VII. Conclusion

Based on the formulated requirements, a prototype of an information system has been developed that can be used as a standard for working with documents in the field of scientific and educational activities, since it solves the main tasks facing these systems: ensuring reliable long-term storage of digital (electronic) documents while preserving all the semantic and functional characteristics of the source documents; ensuring "transparent" search and user access to documents for both familiarization and analysis of the facts contained in them; organizing the collection of information on remote digital repositories.

The developed model of the information system can be used as a standard model of the system for working with documents related to scientific heritage, since it solves the main tasks required for these systems:

- providing a system of reliable long-term storage of digital (electronic) documents with the preservation of all semantic and functional characteristics of the original documents;
- providing a "transparent" search and access to users ' documents both for review and for analysis and scientific work.

To date, all the components necessary to create a qualitatively new scientific information system are available and worked out clearly and systematically. Most of the scientific-centralized distributed systems allow you to create a single environment for the exchange of scientific information.

ACKNOWLEDGMENT

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP09057872).

REFERENCES

[1] Zhizhimov O. L., Fedotov A.M. Ensuring the interoperability of electronic libraries / / Information technologies and mathematical modeling in science, technology and education (Bishkek, Kyrgyzstan, 5-9 October 2011): Izv. Kyrgyz State Technical University. Razzakov State University. 2011. No. 24. Materials of the international conference. pp. 331-335.

[2] Larkov N. S. Documentovedenie: Uchebnoe posobie / N. S. Larkov. - M.: AST: Vostok-Zapad, 2006.

[3] Larkov N. S. Documentovedenie: Electronic textbook. Tomsk: TSU 2002.

[4] Otle P. Biblioteka, bibliografiya, dokumentatsiya: Izbrannye trudy pionera informatiki [Library, bibliography, documentation: Selected Works of the pioneer of Informatics]. - Moscow: FAIR-PRESS: Pashkov House, 2004. - 348, [1] p.- (Special publishing project for libraries). - Bibliogr.: pp. 312-327. - Names. decree: pp. 340-342. - ISBN 5-8183-06.

[5] Fedotov A. M. Metodologii stroeniya razdelennykh sistem [Methodologies for constructing distributed systems]. 2006, vol. 11, Selected reports of the X Russian Conference. Distributed information and computing resources. (DICR-2005), Novosibirsk, October 6-8, 2005, pp. 3-16.

[6] Fedotov A.M., Zhizhimov O. L., Fedotova O. A., Barakhnin V. B. Model of information system for support of scientific and pedagogical activity // Vestn. Novosibirsk State University. Ser. Inform. technologies. 2014. Vol. 12, no. 1. pp. 89-101.

[7] Fedotov A.M., Barakhnin V. B., Zhizhimov O. L., Fedotova O. A. Technology of creating corporate information systems for accounting of scientific workers ' works. Novosibirsk State University. Ser. Inform. technologies. 2011. Vol. 9, issue. 2. pp. 31-41.

[8] Zhizhimov O. L., Fedotov A.M., Fedotova O. A. Building a typical model of an information system for working with documents on scientific heritage. Novosibirsk State University. Ser. Inform. technologies. 2012. Vol. 10, no. 3. pp. 5-14.

[9] Kogalovsky M. R. Metadata, their properties, functions, classification and means of representation // Proceedings of the 14th All-Russian Scientific Conference " Electronic Libraries: Promising Methods and Technologies, Electronic Collections — - RCDL2012, Pereslavl-Zalessky, Russia, October 15-18, 2012.

[10] Kogalovsky M. R. Metadata in computer systems/M. R. Kogalovsky / / Programming, 2013, N # 4. - p. 28-46.

[11] Kogalovsky M. R. Scientific collections of information resources in electronic libraries. Proceedings of the First All-Russian Scientific Conference "Electronic Libraries: Promising Methods and Technologies, Electronic Collections", St. Petersburg, October 1999. St. Petersburg University Press, 1999.

[12] Bezdushny A. N., Bezdushny A. A., Serebryakov V. A., Filippov V. I. Integration of metadata of the Unified Scientific Information Space of the Russian Academy of Sciences. Moscow: Raschut. A. A. Dorodnitsyn Center of the Russian Academy of Sciences, 2006. 258 p.

[13] Functional requirements for bibliographic records: conceptual model: graduate. report / translated from English by V. V. Arefyev. Moscow: Russian State Library,2006. 150 p. Shokin Yu. I., Fedotov A.M., Barakhnin V. B. Problems of information search. Novosibirsk: Nauka, 2010. 198 p.

[14] Kulagin M. V., Lopatenko A. S. Scientific information systems and electronic libraries. The need for integration. // Collection of proceedings of the Third All-Ross. conf. on electronic.libraries. RCDL ' 2001 Petrozavodsk, September 11-13, 2001, pp. 14-19.

[15] Fedotov A. M. Metodologii stroeniya razdelennykh sistem [Methodologies for constructing distributed systems]. 2006, vol. 11, Selected reports of the X Russian Conference. Distributed information and computing resources. (DICR-2005), Novosibirsk, October 6-8, 2005, pp. 3-16.

[16] Fedotov A.M., Zhizhimov O. L., Fedotova O. A., Barakhnin V. B. Model of information system for support of scientific and pedagogical activity // Vestn. Novosibirsk State University. Ser. Inform. technologies. 2014. Vol. 12, no. 1. pp. 89-101.

[17] Fedotov A.M., Barakhnin V. B., Zhizhimov O. L., Fedotova O. A. Technology of creating corporate information systems for accounting of scientific workers ' works. Novosibirsk State University. Ser. Inform. technologies. 2011. Vol. 9, issue. 2. pp. 31-41.

[18] Zhizhimov O. L., Fedotov A.M., Fedotova O. A. Building a typical model of an information system for working with documents on scientific heritage. Novosibirsk State University. Ser. Inform. technologies. 2012. Vol. 10, no. 3. pp. 5-14.

[19] Kogalovsky M. R. Metadata, their properties, functions, classification and means of representation // Proceedings of the 14th All-Russian Scientific Conference " Electronic Libraries: Promising Methods and Technologies, Electronic Collections — - RCDL2012, Pereslavl-Zalessky, Russia, October 15-18, 2012.

[20] Bezdushny A. N., Bezdushny A. A., Serebryakov V. A., Filippov V. I. Integration of metadata of the Unified Scientific Information Space of the Russian Academy of Sciences. Moscow: Raschut. A. A. Dorodnitsyn Center of the Russian Academy of Sciences, 2006. 258 p.

[21] Functional requirements for bibliographic records: conceptual model: graduate. report / translated from English by V. V. Arefyev. Moscow: Russian State Library,2006. 150 p. Shokin Yu. I., Fedotov A.M., Barakhnin V. B. Problems of information search. Novosibirsk: Nauka, 2010. 198 p.

[22] Kulagin M. V., Lopatenko A. S. Scientific information systems and electronic libraries. The need for integration. // Collection of proceedings of the Third All-Ross. Conf. on electronic.libraries. RCDL ' 2001 Petrozavodsk, September 11-13, 2001, pp. 14-19.

[23] S.K.Serikbayeva , D.A.Tussupov, M.A.Sambetbayeva, A.S. Yerimbetova, Taszhurekova ZH.K., Borankulova G.S., EduDIS construction technology based on Z39.50 protocol / / Journal of Theoretical and Applied Information Technology. - 2021. - Vol.99(10), -pp. 2244–2255.

[24] Serikbayeva S.K, Batyrhanov A.G., Sambetbayeva M.A., Sadirmekova Zh.B., Yerimbetova A.S. Development of technology to support large information storage and organization of reduced user access to this information: (IJACSA) International Journal of Advanced Computer Science and Applications, 2021 г.. - 7 : Т. 12. - стр. 493-503.

[25] A.M. Fedotov, I.A. Idrisova, M.A. Sambetbaeva, O.A. Fedotova The use of the thesaurus in the scientific and educational information system // Bulletin of the Novosibirsk State University. Series: Information Technology. - 2015. - T.13. - No. 2. - P.86-102. - ISSN 1818-7900. - EISSN 2410-0420.

[26] Sambetbayeva M.A., Fedotova O.A., Fedotov A.M. Multilingual Thesaurus in Information System for Scientific and Educational Activity Support // Proceedings of the XX International Conference “Data Analytics and Management in Data Intensive Domains” (DAMDID/RCDL’2018). – M.: Lomonosov Moscow State University, 2018. – pp. 360-362.

[27] Fedotov A.M., Tusupov J.A., Sambetbayeva M.A., Sagnayeva S.K., Bapanov A.A., Nurgulzhanova A.N., Yerimbetova A.S. Using the thesaurus to develop it inquiry systems // Journal of Theoretical and Applied Information Technology. - 2016. - Vol.86. - Issue 1. - pp.44-61.

Detection of Intruder in Cloud Computing Environment using Swarm Inspired based Neural Network

Nishika[1]
Ph.D. Research Scholar
UIET, Maharshi Dayanand University
Rohtak, Haryana, India

Kamna Solanki[2]
Assistant Professor
UIET, Maharshi Dayanand University
Rohtak, Haryana, India

Sandeep Dalal[3]
Assistant Professor
DCSA, Maharshi Dayanand University, Rohtak, Haryana, India

***Abstract*—Cloud computing services offered a resource pool with a wide range of storage for large amounts of data. Cloud services are generally used as a demand-driven private or open data forum, and the increase in use has led to security concerns. Therefore, it is necessary to design an accurate Intrusion Detection System (IDS) to identify the suspected node in the cloud computing environment. This is possible by monitoring network traffic so that the quality of service and performance of the system can be maintained. Several researchers have worked on designing valid IDS with the help of a machine learning approach. A single classification algorithm seems to be impossible to detect intruders with high accuracy. Therefore, a hybrid approach is presented. This approach is a combination of Cuckoo Search. CS as an optimization algorithm and Feed Forward Back Propagation Neural Network (FFBPNN) as a multi-class classification approach. The user's request to access cloud data is collected and essential features are selected using CS as an optimization approach. The selected features are used to train FFBPNN with reduced training time and complexity. The experimental analysis has been performed in terms of precision, recall, F-measure, and accuracy. The evaluated value for parameters i.e., precision (85.5%), recall (86.4%), F-measure (85.9%), and accuracy (86.22%) are observed. At last, the parameters are also compared with the existing approach.**

Keywords*—Cloud computing; intrusion detection system; cuckoo search; feed forward back propagation neural network (FFBPNN)

I. Introduction

In this modern era, cloud computing has transformed the IT world with rapidly evolving and extensively accepted computing-based systems. The attractive features of Cloud Computing continue to increase integration in many sectors, such as governments, private, including industry, education, and entertainment [1]. According to the National Institute of Standards and Technology (NIST), cloud computing is defined as the computational model, which delivered services on-demand [2]. Cloud Computing provides a variety of applications and services to customers or users on the Internet. Services are provided remotely from various servers or the cloud, which is far from the users. Cloud Computing allows the user to use different software types in the cloud without installing the user system. Currently, there is a growing demand for clouds due to these devices, which leads to the need to take security measures because, with the increase in the demand, more security is required against threats [3]. There are already many businesses that use cloud computing services with their attractive features such as on-demand services, extensive network access, fast flexibility, and, finally, measurable services. Such features will allow users to focus on business processes while managing computing resources through a cloud service provider (CSP). Using cloud features, the operating costs are reduced by ensuring the compatibility and availability of different computing sources, simplifying device installation, and process with software and hardware updates [4]. There are several service provider models such as the private model, public model, community model, and hybrid model in the market. The cloud model that offers services to their individual or specific users is termed a public cloud. By using this cloud model, the services are delivered to the general public, managed, and controlled by private or government, or semi-government agencies. The private cloud infrastructure is extensive and provided for use by a single organization. The community cloud model is also presented to be used by a defined community. Here, community means a group of organizations with similar interests. A new cloud infrastructure is a hybrid cloud that consists of the right combination of different infrastructures, which can be individual (private), public, or community [5].

The organization, as well as the security of these cloud models, needs to be improved so that the stored data by many cloud users remain safe. This is possible through the utilization of the Intrusion Detection System (IDS). A suspicious entry in the network is known as an intrusion [6]. Therefore, it is necessary to design efficient IDS that can protect the stored data against suspicious users. IDS can be divided into two types, one is Host-based IDS (H-IDS), and the other is Network-based IDS (N-IDS) [7]. The first H-IDS intrusion detection program was developed using the original target system as the primary host computer, where some external interactions are often absent. HIDS will operate based on information collected using a personal computer system. It monitors all incoming and outgoing packets on the computer system and notifies users or the administrator if it is observed that there is a suspicious activity. This can be used commonly to protect personal information, which is valuable for several server-based systems [8].

NIDS will capture all network traffic and be further analyzed to detect all possible intrusions, such as port scans or sometimes Denial of Service (DoS) attacks. NIDS usually performs this detection by effectively processing IP and transport layer headers for all collected network packets simultaneously [9]. Network packets are collected in the presence of an anomaly, and a link will be obtained along with signatures for numerous notorious attacks, and this is used to compare user behaviors with their known profiles. Many hosts are working within a network that is protected from the attacker using the NIDS model. If it is to be run, one must come to know the location of the NIDS that is usually hidden [10]. An IDS is required to classify the malicious nodes in the deployed network. In the presented paper, the author has proposed an improved CS-FFBPNN based IDS system to protect the network from intruders. The authors considered to remove all inappropriate information from training data to increase classification accuracy and provide efficient IDS.

The major purpose of the proposed work is to provide a suitable intrusion detection system (IDS) to detect the suspected node based on the cloud environment. The researcher focused on the two types of attacks detection that are DDoS and Benign.

A. Contribution of the Work

The major contributions of the paper are discussed below:

- Identification of the suspected node in a cloud environment by designing an appropriate Intrusion detection system.
- Apply cuckoo search optimization algorithm for feature selection.
- Apply FFBPNN after feature selection to train the data.
- Design a hybrid approach by integrating cuckoo search algorithm and Feed Forward Back Propagation Neural Network.

II. Related Work

Cloud Computing techniques have become more sensitive as the services are provided in different parts of the world. Hashizume et al. (2013) have mentioned all possible attacks that appear in the cloud environment. They also tried to make recommendations to avoid risks and vulnerabilities [11]. Ghosh et al. (2015) have presented an efficient and fast IDS system for cloud networks combined with N-IDS and H-IDS. Using this approach, the collected network packets are analyzed, and then acknowledgment has been sent to the cloud administrator. K-nearest neighbor (K-NN) and neural networks have been used in combination to train and test the performance of the NSL-KDD dataset. The CSP has created a list of malicious IP addresses and then store that list for future use. The model has been designed to handle large data flow and generate reports accordingly [12]. Pandeeswari and Kumar (2016) have proposed an anomaly detection-based system by integrating the fuzzy c-means clustering algorithm to the ANN approach and named the designed method FCM-ANN. The performance of the intended approach has been compared with the Naïve Bayes approach and simple ANN approach using DARPA's KDD cup dataset 1999. Based on the experiments, it has been concluded that the designed FCM-ANN approach performed well with a high detection rate and a low false alarm rate [13]. Baig et al. (2017) have presented a multi-class ID system combined with the ensemble-based ANN approach to monitor the network traffic of the computer system. The designed system learned the behavior of suspected and regular users by cascading many NN (Neural Networks), and each network is trained using a small training dataset [14].

The small samples of training data have been prepared using a filter, and the system is trained using ANN with a boosting-based learning approach. The designed network has been performed better by comparing the results with the KDD CUP 199 dataset as well as with the UNSW-NB15 dataset. Mahajan et al. (2017) have presented a signature-based IDS approach to detect the anomaly behavior of data based on the traffic data flow. Results show that about 20 to 25 percent of the load has been increased on the CPU in contrast to usual traffic. Also, the CPU load during the virtual networking scenario of about 30% has been observed compared to the general networking case. It has also been concluded that the virtual network is failed to handle high-speed traffic [15].

Deshpande et al. (2018) have presented a host-based IDS model for the cloud environment. The designed model has alerted the cloud user about the suspected activities by observing the system call traces. The model has traced only the system call traces and the failed call traces instead of discovering all the system calls. The designed approach reduced the CPU burden to a great extent using security features and provided cloud security with an average accuracy of about 96% [16].

Hajimirzaei and Navimipour, (2019) have presented a novel IDS system that is cascaded with a multi-layer perceptron (MLP) network, Fuzzy clustering approach, and Artificial Bee Colony (ABC) as an optimization approach. The MLP was trained using the optimized values, which were later used to adjust the weight and bias of the MLP network. The performance has been analyzed based on different parameters to test the performance of the proposed system using the NSL-KDD dataset. Artificial Intelligence (AI) has already been included in this study to detect any interference in a particular cloud environment, and as a result of the application of AI techniques, this self-regulating IDA has been tested using real-time data with high network speed. This algorithm has been used to set up a highly reliable private cloud for the military and banking sector to monitor network activity [17].

Kumari et al. (2017) have presented Particle Swarm Optimization (PSO) in addition to the Genetic Algorithm (GA) as an evolutionary approach to secure data in the cloud network. Cloudsim toolkit has been used to evaluate the performance of the proposed work. The model presented improved results with reduced computation cost [18].

Manickam and Rajagopalan, (2019) have introduced a hybrid structure to prevent cloud networks using Glow Swarm Optimization (GSO) in integration to the Tabu Search (TS) approach. These techniques have been used to minimize

convergence time and resolve the problem of local optima [19].

Velliangiri *et al*. (2020) have focused on detecting DDoS attacks using a deep learning classifier. The collected log information from the number of users has been collected and retained in the log file. An appropriate feature from the log file has been gathered using Bhattacharya distance and then used to train the system using a deep network. Also, to obtained exact and proper features, the Taylor-Elephant Herd optimization-based Deep Belief Network (TEHO-DBN) has been modified using "Taylor Elephant Herd" as an optimization approach. The trained network is then used to detect DDoS attacks with an accuracy of 83%.

The adversary in the node accesses secret information, which may lead to eavesdropping or DDoS attacks. The presence of attack in the cloud environment may cause system overload and dumps packet that results in information losses. The primary attribute that is affected by the DDoS attack is IP address spoofing. The DDoS attack is an intruder that dumps the packets and exhausts the resources of an individual by sending huge data traffic towards the legitimate user. Therefore, it is essential to design a protection system that can detect the intruder and increase the system's performance [20].

III. Methodology Used

The Cuckoo Search (CS) technique as optimization and FFBPNN as a multi-class algorithm are used in this research. A detailed description of both methods is provided below.

A. Cuckoo Search

The brooding parasitism of cuckoo species inspires the CS algorithm. CS is a Swarm Intelligence (SI) approach invented in 2009 by Yang and Deb. CS mainly follows three rule sets:

- One egg is laid by each Cuckoo and dumped that egg in an arbitrarily selected nest.
- The high-quality egg is selected and is responsible for the next generation process.
- The host nest is static, and the egg laid by the Cuckoo is discovered by the host bird with a probability p∈ (0, 1). Depending upon the egg's quality, the host bird either select that egg or abandon the nest and create a new one [21].

As per the above rules, the CS works in the following way. Each egg in the given nest represents a candidate solution. Thus, each nest is supposed to be consists of one egg, although each nest may contain many eggs, generally represents a solution package. CS aims to create a solution in a better way. This is possible by replacing the worst cases with a current new solution presented in the nest. The selection of an appropriate egg is performed based on the objective function. In this research, the aim of using CS is to optimize the data set and hence reduce the searching time [22]. The workflow of CS is shown in Fig. 1.

Generally, the objective function is decided related to the problem with its minimum and maximum limit. Mathematically, the relationship between the minimum and maximum problem can be given by equation (1).

$$min(f(x)) = max(-f(x)) \quad (1)$$

After determining the minimum and maximum limit of an objective function, an additional parameter (p) called the transition probability is determined that measures the change in a new randomly produced nest. This parameter includes two elements of the CS algorithm: exploration and exploitation [23]. The extensive exploitation means premature convergence, whereas considerable investigation results in a decrease the convergence.

A random index egg is selected using the CS algorithm, and then these selected eggs are cross-checked against the fitness function of CS. If the attribute value does not satisfy the fitness function, then the attribute value is changed by equation (2).

$$Modified\ attribute\ value = \frac{\sum_{i=1}^{n} A_v}{n} \quad (2)$$

B. FFBPNN

FFBPNN is a parallel processing network, which is composed of a large number of simple integrated processors. This is one of the most commonly used artificial intelligence schemes. The three-layer structure of FFBPNN is shown in Fig. 2.

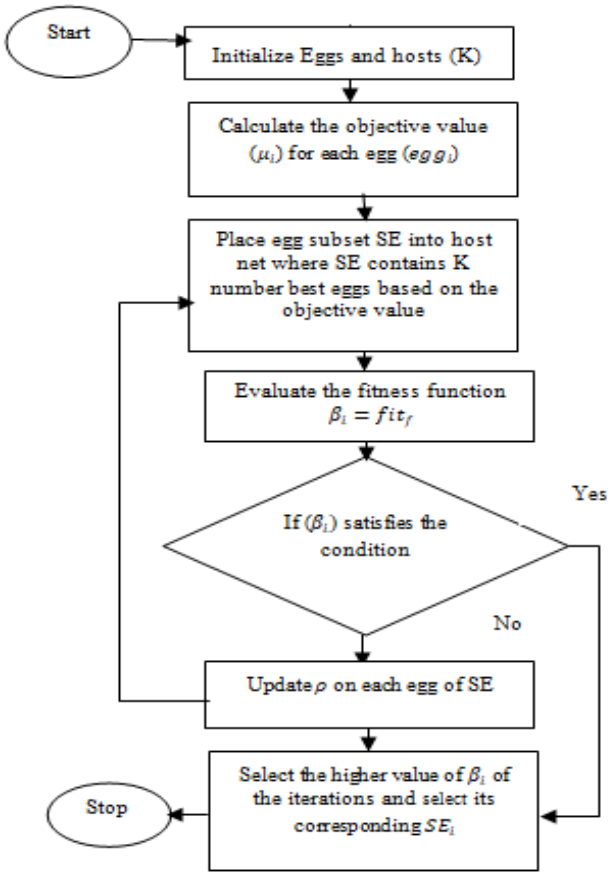


Fig. 1. The Flow of the CS Algorithm.

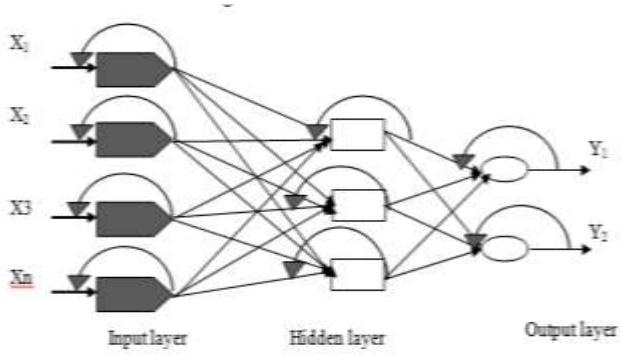


Fig. 2. Architecture of FFBPNN.

As shown in Fig. 2, input data {X=X1, X2, X3….. Xn} is passed as optimized data features to the input layer of FFBPNN. The arrow shows that the error value generated at the output layer is passed to the hidden layer to modify the weight matrix [24].

The standard multi-layer feed-forward network, including three layers, is depicted in Fig. 2. The architecture of the FFBPNN shares a similar characteristic having one layer, which is connected to its nearest layers and sharing neurons in a bidirectional manner. Here, bidirectional means that the information is transmitted or received in both directions. Every connection is assigned with some weight, which can be handled as per certain learning rules.

The generated error, which needs to be minimized, can be defined by equation (3).

$$Error\ (m) = \frac{1}{2}\sum_{r=1}^{r}[O_r - y_r]^2 \quad (3)$$

Where, O_r denotes the ground-truth value or label value. y_r represents the classified output during the simulation. Based on the generated error, the weight of the matrix in the hidden layer is changed as per equation (4).

$$\Delta w_{ij} = -K\frac{derror}{dw_{ij}} \quad (4)$$

IV. Proposed Work

In this research, an automatic intrusion detection system based on the machine learning approach in hybridization with an optimization approach named CS-FFBPNN has been presented. The integrated approach takes the advantage of the FFBPNN as machine learning and CS as a nature-inspired algorithm. CS is used due to its simplicity and ability to resolve non-linear real-world problems [22].

FFBPNN, including single input, 10 hidden and 1 output layers, is used as a multi-class classifier. The input nodes in the input layer correspond to the number of attributes extracted from the dataset and are being optimized by the CS algorithm, whereas the number of nodes in the hidden layer corresponds to the feedback taken from the output layer to obtain the desired output. The output layer of the FFBPNN structure consists of a single neuron, which corresponds to the single output. The value '1' corresponds to normal data, and the value '0' indicates the presence of intrusion in the network. The designed model mainly includes four steps: upload data, optimization using CS, training, and data validation using FFBPNN. The general architecture of the work is shown in Fig. 3.

In the Fig. 4, features are selected based on the fitness function of the cuckoo search algorithm. The data that satisfies the fitness function is selected as the best feature of the data.

For any classification and prediction model, uniqueness and relevant feature selection is an essential step and compulsory to achieve better accuracy during classification or prediction. So, here cuckoo search algorithm is used to select the best set of features using their fitness according to the system requirements. For selecting a set of required and relevant feature, cuckoo search algorithm need a condition in terms of fitness that should be satisfied for selection of a feature as a set of optimized features.

A. Dataset

The CICIDS2017 data set includes benign and the latest common attacks, similar to real-world data (PCAP). The dataset is also composed of using CIC Flow Meter for network traffic analysis that is being labeled based on timestamp, the source IP, the destination IP, the source and destination address of port, protocol, and attack in the form of (CSV file). The extracted feature definition of the dataset is also available.

B. Upload Data

The foremost step in the proposed work is to upload data from the available dataset, as discussed above. After uploading the dataset, pre-processing has been performed to remove the undesired data.

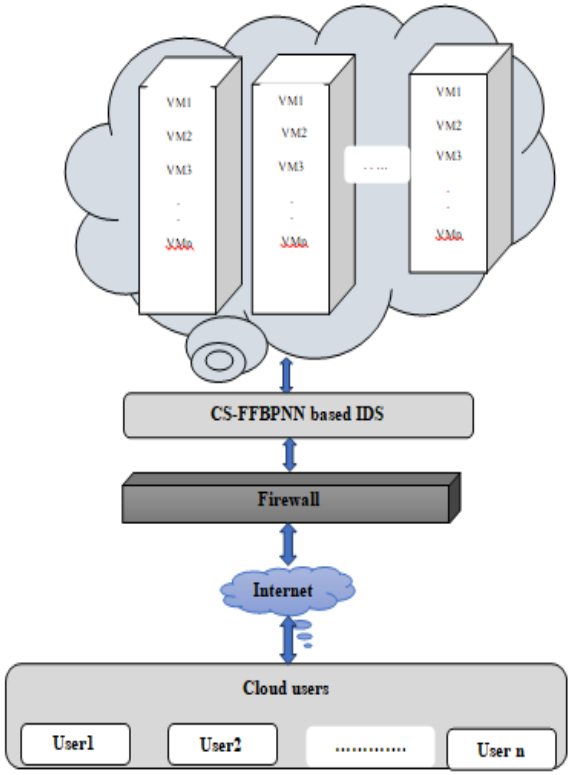


Fig. 3. General Workflow [25].

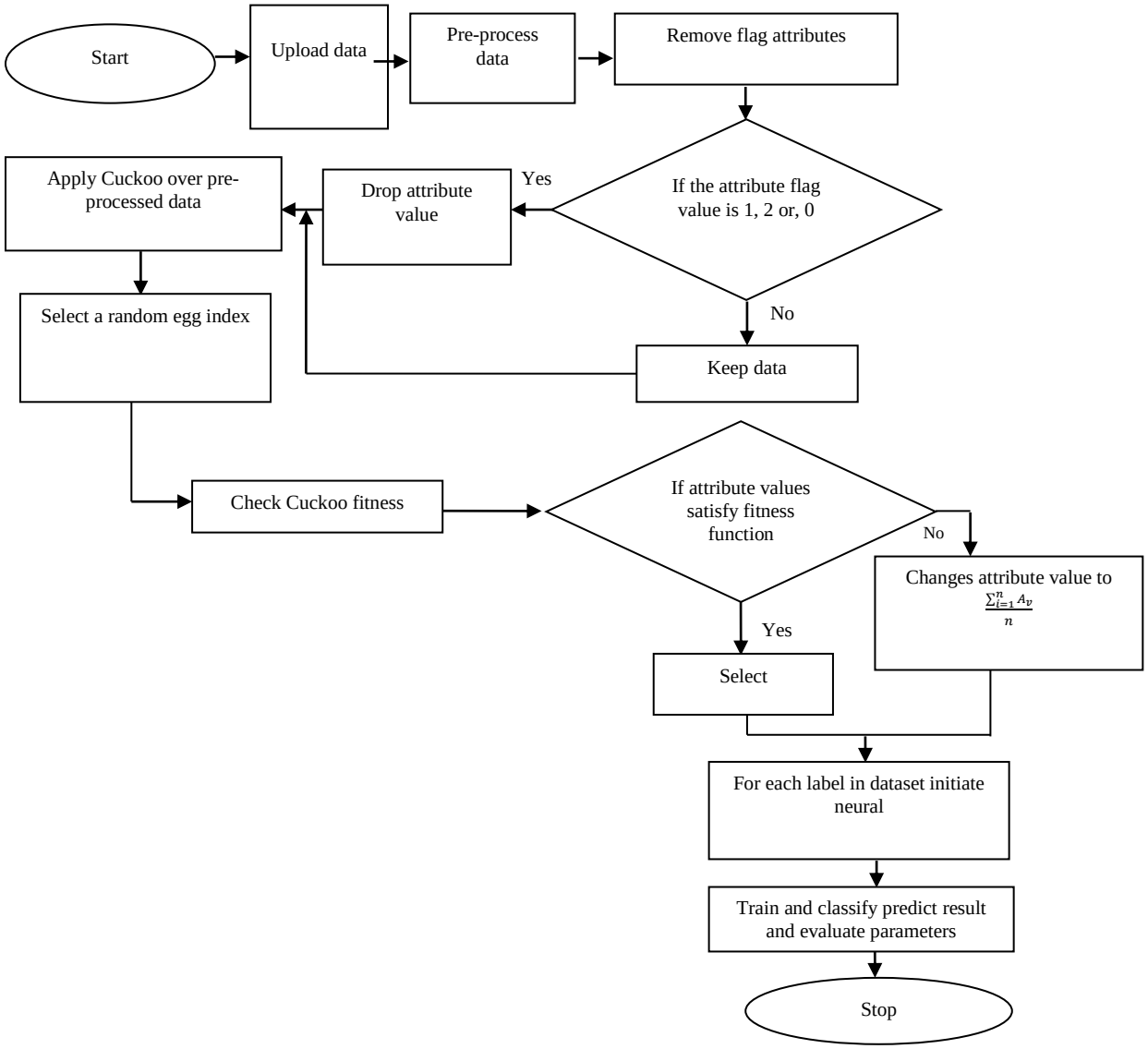


Fig. 4. Proposed Work Flow.

C. Pre-processing

Large disk space is required to store such records coming from multiple users in the cloud network. Therefore, to analyze each record in the log file, pre-processing, or the filtering of the desired record from the unwanted record is an essential step in data processing. It helps researchers to reduce computation time as well as saving the available resources. In other words, one can say that data Pre-processing is the way by which the raw data is processed for further processing. It converts complex/ unstructured data into a structured data form [26]. Here, pre-processing has been applied to remove the flag attributes. The case, when the value of flag attribute is 0, 1, and 2, then dropped that attribute value. Otherwise, passed data as input to the Cuckoo Search (CS) as an optimization algorithm.

Feature extraction plays an essential role in the IDS system. In feature extraction, mainly two processes are being carried out (i) feature construction and (ii) feature selection. The quality of both feature construction and feature selection is essential as they affect the classification accuracy of IDS. The feature construction, as well as the feature selection process, can be carried out manually or automatically using domain knowledge and machine learning approaches, respectively [27].

D. Cuckoo Search (CS) as an Attribute Optimizing Approach

CS is a metaheuristic approach used to optimize data based on the selected fitness function. The fitness function chosen is given in equation (5).

$$Ftness\ function = \begin{Bmatrix} S_b(true) & if S_b > Th_b \\ Th_b\ (false) & Otherwise \end{Bmatrix} \quad (5)$$

Here, S_b is the selected behavior of nodes, and Th_b denotes the threshold behavior of node, which is measured based on the average values of eggs generated by the cuckoos [28].

Based on the above-mentioned fitness criteria, a sensor node is considered as an intruder, if node requires more

energy or transmission time (delay) to forward or receive a data packet, otherwise considered as a normal node. After the segregation of all sensor nodes in the network in two categories, like normal and intruders, here FFBPNN is used to train the network of further classification of the intruders in the network and helps to prevent the network from different kinds of intruders. The working of which is explained in the subsequent section.

E. Feed Forward Back Propagation Neural Network (FFBPNN) as a Classifier

To prepare FFBPNN for detection of intruder automatically, the selection of the number of input neurons, hidden layer neurons must be decided accurately. To obtain the best performance, the modification of hidden neurons must be adequately done by following error detection and correction procedure, i.e., the generated error at the output layer is fed back to the hidden layer, which is again used for the modification of neurons in the hidden layer and then updates the weight matrix. The tangent function is used as an activation function in the hidden layer, whereas for output, the desired output is monitored using the linear activation function. After obtaining the desired output that is '1' for normal and '0' is for an intruder, it is then stored into the FFBPNN as training data. After selecting the appropriate structure, the designed IDS model for the cloud is trained based on the relationship between the input fed to the FFBPNN and the obtained output [13].

Where k is the constant of proportionality, 'error' is the error function, and w_{ij} denotes the weight between neuron i and j, respectively. The process of adjusting the weight matrix is repeated until the desired results are obtained, or the difference between the actual value and the node output is minimum as per acceptance level [13, 24]. The training scenario with a total 20 number of attributes and 30 hidden layers is shown on the right-hand side of Fig. 5.

The trained structure of FFBPNN with the generated Mean Square Error (MSE) is shown in Fig. 5. The network is trained at the 4th iteration with the minimum error of 1.2925 e^{-26} that is acceptance error, and after this, the network is trained, and the data is stored in its database.

After that the training testing process is performed to check the efficiency of the designed model, which is explained in the next section.

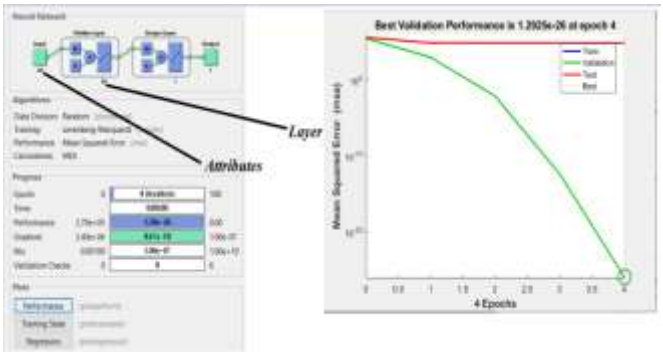


Fig. 5. Trained Structure of FFBPNN.

Algorithm: Enhanced CS_FFBPNN

Input: Pre-processed data (P-Data)
Output: Predicted Results (PR)
Initialization of variables
P-Data: Pre-processed data after the dataset loading
O-Data: Optimized Data as selected attributes from the dataset
Start selection
To optimized the P-data, CS Algorithm is used
Set up basic parameters of CS: Egg population in the nest ($E_{P\text{-}Data}$) = Number of P-Data
Define Fitness function for the selection best P-Data,

Fitness Function: $F(f) = \begin{cases} True; & if\ Selected_{P-Data} < Threshold_{P-Data} \\ False; & Otherwise \end{cases}$

In the fitness function, $selected_{-Data}$ is pre-processed current data present in the P-Data and $Threshold_{P-Data}$ is the threshold properties and the average of P-Data
Calculate Length of P-Data in terms of Row and Column, [Roz, Column] = Size (P-Data)
Set,O-Data= [] // Initiate an empty variable to store selected data
For i in rang of Row
For j in rang of Column
C_E = P-Data (i, j) = E_{P-Data} // Current egg from $E_{P\text{-}Data}$
$M_E = Threshold_{P-Data} = \sum_{i=1}^{Row} \sum_{j=1}^{Column} E_{P-Data}(i,j)$ // Mean of all $E_{P\text{-}Data}$
$F(f) = Fit\ Fun\ (C_E, M_E)$
O-Data = Cuckoo Search (F (f), FR (i, j))
End – For
End – For
Check the index of O-Data
If O-Data (index) = True
O-Data = Select (P-Data)
Else
O-Data = Reject (P-Data)
End –If
Returns: O-Data as a selected data
Initialize FFBPNN
Initialization of variables
→O-Data: Optimized Data as selected attributes from the dataset
→Cat: Target or Category according to the O-Data class
→N: Carrier Neurons Number
→PR: Predicted Results
Start training
Initialize FFBPNN- with O-Data: – Number of Epochs (E) // Iterations used by FFBPNN
– Number of Neurons (N) // Used as a carrier
– Performance: MSE, Gradient, Error Histogram, and Validation
– Data Division: Random
For i in range of all O-Data
If O-Data belongs to Type 1
Cat (1) = Feature from the O-Data of 1st Part // 1st Class of Dataset
Else (Others)
Cat (2) = Feature from the O-Data of 2nd Part // 2nd Class of Dataset
End – If
End – For
Initialized the pattern net using O-Data and Cat
FFBPNN--Structure = FFBPNN- (N)
Set the training parameters according to the requirements and train the system
FFBPNN-Structure = Train (FFBPNN-Structure, O-Data, Cat)
Test Data Group = Sim (FFBPNN-Structure, Current Data for testing)
If Test Data Group = 1
Predicted Results, PR = 1st with performance parameters
Else
Predicted Results, PR = 2nd with performance parameters
End – If
Return: FFBPNN--Structure as a trained structure with PR as a Predicted Results
End – Function
End – Function

V. Experimental Evaluation

Initially, the setup used to perform the analysis of the designed CS-FFBPNN based IDS for the cloud network is presented. Then the evaluated parameters are explained before and after the detection of an intruder in the cloud network.

A. Experimental Setup

The work is designed and implemented in MATLAB 2016 a. The evaluation was performed using an Intel core processor with 4 GB RAM. The designed IDS model's performance has been performed on two datasets, namely DoS attacks and BENIGN. The description of those datasets is provided in Table I.

TABLE I. Dataset Description

Category	Training data	Testing data
DoS attacks	241671	121451
BENIGN	92000	70000
Total	347241	199591

B. Experimental Result

The performance of the proposed research is evaluated in terms of different computation parameters, such as Classification Accuracy, Precision, Recall, and F-measure. All parameters are defined below.

1) Accuracy: It indicates the closeness of the detected output to the actual performance. Mathematically, it is given by equation (6).

$$Accuracy = \frac{T_p+T_N}{T_p+T_N+F_p+F_N} \quad (6)$$

where T_p → True Positive

T_N → True Negative

F_p → False Positive

F_N → False Negative

2) Precision: It represents the number of positively detected classes (normal or intruder) that belong to the positive class.

$$Precision = \frac{T_p}{T_p+F_p} \quad (7)$$

3) Recall: It represents the number of positive class (either normal or intruder) predictions made out of all positive samples saved in the FFBPNN database.

$$Recall = \frac{T_p}{T_p+F_N} \quad (8)$$

4) F-Measure: It represents a single score value that indicates the balance between the precision and recall parameter.

$$F\_measure = 2\times \frac{Precision*Recall}{Precision+Recall} \quad (9)$$

The comparison of precision parameters examined by the proposed work and existing work [20] with respect to a number of records is presented in Fig. 6. The X-axis represents the number of records, and the y-axis represents the precision values for proposed and existing work. The evaluated metrics values are summarized in Table II. Fig. 6 shows that with the increase in the number of records, the precision value also increases. This is due to the appropriate selection of data attributes so that the presence of an intruder is detected at the early stage and hence should be sorted before it affects the data transmission in the cloud network. From the graph, it is seen that precision for the proposed work is higher than existing work performed by Velliangiri *et al.* [20].

Proposed and the existing work for precision is 0.855and 0.775, respectively. Therefore, the percentage increase in precision from the existing work is 10.32 %.

The comparison of recall factor represents the TP entities corresponding to FN entities that are not at all categorized. In Fig. 7, the highest recall value for the proposed as well as for the existing work is recognized at 70,000 records, whereas the minimum recall value is analyzed at 10,000 records. The overall recall for multiple cloud users is analyzed as 0.8648. Also, the average value analyzed for the recall metric of the proposed work and existing work (TEHO-DBN) (Velliangiri et al., 2020) is 0.8648 and 0.812, respectively. Thus, the improvement in the recall values of 6.5% has been achieved against the existing work.

The values of F-measure after the detection of an intruder in the cloud network, for proposed as well as for the existing work for the same dataset, are shown in Fig. 8. The average F-score analyzed for proposed and existing work are 0.859, 0.790 respectively. Therefore, there is an improvement of 8.73% against the existing work.

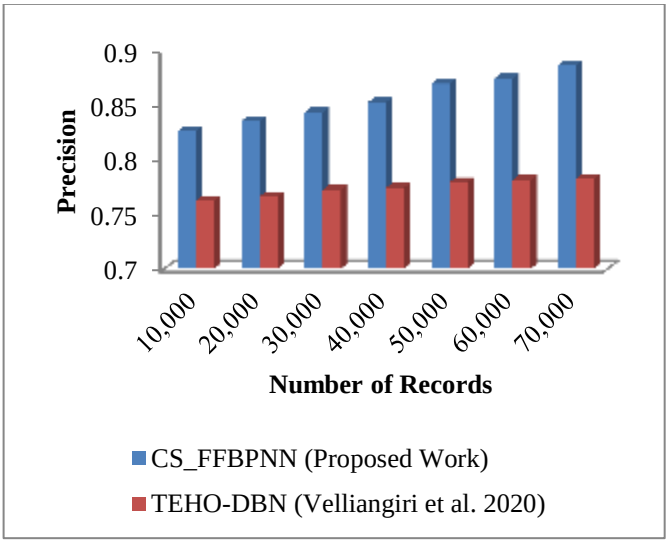


Fig. 6. Precision.

TABLE II. COMPUTED PARAMETRIC ANALYSIS

	Precision		Recall		F-measure		Classification Accuracy	
Number of Records	**CS_FFBPNN (Proposed Work)**	**TEHO-DBN (Velliangiri *et al.*, 2020)**	**CS_FFBPNN (Proposed Work)**	**TEHO-DBN (Velliangiri *et al.*, 2020)**	**CS_FFBPNN (Proposed Work)**	**TEHO-DBN (Velliangiri *et al.*, 2020)**	**CS_FFBPNN (Proposed Work)**	**TEHO-DBN (Velliangiri *et al.*, 2020)**
10,000	0.826	0.762	0.831	0.795	0.828	0.779	0.826	0.725
20,000	0.835	0.766	0.847	0.797	0.840	0.789	0.835	0.729
30,000	0.843	0.772	0.859	0.802	0.852	0.788	0.857	0.734
40,000	0.852	0.774	0.868	0.808	0.859	0.792	0.862	0.736
50,000	0.869	0.779	0.874	0.814	0.871	0.80	0.878	0.742
60,000	0.874	0.781	0.884	0.816	0.878	0.812	0.882	0.745
70,000	0.886	0.782	0.891	0.824	0.888	0.825	0.896	0.748

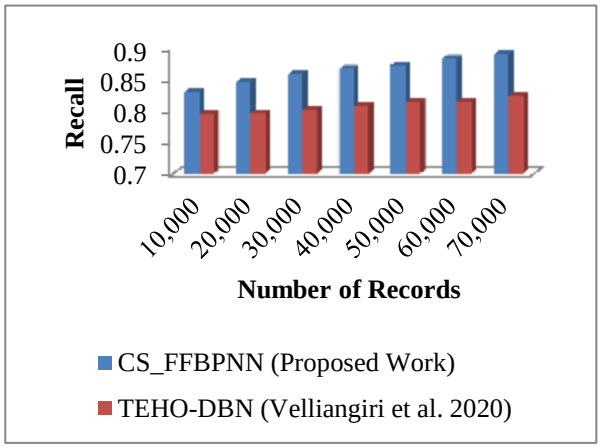


Fig. 7. Recall.

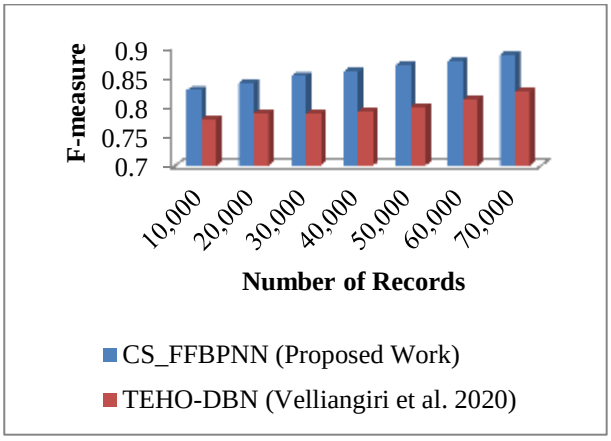


Fig. 8. F-measure.

To categorize the attacks considered in the database efficiently, IDS is used to ensure the security of the cloud data. To achieve better accuracy, proper training of the classification algorithm (FFBPNN) in this research is required. As the presence of irrelevant features in the dataset leads to increase in computation time, increase in error, and decrease classification accuracy. To solve this problem, the dataset is pre-processed and then optimized using a well-known Swarm Intelligence Cuckoo Search Approach (SI-CS) approach. The obtained results for the proposed IDS system for classification accuracy compared to existing work are shown in Fig. 9. Using CS-FFBPNN as IDS for the cloud system, the classification accuracy increases successfully. The graph shows improvement in the proposed work compared to the (Velliangiri *et al.*, 2020) work. The average classification accuracy analyzed for the proposed work, and the existing work are 0.8622 and 0.734. Therefore, the percentage increase in the classification rate from existing work is 17.47%.

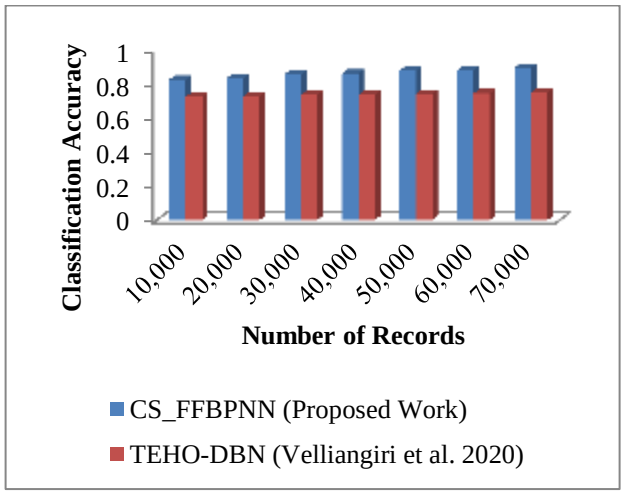


Fig. 9. Classification Accuracy.

The existing work represented by Velliangiri *et al.* (2020) is only suitable for DDoS attacks, but the proposed work is implemented on two different types of attacks which are DDoS and BENIGN. The proposed work provides better results as compared to existing approach. The proposed work also obtained the better values as compared to the existing approach against different parametric values, namely, precision, recall, classification accuracy, and F-measure.

VI. CONCLUSION

Cloud computing is used as a shared pool of resources that provides fast computing and aims to give convenient and, at the same time, required network access with minimal effort. The machine learning approach has offered the advantage of being interested in computing needs, and there is a suggestion for optimizing the unstructured data using the CS algorithm has also been presented. CS selected the optimal features from the pre-processed data and contributed to enhance the classification accuracy of the training and the testing phase of

the designed IDS cloud network. The proposed approach performed well in contrast to the existing work. As in existing work, authors have used a deep learning-based classifier in addition to Taylor-Elephant Herd Optimization, and the system time complexity increases as well as provides inadequate response for large network traffic. The work is evaluated against four performance metrics namely, precision, recall, f-measure, and classification accuracy of the designed IDS. From the test results, the examined accuracy has been observed as 86.22%. This accuracy might be low because the research has focused on detecting two types of attacks, such as DDoS attacks, and BENIGN attacks, for a sample of extensive data to train and test the network for intruder detection.

The proposed work is not implemented on real world data set. In future work, real world data set will be implemented by applying metaheuristic algorithm to achieve more accuracy based upon different types of attacks.

REFERENCES

[1] D.,Callegari, E., Conte, T., Ferreto, D., Fernandes, F., Moraes, F., Burmeister, & R. Severino. EpiCare—"A home care platform based on mobile cloud computing to assist epilepsy diagnosis", In 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH) , IEEE , pp. 148-151, November 2014.

[2] P., Mell, & T. Grance. "The NIST definition of cloud computing", 2011.

[3] M., Moorthy, & M, Rajeswari. "Virtual host based intrusion detection system for cloud. International Journal of Engineering & Technology", pp. 0975-4024, 2013.

[4] A., Carlin, M., Hammoudeh, & O., Aldabbas. "Defence for distributed denial of service attacks in cloud computing. Procedia computer science", vol. 73, pp. 490-497, 2015.

[5] S., Goyal. "Public vs private vs hybrid vs community-cloud computing: a critical review", International Journal of Computer Network and Information Security, vol. 6, no. 3, 2016.

[6] P., Ghosh, S., Shakti, & S, .Phadikar. "A cloud intrusion detection system using novel PRFCM clustering and KNN based dempster-shafer rule", International Journal of Cloud Applications and Computing" (IJCAC), vol. 6, no. 4, pp. 18-35, 2016.

[7] K., Indira, D., UshaNandini, & A., Sivasangari. "An efficient hybrid intrusion detection system for wireless sensor networks. Int J Pure Appl Math", vol. 119, no. 7, pp. 539-556 2018.

[8] M. A., Kumbhare & M. M., Chaudhari. IDS: "survey on intrusion detection system in cloud computing. Int. J. Comput. Sci. Mob. Comput"., vol. 3, no.4, pp. 497-502, 2014.

[9] E., Besharati, M., Naderan, & E., Namjoo, "LR-HIDS: logistic regression host-based intrusion detection system for cloud environments. Journal of Ambient Intelligence and Humanized Computing, vol. 10, no.9, pp. 3669-3692, 2019.

[10] T., Nathiya, & G., Suseendran. "An effective hybrid intrusion detection system for use in security monitoring in the virtual network layer of cloud computing technology". In Data management, analytics and innovation, Springer, Singapore, pp. 483-497, 2019.

[11] K., Hashizume, D., Rosado, G., Fernández-Medina, E., & E. B., Fernandez. "An analysis of security issues for cloud computing. Journal of internet services and applications", vol.4, no., pp. 1-13, 2013.

[12] P., Ghosh, A. K., Mandal, & R., Kumar. "An efficient cloud network intrusion detection system. In Information systems design and intelligent applications, Springer, New Delhi, pp. 91-99, 2015.

[13] N., Pandeeswari, & G., Kumar, "An anomaly detection system in a cloud environment using a fuzzy clustering-based ANN. Mobile Networks and Applications," vol. 21, no.3, pp. 494-505, 2016.

[14] M. M., Baig, M. M., Awais & E. S. M., El-Alfy. "A multiclass cascade of artificial neural network for network intrusion detection. Journal of Intelligent & Fuzzy Systems", vol. 32, no.4, pp. 2875-2883, 2017.

[15] V., Mahajan & S. K. Peddoju, "Deployment of the intrusion detection system in the cloud: a performance-based study", In IEEE Trust com/Big Data SE/ICESS, pp. 1103-1108. IEEE, August 2017.

[16] P., Deshpande, S. C., Sharma, S. K., Peddoju, & S. Junaid, HIDS: "A host-based intrusion detection system for a cloud computing environment", International Journal of System Assurance Engineering and Management, vol. 9, no. 3, pp. 567-576, 2018.

[17] B., Hajimirzaei, & N. J. Navimipour, "Intrusion detection for cloud computing using neural networks and artificial bee colony optimization algorithm", ICT Express, vol. 5, no. 1, pp. 56-59. 2019.

[18] K. R., Kumari, P., Sengottuvelan, & J. Shanthini, "A hybrid approach of genetic algorithm and multi-objective PSO task scheduling in cloud computing.", Asian Journal of Research in Social Sciences and Humanities, vol. 7, no. 3, pp. 1260-1271, 2017.

[19] M., Manickam, & S. P. Rajagopalan, "A hybrid multi-layer intrusion detection system in the cloud", Cluster Computing, vol. 22, no. 2, pp. 3961-3969, 2019.

[20] S., Velliangir, P., Karthikeyan & V. Vinoth Kumar, "Detection of distributed denial of service attack in cloud computing using the optimization-based deep networks", Journal of Experimental & Theoretical Artificial Intelligence, pp. 1-20, 2020.

[21] X. S., Yang, & S. Deb, "Cuckoo search via Lévy flights. In 2009 World congress on nature & biologically inspired computing", (NaBIC) pp. 210-214, IEEE, December, 2009.

[22] Jr, I., Fister Fister, D., & I. Fister, "A comprehensive review of cuckoo search: variants and hybrids", International Journal of Mathematical Modelling and Numerical Optimisation, vol. 4, no. 4, pp. 387-409, 2013.

[23] I., Fister, X. S., Yang, & D. Fister, "Cuckoo search: a brief literature review. In Cuckoo search and firefly algorithm", Springer, Cham, pp. 49-62, 2014.

[24] S. M., Mehibs & S. H. Hashim "Proposed network intrusion detection system in a cloud environment based on backpropagation neural network", Journal of the University of Babylon for Pure and Applied Sciences, vol. 26, no. 1, pp. 29-40, 2018.

[25] D. A. A. G., Singh, R., Priyadharshini, & E. J., Leavline "Cuckoo optimisation-based intrusion detection system for cloud computing." International Journal of Computer Network and Information Security, vol. 9, no.11, pp. 42-49, 2018.

[26] N., Paulauskas, & J., Auskalnis, "Analysis of data pre-processing influence on intrusion detection using NSL-KDD dataset. Open conference of electrical, electronic and information sciences (eStream), IEEE, pp. 1-5, 2017.

[27] G., Serpen, & E., Aghaei, "Host-based misuse intrusion detection using PCA feature extraction and kNN classification algorithms. Intelligent Data Analysis", vol. 22, no.5, pp. 1101-1114, 2018.

[28] V., Ravindranath, S., Ramasamy, R., Somula, K. S, Sahoo, & A. H, Gandomi, "Swarm intelligence-based feature selection for intrusion and detection system in cloud infrastructure." In IEEE Congress on Evolutionary Computation (CEC), pp.1-6, 2020.

Construction of a Model and Development of an Algorithm for Solving the Wave Problem under Pulsed Loading

Khabdolda Bolat[1], Zhuzbayev Serik[2], Sabitova Diana S[3]
Aitkenova Ailazzat A[4], Serikbayeva Sandugash[5], Badekova Karakoz Zh[6], Yerzhanova Akbota Y[7]
Department of Information Systems, L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan[1, 2, 5]
Department of Information Systems and Computer Engineering, Sh.Ualikhanov Kokshetau University[3]
Department of Computer Science and Biostatistics, Karaganda Medical University[4, 6]
S. Seifullin Kazakh Agro Technical University, Nur-Sultan, Kazakhstan[7]

***Abstract*—The article considers approaches and methods for modeling the wave process resulting from blasting operations. The analysis of modeling methods has shown that in the context of the task it is advisable to conduct a study based on the application of the method of behavioral characteristics, which was optimized using the splitting method. The defining equations were calculated, the point scheme of the template was selected, the resolving difference equations for dynamic boundary value problems of a seismic nature were calculated. Based on the method, an algorithm for calculating the relationship between the voltage and the seismic medium was developed, which allowed generating a code and designing an information system for calculating the wave process.**

***Keywords*—Information systems; wave process; explosive technologies; method of bicharacteristics; stress tensor**

I. Conduction

When compacting the foundations of structures, sinking underground developments, the behavior of embankments, dams, dams, etc., explosion energy is widely used. To achieve the maximum effect of blasting, it is necessary to properly study the effect of explosives on soils as a special dynamic effect. In the practice of explosives, both individual charges and very complex systems of them are used, which is undoubtedly of wide technical interest [1,2]. In [3,4], the influence of the pile driving depth on the stress-strain state of the base is studied, the interaction of the pile structure with the ground is studied, and the stresses arising in the piles themselves during driving into the ground are analyzed.

The increasing volume of industrial, mining, hydraulic engineering and aviation engineering makes it necessary to improve the methods of studying wave problems. The known methods of solving problems [5-13] can not always fully reveal the features of contact problems of dynamics.

As part of the study of wave processes using explosive technologies, the problem of propagation of elastic waves under pulsed loading was chosen. To solve this problem, an optimized method of behavioral characteristics is used with the addition of the ideas of the splitting method [14, 15]. The solution of a number of problems based on this method contributed to the writing of an algorithm and the development of information for the analysis of wave processes in various media [16,17], including using composite materials technology [18].

II. Problem Statement

Let a repeated dynamic load (Fig. 2), which is the result of blasting operations, act on the resting quarter plane with the insert (Fig. 1) at some depth from the free surface and at some distance from the insert. At the same time, a shock wave propagates from the source of the explosion, which, as it moves away from it, turns into a continuous compression wave.

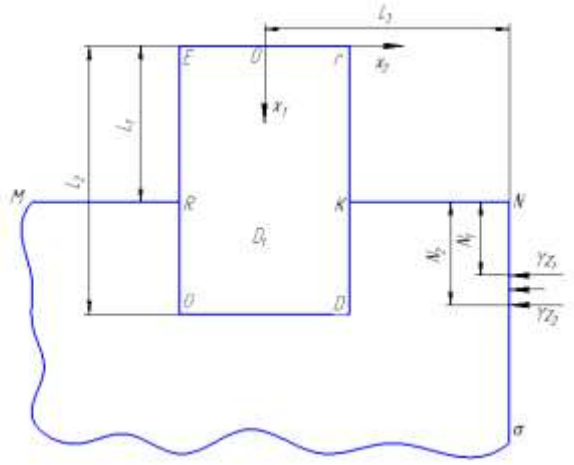


Fig. 1. The Study Area.

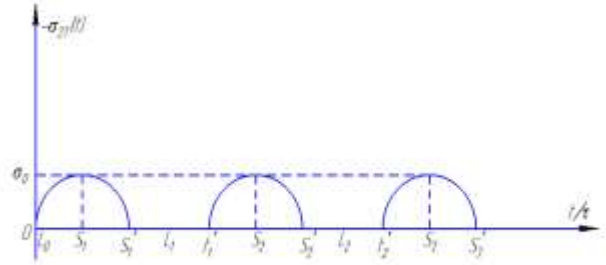


Fig. 2. The Shape of the Current Load.

It reaches the insertion and determines its tense state. There are many technical problems here, for example, insertion fluctuations. In any case, the initial urgent task is to study the features and critical values of stress fields in the perturbed region [24].

In mathematical modeling of the problem of underground blasting, the soil is considered as an elastic medium occupying a quarter of the plane, and the insert as an elastic structure. The shock load is represented as a harmonic distributed load that acts along the normal to the vertical surface of the quarter plane on a certain section. In the Cartesian coordinate system $x_i\ (i = 1, 2)$ the rectangular insert (body 1) occupies the area$D_1(0 \le x_1 \le L_2 \cap |x_2| \ge l)$ and the quarter plane (body 2) is the area $D_2(L_1 \le x_1 \le \infty \cap -\infty \le x_2 \le L_3)$ (Fig. 1), and the quarter plane (body 2) is the $D_2(L_1 \le L_2; l \ge L_3)$.

The boundary value problem assumes that the initial:

$$v_\alpha^{(k)} = \sigma_{\alpha j}^{(k)} = 0, (\alpha, j = 1\ ,2; k = 1, 2), \quad (1)$$

boundary conditions:

$$\sigma_{1j}^{(2)} = 0, (j = 1, 2) \text{ by } x_1 = L_1, l \le |x_2| < \infty \quad (2)$$

$$\sigma_{j2}^{(1)} = 0, (j = 1, 2) \text{ by } 0 \le x_1 \le L_1, |x_2| = l \quad (3)$$

and contact conditions:

$$v_\alpha^{(1)} = v_\alpha^{(2)}, \sigma_{1j}^{(1)} = \sigma_{1j}^{(2)}, (\alpha, j = 1{,}2) \text{ by } x_1 = L_2, |x_2| \le l \quad (4)$$

$$v_\alpha^{(1)} = v_\alpha^{(2)}, \sigma_{j2}^{(1)} = \sigma_{j2}^{(2)}, (\alpha, j = 1{,}2) \text{ by } L_1 \le x_1 \le L_2, |x_2| = l \quad (5)$$

At the boundary $EF\ (x_1 = 0, |x_2| \le 1)$, the stress-free surface conditions are assumed (Fig. 1).

$$\sigma_{11}^1 = 0, \sigma_{12}^1 = 0 \quad (6)$$

It is assumed that the NB boundary $NB\ (0 \le x_1 \le \infty, x_2 = L_3)$ is also free from stresses (3) and only on the local section $(N_1 \le x_1 \le N_2, x_2 = L_3)$of the NB boundary of the quarter plane, the normal componentσ_{22}^2 acts at certain time intervals, changing according to the la.

$$\sigma_{22}^2(t) = \begin{cases} 0, t \le 0\ S_i \le t \le t_i^* \\ A_0 \sin\left(\frac{\omega(t-S_i)}{B_0}\right), t_{i-1}^* \le t \le S_i \end{cases} \quad (7)$$

$$\sigma_{12}^2(t) = 0,$$

where $i = 1{,}2, \dots, n; B_0 = S_i - t_{i-1}^*; S_i = \frac{B_0}{2+t_{i-1}^*}$. The amplitude A_0 and the frequency ω the oscillation frequency remains constant. The boundary conditions (7) simulate the conditions of blasting operations in the well, acting at a certain $(x_1 \le L_1, x_2 = L_3 - l\ )$distance from the structure D_1.

Numerical calculations of the problem were carried out with the following initial data: $f(t) = -A_0 t \exp(-t)\ by\ t \ge 0; \tau = \Delta t = 0.002; l = 10h; L_1 = 10h; L_2 = 20h; h = \Delta x_1 = \Delta x_2 = 0.05$. External loads $A_0 = -0.5, \omega = 9$. The insertion configuration (region D_1)) and the local impact site varied.

Under the described conditions, it is necessary to investigate the stress-strain state of an inhomogeneous medium $D_1 \cap D_2$.

III. Method of Behavioral Characteristics with the Ideas of the Splitting Method

To solve this problem, along with the initial (1) and boundary conditions (1) - (5), a system of equations consisting of equations of motion and relations of the generalized Hooke's law is used.

$$\rho_k \ddot{u}_\alpha^{(k)} = \sigma_{\alpha\beta,\beta}^{(k)}$$

$$\sigma_{\alpha j}^{(k)} = \lambda_k \varepsilon_{\beta\beta}^{(k)} \delta_{\alpha j} + 2\mu_k \varepsilon_{\alpha j}^{(k)} \quad (8)$$

where $\varepsilon_{\alpha j}^{(k)} = 0.5\left(u_{\alpha,j}^{(k)} + u_{j,\alpha}^{(k)}\right);\ \delta_\alpha$ – Kronecker symbol, $u_\alpha^{(k)}, \varepsilon_{\alpha j}^{(k)}$ – components of the displacement vector and the strain tensor.

It is convenient to calculate the solution of the problem in a dimensionless space of variables and the desired parameters, which are obtained after the introduction of the notation described in [19].

A. Defining Equations of the Dynamic Problem of the Theory of Elasticity

Using the relations from [19] for dimensionless quantities, one can obtain (i≠j) from equations (8) after simple transformations:

$$\rho_k \dot{v}_\alpha^{(k)} = \sigma_{\alpha\beta,\beta}^{(k)}; \dot{\sigma}_{\alpha j}^{(k)} = \left(\gamma_{11}^{(k)} v_{j,j}^{(k)} + \gamma_{33}^{(k)} v_{i,i}^{(k)}\right)\delta_{\alpha j} + \gamma_{12}^{(k)}\left(v_{\alpha,j}^{(k)} + v_{j,\alpha}^{(k)}\right)\left(1 - \delta_{\alpha j}\right) \quad (9)$$

Equations (9) are a linear homogeneous hyperbolic system of first-order differential equations with constant coefficients [16]. Its characteristic surfaces in three-dimensional space $(x_1, x_2; t)$ are hyperconuses with axes parallel to the time axis (Fig. 3).

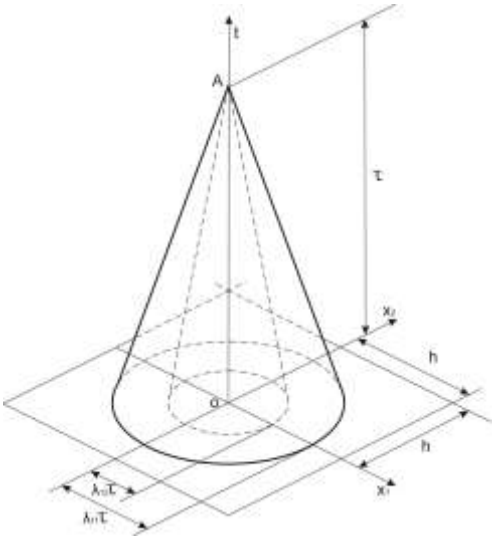


Fig. 3. Characteristic Cones on the Plane.

The system of equations (9) has two families of characteristic cones. These cones coincide with the behavioral characteristics of equations (9). Along the behavioral characteristics lying in the plane$x_i = const$, equations (9) are functions of only two variables $(x_j;\ t)$. This circumstance indicates that the conditions on the behavioral characteristics can be obtained as conditions on the characteristics in the corresponding one-dimensional problem. The corresponding transformations can be performed if one of the spatial variables is alternately fixed in the system of equations (9) [20]. In this case, the system of equations (9) is split into two systems of equations corresponding to the directions j=l and j=2 (i≠j):

$$\dot{v}_\alpha^{(k)} - \rho_k^{-1}\sigma_{\alpha j,j}^{(k)} = A_{\alpha i}^{(k)}$$

$$\dot{\sigma}_{\alpha j}^{(k)} - \gamma_{\alpha j}^{(k)} v_{\alpha j}^{(k)} = B_{\alpha i}^{(k)} \qquad (10)$$

where the notation is entered:

$$A_{\alpha i}^{(k)} = \rho_k^{-1}\left(\sigma_{\alpha\beta,\beta}^{(k)} - \sigma_{\alpha j,j}^{(k)}\right)$$

$$B_{\alpha i}^{(k)} = \gamma_{33}^{(k)}\left(v_{\alpha,\alpha}^{(k)} - v_{\alpha,j}^{(k)}\right)\delta_{\alpha j} - \gamma_{12}^{(k)}\left(v_{i,j}^{(k)} - v_{j,i}^{(k)} - v_{\alpha,j}^{(k)}\right)(1 - \delta_{\alpha j}) \qquad (11)$$

The differential equations of characteristics have the form:

$$dx_\alpha = \pm\lambda_{\alpha j}^{(k)} dt, \qquad (12)$$

and the conditions on the bicharacteristics are:

$$d\sigma_{\alpha j,j}^{(k)} \pm (-1)\rho_k\lambda_{\alpha j}^{(k)}\, dv_\alpha^{(k)} = \left(B_{\alpha i}^{(k)} \pm (-1)\rho_k\lambda_{\alpha j}^{(k)}\, A_{\alpha i}^{(k)}\right)dt \qquad (13)$$

There $\lambda_{\alpha j}^{(k)} = \gamma_{11}^{(k)}$ if α=j and$\lambda_{\alpha j}^{(k)} = \gamma_{12}^{(k)}$ if α≠j. It can be seen from (12) that on each of the two hyperplanes there are two pairs of families of bicharacteristics that determine the longitudinal $\lambda_{ii}^{(k)}$ and shear$\lambda_{ij}^{(k)}$ $(i \neq j;\ i,j = 1,2)$ wave propagation velocities (Fig. 3). In each of the two planes (x_i; t) there are two families of behavioral characteristics of the positive and negative directions. The upper sign corresponds to the behavioral characteristics of the positive, and the lower sign corresponds to the negative directions. Equations (12) and (13) correspond to each other with the same pair of indices and with the same arrangement of signs. Equations (10) and conditions (13) are used to find a solution to the formulated problem (1) -(5) [23].

B. Selecting the Point Scheme of the Template

To perform numerical calculations of the formulated problem for a region with a given configuration$D_1 \cap D_2$, it is necessary to study the characteristic surfaces. The body$D_1 \cap D_2$ is exposed to non-stationary loads. The initial conditions (1) are given by stresses and displacement velocities in the entire body, and the boundary conditions are given by stresses on the surface (2) - (3). It is assumed that both are continuous differentiable functions. The shape of the body is such that it admits the existence of a coordinate system x_i $(i = 1,2)$, in which the boundary surfaces are coordinate [25].

Let the body $D_1 \cap D_2$be divided into cells formed by intersections of coordinate surfaces$x_i = const$ $(i = 1,2)$. The linear dimensions of these cells in the direction of the axes x_1 and x_2 are considered uniform and equal to h. The intersections of the lines $x_i = const$ $(i = 1,2)$ form nodes. At these nodal points, the values of the desired functions $v_\alpha^{(k)}, \sigma_{\alpha j}^{(k)}$ $(\alpha, j = 1,2)$ are found at various time points $t_n - \tau, t_n, t_n + \tau$ $(n = 1,2,\dots,N)$ in time increments τ.

A template consisting of a node *O* and points $E_{\alpha j}^{\pm(k)}$lying on the coordinate lines $x_j = const$ and spaced from the point *O* at distances $\lambda_{11}^{(k)}\tau$ and $\lambda_{12}^{(k)}\tau$ is accepted (Fig. 4). Oblique lines originating from point A are bicharacteristics. In the future, the values of the functions at point *O* are assigned the upper sign "0"; at points $E_{\alpha j}^{\pm(k)}$ - the lower and upper signs ± (for example, $\sigma_{\alpha j}^{\pm(k)}$), and at point A an additional index is not assigned [21-24].

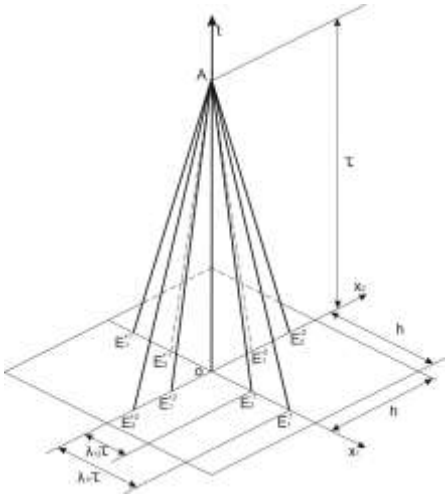


Fig. 4. Point Grid of the difference Scheme for Internal Points.

Based on the described point schemes, the method developed below for solving dynamic problems allows us to determine the particle velocities $v_\alpha^{(k)}$ and the components of the stress tensor $\sigma_{\alpha j}^{(k)}$ at point *A* on the calculated layer in time t_n if their values are known on the previous layert_{n-1} $(n = 1,2,\dots,N)$ at point O and at adjacent points $E_{\alpha j}^{\pm(k)}$. Difference schemes of this type are called explicit. Explicit schemes are convenient because there are no difficulties in solving the systems of difference equations associated with them. These systems are solved sequentially from one-time layer to the next. In this case, the desired values at each node point, unlike the implicit difference scheme, are calculated independently of the others [20].

C. Resolving difference Equations for Solving Dynamic Boundary Value Problems

Resolving difference equations at internal points. The calculation algorithm of the second order of accuracy is constructed below [20]. The integration of the system of equations (10) from point O to point A and the relations (13) from point $E_{\alpha j}^{\pm(k)}$to point A by the trapezoid method (Fig. 4) allows us to obtain expressions of the following type:

$$v_\alpha^{(k)} = v_\alpha^{0(k)} + \frac{\tau}{2}\left(\dot{v}_\alpha^{0(k)} + \rho_k^{-1}\sigma_{\alpha j,j}^{(k)} + A_{\alpha i}^{(k)}\right);$$

$$\sigma_{\alpha j}^{(k)} = \sigma_{\alpha j}^{0(k)} + \frac{\tau}{2}\left(\dot{\sigma}_{\alpha j}^{0(k)} + \gamma_{\alpha j}^{(k)} v_{\alpha j}^{(k)} + B_{\alpha i}^{(k)}\right) \quad (14)$$

and

$$\sigma_{\alpha j,j}^{(k)} - \sigma_{\alpha j}^{\pm(k)} \pm (-1)\rho_k \lambda_{\alpha j}^{(k)}\left(v_\alpha^{(k)} - v_\alpha^{\pm(k)}\right) =$$

$$= \frac{\tau}{2}\left(B_{\alpha i}^{(k)} + B_{\alpha i}^{\pm(k)} \pm (-1)\rho_k \lambda_{\alpha j}^{(k)}\left(A_{\alpha i}^{(k)} + A_{\alpha i}^{\pm(k)}\right)\right) \quad (15)$$

where the unknown values at point A are taken without additional indexes.

The values of functions at non-node points $E_{\alpha j}^{\pm(k)}$ are replaced by the values calculated by the Taylor formula with first-order accuracy for the functions $A_{\alpha i}^{\pm(k)}$and $B_{\alpha i}^{(k)}$ and with second-order accuracy for the functions $v_\alpha^{(k)}$ and $\sigma_{\alpha j}^{(k)}$ through their values at the node points $O\,(x_1, x_2; t)$ [20]:

$$A_{\alpha i}^{\pm(k)} = A_{\alpha i}^{0(k)} \pm (-1)\lambda_{\alpha j}^{(k)}\tau\frac{\partial A_{\alpha i}^{(k)}}{\partial x_j};$$

$$B_{\alpha i}^{\pm(k)} = B_{\alpha i}^{0(k)} \pm (-1)\lambda_{\alpha j}^{(k)}\tau\frac{\partial B_{\alpha i}^{(k)}}{\partial x_j} \quad (16)$$

and

$$\sigma_{\alpha j}^{\pm(k)} = \sigma_{\alpha j}^{0(k)} \pm (-1)\lambda_{\alpha j}^{(k)}\tau\frac{\partial \sigma_{\alpha j}^{0(k)}}{\partial x_j} + \frac{1}{2}\left(\lambda_{\alpha j}^{(k)}\tau\right)^2\frac{\partial^2 \sigma_{\alpha j}^{0(k)}}{\partial x_j^2};$$

$$v_\alpha^{\pm(k)} = v_\alpha^{0(k)} \pm (-1)\lambda_{\alpha j}^{(k)}\tau\frac{\partial v_\alpha^{0(k)}}{\partial x_j} + \frac{1}{2}\left(\lambda_{\alpha j}^{(k)}\tau\right)^2\frac{\partial^2 v_\alpha^{0(k)}}{\partial x_j^2} \quad (17)$$

Partial derivatives of the system of equations (10) with respect to the variable x_j are written as:

$$\frac{\partial \dot{v}_\alpha^{0(k)}}{\partial x_j} = \rho_k^{-1}\frac{\partial \sigma_{\alpha j,j}^{0(k)}}{\partial x_j} + \frac{\partial A_{\alpha i}^{0(k)}}{\partial x_j}$$

$$\frac{\partial \dot{\sigma}_{\alpha j}^{0(k)}}{\partial x_j} = \gamma_{\alpha j}^{(k)}\frac{\partial v_{\alpha j}^{(k)}}{\partial x_j} + \frac{\partial B_{\alpha i}^{(k)}}{\partial x_j} \quad (18)$$

Substituting the relations (16), (17) in (15), then excluding with the help of (14) the variables $v_\alpha^{(k)}$, $\sigma_{\alpha j}^{(k)}$and taking into account (18) , we can obtain eight equations with respect to the derivatives $v_{\alpha j}^{(k)}$, $\sigma_{\alpha j,j}^{(k)}$, $A_{\alpha i}^{(k)}$, $B_{\alpha i}^{(k)}$:

$$\gamma_{\alpha j}^{(k)\,2} v_{\alpha j}^{(k)} \pm (-1)\lambda_{\alpha j}^{(k)}\sigma_{\alpha j,j}^{(k)} =$$

$$= \gamma_{\alpha j}^{(k)\,2}\left(v_{\alpha j}^{0(k)} + \tau\frac{\partial \dot{v}_\alpha^{0(k)}}{\partial x_j}\right) \pm \lambda_{\alpha j}^{(k)}\left(\sigma_{\alpha j,j}^{0(k)} + \tau\frac{\partial \dot{\sigma}_{\alpha j}^{0(k)}}{\partial x_j}\right) \quad (19)$$

Adding and subtracting the corresponding pairs of equations (19) in turn, we can find unknown derivatives:

$$v_{\alpha j}^{(k)} = v_{\alpha j}^{0(k)} + \tau\frac{\partial \dot{v}_\alpha^{0(k)}}{\partial x_j};$$

$$\sigma_{\alpha j,j}^{(k)} = \sigma_{\alpha j,j}^{0(k)} + \tau\frac{\partial \dot{\sigma}_{\alpha j}^{0(k)}}{\partial x_j} \quad (20)$$

The system of equations (20) can be used to determine unknown derivatives, both in the internal and boundary nodal points of the studied region $D_1 \cap D_2$. Such expressions can be obtained directly by integrating the system of equations (9) according to the Euler scheme, having previously differentiated them by x_i. However, it is important to have intermediate relations (19), which are used in solving systems of equations where boundary functions are given. The substitution of equalities (20) in (14) allows us to obtain unknown functions $v_\alpha^{(k)}$, $\sigma_{\alpha j}^{(k)}$ at the internal nodal points of an inhomogeneous body at time $t_{n-1} + \tau$ $(n = 1, 2, \ldots, N)$ [21-22].

IV. Analysis of the Result

Based on the developed information system, the calculation results were obtained, shown in Fig. 5 to 7, when $l = 5h, L_1 = 5h, L_2 = 10h, N_1 = 8h, N_2 = 11h,$ calculation results.

Fig. 5 shows the changes in the normal stresses along $\sigma_{22}^k(k = 1,2)$ the $\frac{x_2}{h}$ coordinate at time t=60τ in different sections:

$1 - x_1 = 5h$ (this section passes through both the insert and the quarter plane);

$2 - x_2 = 10h$ (the cross-section runs along a quarter of the plane);

$2' - x_1 = 10h$ (the cross-section passes through the insert);

$3 - x_1 = 0h$ (the cross-section passes through the insert).

The normal stresses $\sigma_{22}^k(k = 1,2)$ in Section 1 near the angular point $N(x_1 = L_1, x_2 = L_3)$, at the boundary of the free surface $PN(x_1 = L_1, l \le x_2 \le L_3$ and in the insertion region $(x_1 = L_1, |x_2| \le l)$ change abruptly with different local extrema. The normal stresses σ_{22}^2 in the section 2 in the quarter plane reach the maximum extreme value (-0.22). It is caused by a given load. On the contact surface of the insert and the quarter plane, the normal stresses σ_{22}^k $(k = 1,2)$ suffer a break, taking different values at the insertion points and at the points of the quarter plane. At the exit from the contact area, the normal voltages σ_{22}^2again take on a local extreme value. It should be noted that in the contact region, the value of the extremum of the normal voltage σ_{22}^1 (in the Section 2') is almost twice as high as its value in the material of the quarter plane. This result is associated with a greater rigidity of the insert material.

The evolutions of the wave pattern of tangential stresses σ_{12}^k $(k = 1,2)$ at times t=40τ and t=60τ are shown in Fig. 6 and 7.

The isolines constructed in the figures are in good agreement with the nature of the change in pulse loads specified at the boundary with a quarter of the plane. At time t=40τ (Fig. 6), the leading edge of the boundary wave reaches the middle of the insert (x_1=0 axis), i.e., the wave has passed the distance x_2=20h with some delay due to the weakening of elastic perturbations during the transition to the insert material. This moment of time t=40τ corresponds to the action

time of the first pulse of the boundary load. The isolated traces of extensive single extremums in the entire region $D_1 \cap D_2$ of an inhomogeneous medium are explained by the influence of angular points, points of discontinuity of boundary conditions and contact surfaces.

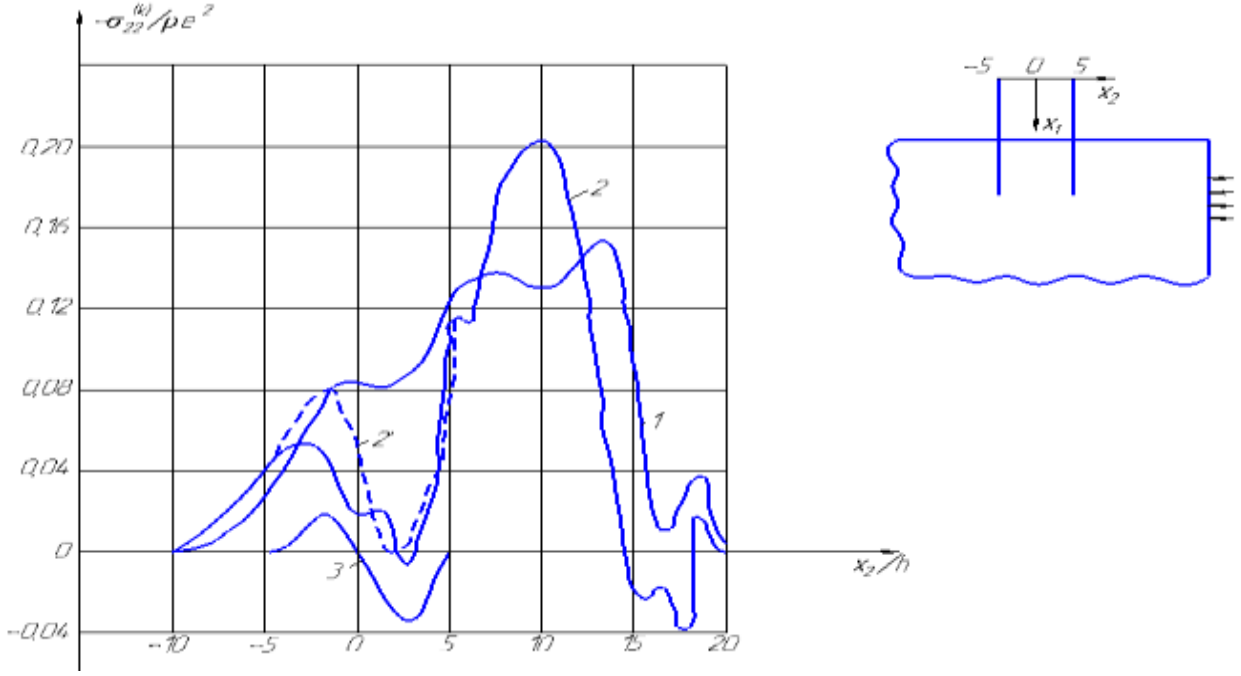


Fig. 5. Change of Normal Stresses σ_{22}^k at Time t=60τ along the x_2 Coordinate in Sections $1 - x_1 = 5h$, $2 - x_2 = 10h$, $2' - x_1 = 10h$, $3 - x_1 = 0h$.

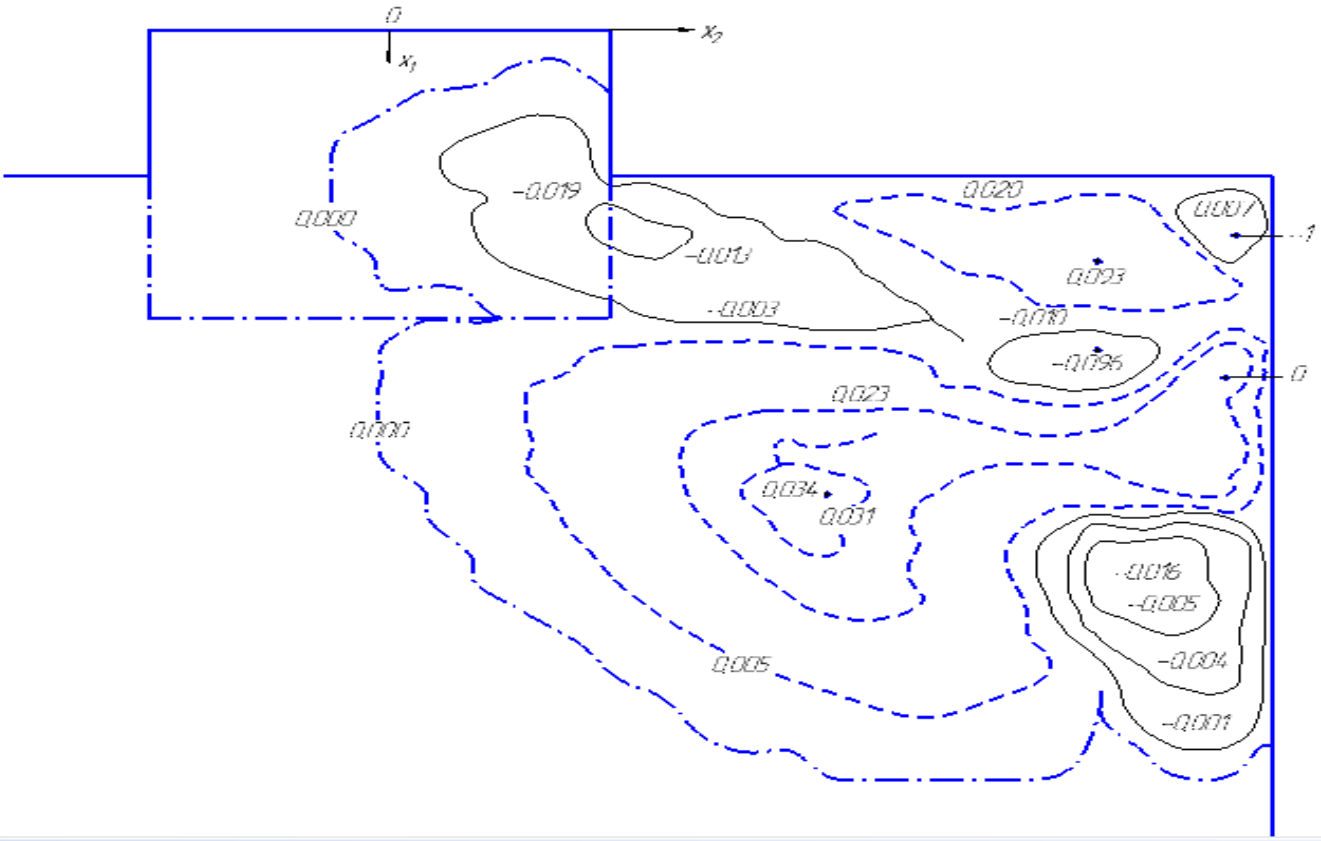


Fig. 6. Tangential Stress Isolines σ_{12}^k $(k = 1,2)$ at Time t=40τ.

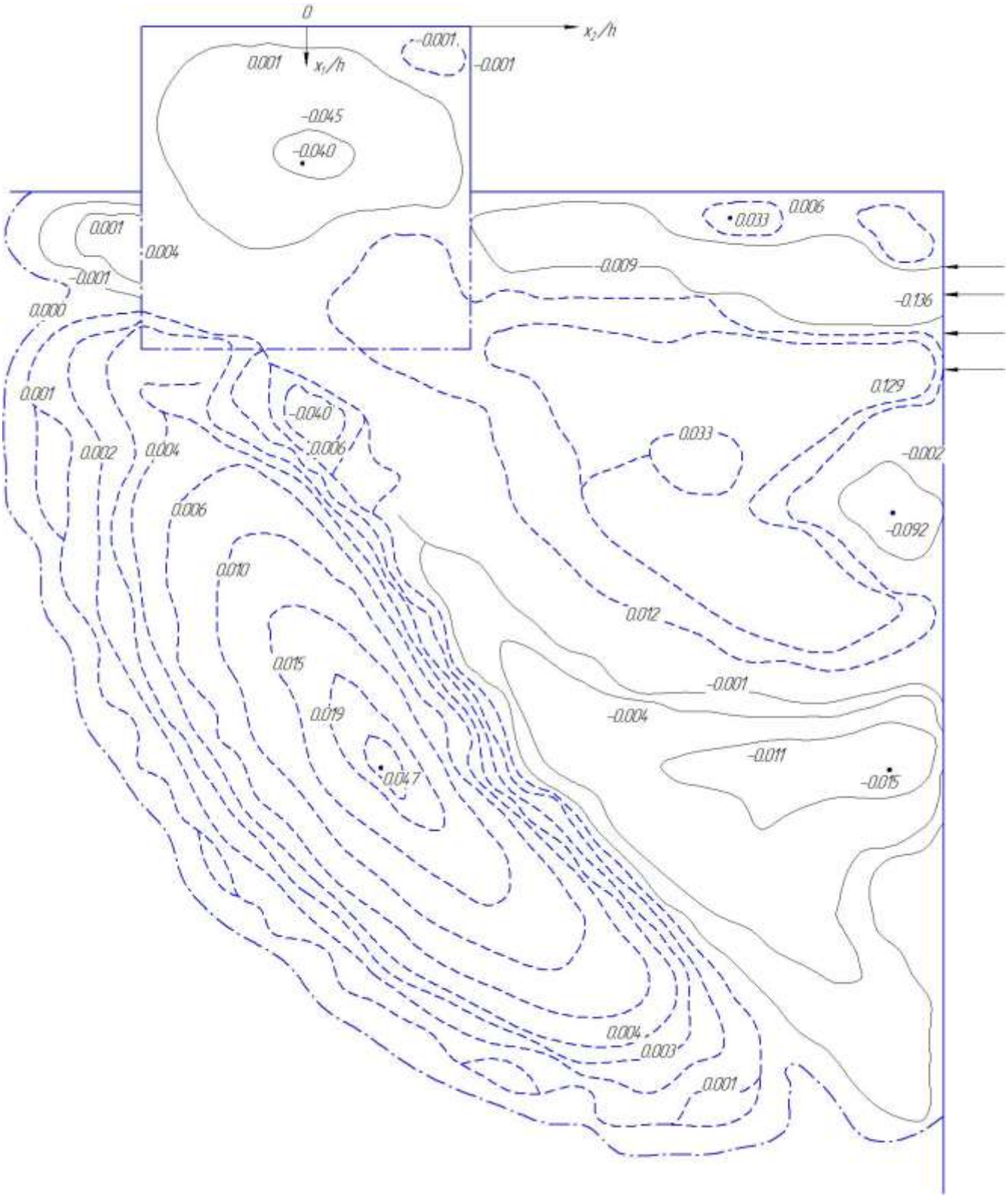


Fig. 7. Tangential Stress Isolines σ_{12}^{k} $(k = 1,2)$ at Time t=60τ.

In Fig. 7, the tangential stresses σ_{12}^{k} $(k = 1,2)$ reach their maximum values in absolute magnitude near the load application points $(N_1 \le x_1 \le N_2, x_2 = L_3;;$ line $B_1B_2)$. At this moment of time t=60τ, the second pulse is affected. This can explain the appearance of the selected double zones of stress extremes in the entire region of the inhomogeneous medium $D_1 \cap D_2$.

V. Conclusion

The developed information system made it possible to visualize the results of calculating wave propagation during blasting operations, to obtain graphs of changes in normal stresses and tangential stress isoclines [26].

Scientific and theoretical interest in the study of the relationship between stress (stress tensor) and deformations occurring in the medium as a result of blasting remains very relevant. The problems of seismically unstable areas require consideration of this issue from the point of view of modeling the wave process.

Modeling the reaction of the soil to changes resulting from an explosion, using information and communication technologies, allows solving problems not only of an economic, construction, geological nature, but also of the safety of human life.

References

[1] Semenova V. Y. modeling of the response of soil under seismic mikroraionirovanii construction sites // Geophysical journal. — 2015. — V. 37, № 1. — P. 137-153.

[2] Juzbaev S. S., Bayaliev B. T., Mortalin N. To. the calculation of the strain state of the system "structure-soil" under the influence of energy strombolian explosions // abstracts of the Republican scientific-practical conference on problems of construction on the landing grounds in southern Kazakhstan. Shymkent, 1991, p. 60-68.

[3] Sabitova D.S., Zhuzbayev S.S., Juzbayeva B.G. Development of an information analysis system for analyzing wave processes in a homogeneous medium // International Journal of Mechanical Engineering and Technology. – 2018. –Vol. 9(12). – P. 499-508.

[4] Akhmetova Z., Zhuzbaev S., Boranbayev S., Sarsenov B. Development of the System with Component for the Numerical Calculation and Visualization of Non-Stationary Waves Propagation in Solids. IOS Book Series: Frontiers in Artificial Intelligence and Applications (FAIA), 2016, Volume293, pp. 353-359, ISSN 0922-6389.

[5] Clifton R. Difference method in plane problems of dynamic elasticity // Mechanics. Collection of translations. 1968. - No. 1. - pp. 103-122.

[6] Tarabrin G. T. Difference schemes of wave problems of elasticity theory: monograph / G. T. Tarabrin; VolgSTU. - Volgograd: RPK "Polytechnic", 2000. - 148 p.

[7] Kukudzhanov V.N. Numerical Continuum Mechanics. De Gruyter, 2012. XVIII, 425 pages.

[8] Reker V. V. Applied mechanics// Series E.-1970. - No. 1.-B. 121-129.

[9] Italien S. M., Masanov, J. K., Baimahanov I. B., N. Makhmetova.M.//Numerical methods for solving problems in solid mechanics, Karaganda, 1987. pp. 3-15.

[10] Bazhenov V. G., gonik E. G., Kibar A. I., D. V. Shoshin Stability and limit state elastic-plastic spherical shells under static and dynamic loadings. Applied Mechanics and Technical Physics, 2014. T. 55, No. 1, pp. 13-22.

[11] Aitaliev Sh. M., Alekseeva L. A., Dildabaev Sh. A., Zhanbyrbaev N. B. The method of boundary integral equations in problems of dynamics of elastic multi-connected bodies. - Alma-Ata: Gylym, 1992. - p. 228.

[12] S. Tleukenov, A. Bobeev, D. Sabitova. Structure of the matriciant for systems of ordinary differential equations of first order and its applications. International journal of applied mathematics & statistics. – 2018. – V.57 (1). – P. 209-217.

[13] Rysbaiuly B., Yunicheva N., Rysbayeva N. An iterative method to calculate the thermal characteristics of the rock mass with inaccurate initial data. Open Engineering, 2016, 6(1).

[14] Zhuzbaev S. S., Sabitova D. S. Computer modeling of elastic wave propagation in a homogeneous medium. BULLETIN OF THE PSU. Physics and Mathematics Series, 2018, No. 1, pp. 68-82. ISSN: 1811-1807.

[15] Akhmetova Z., Zhuzbaev S., Boranbayev S. The method and software for the solution of dynamic waves propagation problem in elastic medium. Acta Physica Polonica A, Polish Academy of Sciences. -2016, Vol.130. - pp. 352-354, ISSN 0587-4246.

[16] Zhuzbayev S., Sabitova D., Sarsenov B. Computer mathematical modeling of wave processes // 4th intern. conf. on computer and technology applications (ICCTA). – Istanbul, 2018. – P. 1-5.

[17] Zhuzbayev S., Sabitova D. Mathematical modeling of elastic wave propagation in a homogeneous media // Вестник Восточно-Казахстанского государственного технического университета им. Д. Серикбаева. – 2018. – Т. 1, №3, ч. 3. – С. 78-90.

[18] Zhuzbayev S., Adilova A., Akhmetzhanova S., Juzbayeva B., Sabitova D. Design of composite materials using information technology. Journal of Theoretical and Applied Information Technology, 2020, 98(18), c. 3698-3711.

[19] Ashirbayev N. K. Features of the propagation of dynamic perturbations in bodies with inhomogeneities. / / Dissertation work for the degree of Candidate of Physical and Mathematical Sciences, Shymkent, 1986, 207 p.

[20] Tarabrin G. T. Numerical solution of nonstationary problems of dynamics of an anisotropic elastic medium. / / Moscow, Izv. of the USSR Academy of Sciences, Mechanics of a solid body, 1982, No. 2, pp. 83-95.

[21] Y. Ru, G.F. Wang, T.J. Wang, “Diffractions of elastic waves and stress concentration near a cylindrical nano-inclusion incorporating surface effect,” Journal of Vibration and Acoustics, vol. 131, 2009: Art. 061011.

[22] J.-Y. Kim, “On the generalized self-consistent model for elastic wave propagation in composite materials,” International Journal of Solids and Structures, 2004, pp. 4349-4360.

[23] L. Ya. Cosachevskiy, “On propagation of elastic waves in two-component media”, Applied Mathematics and Mechanics, vol. 23, 1959, pp. 1115-1123.

[24] S.K.Serikbayeva, D.A.Tussupov, M.A.Sambetbayeva, A.S. Yerimbetova, Taszhurekova ZH.K., Borankulova G.S. EduDIS construction technology based on Z39.50 protocol: Journal of Theoretical and Applied Information Technology, 31st May 2021 г.. - Т. - Vol. 99. No. 10, P.2224-2255.

[25] Serikbayeva S.K, Batyrhanov A.G., Sambetbayeva M.A., Sadirmekova Zh.B., Yerimbetova A.S. Development of technology to support large information storage and organization of reduced user access to this information: (IJACSA) International Journal of Advanced Computer Science and Applications, 2021 г.. - 7 : Т. 12. - стр. 493-503.

[26] S.K. Kanaun, V.M. Levin, F.J. Sabina, “Propagation of elastic waves in composites with random set of spherical inclusions (effective medium approach),” Wave Motion, vol 40(1), 2004, pp. 69-88.

Evaluation Study of Elliptic Curve Cryptography Scalar Multiplication on Raspberry Pi4

Fatimah Alkhudhayr[1], Tarek Moulahi[2]
Department of Information Technology
College of Computer, Qassim University
Buraydah, Saudi Arabia

Abdulatif Alabdulatif[3]
Department of Computer Science
College of Computer, Qassim University
Buraydah, Saudi Arabia

***Abstract*—The internet of things (IoT) is defined as a collection of autonomous devices that connect and network with each other via the Internet without the requirement for human interaction. It enhances daily our lives such as through personal devices, healthcare sensing, retail sensing, and industrial control, as well as the smart homes, smart cities, and smart supply chains. Although the IoT offers significant benefits, it has inherent issues, including security and privacy risks, memory size limitations, and processing capability challenges. This paper describes the application of elliptic curve cryptography (ECC) in a simulated IoT environment to ensure the confidentiality of data passed between the connected devices. Scalar multiplication represents the main operation of ECC, and it is primarily used for key generation, encryption, and decryption. The aim of this paper is to evaluate and show the efficiency of adapt lightweight ECC with an IoT devices. In the study outlined in this paper, scalar multiplication was implemented on Raspberry Pi4 and processing time and consumed energy were measured to compare the performance. The comparison was made on the scalar multiplication of both fast and basic ECC algorithms. The result of the performance test revealed that a fast scalar multiplication reduced the computation time in comparison with basic scalar multiplication while consuming a similar level of energy.**

***Keywords*—IoT; elliptic curve cryptography; fast scalar multiplication; raspberry Pi4**

I. Introduction

Since its conception by Kevin Ashton in 1999, the increasing popularity of the internet of things (IoT) has led to rapid changes in fields as varied as lifestyles, standards, and business models. The IoT refers to the connection of autonomous devices to the internet with the capacity to transmit data via a network without human intervention. IoT devices range from small accessories to large machines, including smartphones, tablets, laptops, personal computers, and similar portable embedded devices [1]. The IoT is not one single technology, but an agglomeration of technologies, by which embedded sensors, actuators, processors, and transceivers of connected devices comprise the IoT [2]. The communication system facilitating the communication between IoT devices can be based on sensors or wireless technologies, further enabling the transfer of data to a centralized system following processing [1].

As IoT devices imply a constant internet connection, privacy, and security issues are paramount. For example, it was demonstrated that 70 % of IoT devices are unable to resist attacks [1]. This underlines the need for security mechanisms that can ensure IoT security, e.g., in terms of access control, authentication, data integrity, confidentiality, and secrecy as well as protecting connected devices from attack.

Conventional security mechanisms and protocols designed to protect computers against cyberattacks are not appropriate for use with the IoT, primarily as the connected devices have insufficient memory size and processing capability. Hence, efficiently protecting such low-resource devices requires the consideration of other security protocols, with cryptography offering a suitable solution [3].

Cryptography refers to the encryption and decryption process, i.e., converting plain text (readable form) into ciphertext (encoded form) and vice versa, respectively using cryptographic algorithms. These algorithms can be symmetric or asymmetric: If the same key is used for both the encryption and decryption, then it is a symmetric encryption, while using a public key to encrypt and a private key to decrypt is asymmetric encryption. Cryptography strengthens computational security by making the cost of breaking the encryption exceed the value of the information that is encrypted or making the breaking time exceed the information's useful lifetime [4].

Elliptic Curve Cryptography (ECC), proposed by Miller and Koblitz in 1985, is among the most popular cryptography protocols [5]. It is like the Rivest–Shamir–Adleman (RSA) public-key cryptosystem in terms of the security level, although it has a smaller key size. The security strength of ECC relies on the Elliptic Curve Discrete Logarithm Problem (ECDLP) difficulty, which includes point doubling and adding operations, making it more computationally efficient than RSA exponentiation. Furthermore, as it consumes less memory, ECC leads to reduced performance costs and computational costs [6].

The paper explores the use of ECC to enhance data security in the IoT. It hereby aims to show that ECC is applicable for the IoT due to its efficiency and performance regarding time and energy. In achieving this aim, the following motivations are considered:

- Highlight previous studies that concern with lightweight ECC to shows the importance of an ECC and evaluate the efficient technique.

- The importance of adapting a security mechanism to the IoT environment to allow it to operate in the fastest possible way.
- The application of fast scalar multiplication to the finite field of ECC, presenting a highly suitable technique for embedded devices in the context of the IoT.
- Deduce the gap present in the literature since, few studies concern with applied fast scalar multiplication alongside with IoT devices.

The rest of this paper is organized as follows. Section II discusses similar prior works in Internet of Things' security and Elliptic Curve Cryptography. Section III describes the methodology study, while the experimental study is explained in Section IV. Section V presents the results and discusses it. Section VI concludes the paper and section VII points out our research contribution to the future work.

II. Related Work

This section mainly highlights related works on the security of Internet of Things as well as application of an Elliptic Curve Cryptography. These two subsections are considered in this paper due to its applications in many aspects in real life, in addition to its security and privacy aspects.

A. IoT

The growth of the IoT provides opportunities to enhance many aspects of our lives, such as through personal devices, healthcare sensing, retail sensing, and industrial control, as well as the smart homes, smart cities, and smart supply chains [7]. The popularity of smart home technology is made possible by the development of sensors and actuators that can be utilized in a wireless sensor network. At the same time, people have become more comfortable with and trusting toward technology, allowing companies to overcome concerns by offering benefits to the security and quality of life. Smart homes require sensors to provide intelligent services to the user. Their incorporation into domestic environments can assist with many aspects of daily living, such as by automating tasks, saving energy, and enhancing security. However, smart home technology presents issues and challenges, among which security and privacy remain the most pressing and problematic because data are recorded regarding many activities around the home. Systems must be safeguarded from attack [2].

In [8], Ayoub et al. proposed a lightweight secure scheme for IoT objects and cloud computing. Their recommendations depend on ECC and message queuing telemetry transport (MQTT), for which the key attributes are publishing and subscription. The driving factor behind their techniques is the provision of secure interactions between the IoT and cloud computing, along with enhanced communication speeds. Their security procedures consist of initialization, sub- scription, and publication, and they validated the scheme's performance by comparing it with TLS/SSL. They automatically verified the protocol's safety using the Automated Validation of Internet Security Protocols and Applications (AVISP) tool.

In [9], Sha et al. analyzed numerous issues and security challenges presented by IoT systems and found that the IoT faces more issues than wireless sensor networks (WSNs). They identified the security architecture factors of end-to- end security, edge computing-based designs, and distributed security models, discussed their benefits and restrictions, and provided examples of implementations of each design. They found that to achieve comprehensive security for IoT systems, capable low-end devices must be supported from higher up in the command structure.

In [10], Hossain et al. proposed a technique to ensure quality end-to end security for IoT systems based on biometrics and cryptography. They depicted an infrastructure of biometric-based end to end security solution for IoT with four layers: device, communication, cloud, and application. They discussed the security challenges and possible solutions for each layer and determined that their proposal based on the biometrics of facial recognition was 99 % effective by comparing face recognition in a local server and in cloud-server. Consequently, they demonstrated that to use biometrics for authentication ensures IoT system security to a greater extent than password authentication.

B. ECC

In terms of lightweight Elliptic curve cryptography various studies and experiments conducted to evaluate the performance of lightweight ECC in terms of efficiency and security with different techniques, for both hardware and software implementations. This section highlights several studies that involve for enhance scalar multiplication operation of an elliptic curve in different environments.

Firstly, number of studies have applied a parallel implantation technique for speed up scalar multiplication. In [11], Yanbo Shou et al. applied ECC cryptography to network security. To optimize the performance of scalar multiplication, which is the most expensive ECC operation, they applied it in parallel via distributed tasks that were split into neighbor nodes and executed simultaneously. Due to the required energy consumption, they found that parallel computing is only suitable when execution time is the critical factor.

Another study applied parallel technique provided by Albahri et al. [12] proposed a new algorithm that enable parallel implementation of ECC on multi-core platforms by modified algorithms that overcome data dependencies in ECC computation. Their work aims to explore the efficiency of parallel implementation of ECC as well as enhance point multiplication operation on ECC. Their proposed modifications based on two novel algorithm modifications for performing ECC point multiplication. They perform a vertical parallelization for operations of point doubling and point additions, which is they first modification. It depends on perform multiple finite field operations with no data dependency by different parallel logic cores. Their second modification to remove data dependencies by modifying the left to right double and add binary point multiplication. They implement modified algorithms with pure software implementation for ECC scalar multiplication over GF(2^{163}) using Xmos multi-core microcontroller, the result of their proposed multi-core implementation to enhance operations of point multiplication and point addition up to 60% and around 50%, respectively. Finally, their experiments show the feasibility and efficiency of adapting parallelism in ECC implementation.

In [13], Faye et al. proposed an approach to improve ECC performance for WSN capabilities, aiming to accelerate ECC scalar multiplication over primary fields, as well as avoid the storage of precomputation point by accelerating computation. They firstly proposed a new technique based on the negative of point and a point order to run fast computation of scalar multiplication. Secondly, they accelerate computation in parallel scalar multiplication on KP to avoid storage of precomputation by proposed an efficient algorithm depend on improvement of the double and add (DA) and quadrable-ana-quadrable algorithms. Finally, they showed that their proposed algorithm accelerates the computation of scalar multiplication on NIST-192 parameters for ECC. As well as they showed their technique for avoid storage computation very efficient especially for Jacobian coordinates.

As well several studies applied various technique to improve scalar multiplication operation. To confirm complete encryption for users of Online Social Networks (OSN), in [14], Rajam and Kumar proposed improved elliptic curve cryptography (IECC) to confirm complete encryption for users of online social networks (OSNs) and compared it with standard ECC. The algorithm requires the replacement of each repetitive text character with different ciphertext. Time, size, and security were used to evaluate their proposed algorithm and standard ECC based on time of key generation, time of encryption, time of decryption, and size of plaintext compared to ciphertext. They found that IECC performed better than the standard ECC.

In [15], Kalra and Dhillon conducted a comparative study to highlight viable solutions to security issues of real-time embedded systems. The characteristics of embedded systems, security issues, and threat models were presented. Public key cryptography (PKC) is the main challenge to security implementation in embedded systems; hence, efficient PKC solutions require an ECC that uses smaller key sizes. Suitable solutions suggested by this analysis of resource- constrained real-time embedded systems were all ECC-related.

In [16], Mingquan Hong et al. proposed an encryption scheme that uses ECC-based homomorphic encryption to solve the problem of secure multiparty computation (SMC). They compared their scheme's performance with RSA and Paillier homomorphic encryption, particularly in the context of computation time and communication. ECC was found better than either of these options, and they applied their scheme to GPS earthquake data to show that it is efficient and provides high security.

In [6], Arora and Chhabra proposed a security scheme to prevent eavesdropping attacks in the Cloud environment. They compared the ECC-based scheme's performance, such as the time taken to process encryption and decryption data, with that of traditional RSA schemes, and found it much faster.

In [17], Javed R. Shaikh et al. focused on implementing ECC in resource- constrained e-commerce applications. They analyzed a set of selected curves recommended by several sources to find an efficient option for constrained resources. The elliptic-curve Diffie–Hellman (ECDH) algorithm and elliptic curve digital signature algorithm (ECDSA) were applied to selected curves. They found that SECP256r1 and M221 can be used to implement ECDH and ECDSA algorithms, respectively, offering suitable curves for e-commerce applications.

Additionally, Liu et al. [18] optimized an efficient scalar multiplication algorithm for wireless sensor networks based on symmetric ternary. Their optimization depends on eliminate the modulo inversion operation in the fundamental operation by using a Jacobi coordinate. As well as a nonzero weight reduced which can reduce the operation of point addition by optimized the symmetric ternary representation of the positive integer. So that, the efficiency of the scalar multiplication algorithm is improved. The basic idea of the symmetric ternary scalar multiplication model is computing the calculation of the scalar multiplication by calling the operation of point tripling and point addition constantly. By applying symmetric ternary scalar multiplication based on Jacobi coordinate their study indicates that it is better than the scalar multiplication based on affine coordinate. They found that it has an efficiency improvement of 4.3%.

Recognizing the importance of security for mobile devices, in [19] Mullai and Mani focused on enhancing crucial aspects and optimizing the cryptography operations of RSA and ECC. To suit these two algorithms, they proposed generating addition chains (ACs) based on particle swarm optimization (PSO) and simplified swarm optimization (SSO) before measuring performance using mobile emulators of Android and Windows. The two algorithms were compared based on time of processing, power consumption for encryption and decryption, and level of security. They observed that, when considering the security aspects of SSO-optimized AC, ECC provides more security, although RSA consumes less power.

Lastly, Javeed et al. [20] proposed hardware architecture with high performance for elliptic curve scalar multiplication over prime field. It depends on parallel technique, which can either execute modular addition or modular subtraction in parallel to four modular multiplication operations. Totally, it executed five mathematic instructions in parallel. Then, presented the scalar multiplication using Jacobian coordinate. Finally, it implemented on Xilinx Virtex-6, Virtex-5 and Virtex-4 FPGA (field-programmable gate array) platforms. Their tested shows that, for one operation of elliptic curve scalar multiplication with 256-bit it takes 2.01 ms, 2.62 ms and 3.91 ms for three platforms respectively. Significantly, it is 1.96 times faster and it is applicable for any value of prime p less than 256 bits.

From above presented studies, we can observe that an improved elliptic curve cryptography has gained attention of the researchers for various platforms or environments such as WSN either for implemented in hardware or software. As well, technique applied for accelerate scalar multiplication divers some depends on parallel or sequential. Also, to the bast of our knowledge, few studies have applied a lightweight cryptography especially in term of fast scalar multiplication in an IoT devices. That encourage the needed for further study and research in this field. So, in this paper a fast scalar multiplication is applied in an IoT device with various key sizes to evaluate scalar multiplication in terms of energy and running time.

III. Contribution Theoretical Study

The following subsections explain and discuss the main ideas behind IoT scenarios in these contexts, together with ECC methods and fast scalar multiplication.

A. Architecture of IoT

The main components of IoT are presented by the three-layered architecture shown in Fig. 1. These factors can be considered as the most basic architecture [2], consisting of the three layers of perception, network, and application.

- The perception, or physical layer, is responsible for sending and receiving environmental information via sensors.
- The network layer connects with smart applications, other network de- vices, and servers. It is also used to transmit and process sensor data.
- The application layer interacts with and delivers application services to the user.

B. IoT Scenario

Smart home systems, or automated homes, allow the remote control and operation of electrical devices via electronic devices, such as smartphones or laptops, that have applications with user-friendly interfaces. A related form, the intelligent home, acts based on predefined information [21].

In our scenario, the smart home could be an IoT virtual environment, as presented in Fig. 2. Smartphone applications can interact with various applications, making them easier to control and more efficient. User registration is required for security reasons, but notifications to control and interact can be conveniently sent by email or text messages. The technology consists of wireless sensor nodes, such as for lighting, temperature, motions, and cameras that provide connections to gather and send data between users and applications via a base station that functions as a gateway. In our experiments, we used Raspberry Pi 4 for this gateway, which allowed communication with the private cloud or a specific database. For example, to check whether a light is on, the user accesses the appropriate application and chooses the light option. The gateway will allow the light sensor to send information to the user, who can then read the data, via email or text message. This is a typical example, but what if the lighting sensor reads and sends the wrong data? Such an error can cause the system to work in abnormal or malicious ways.

C. ECC Cryptosystem

Like RSA, ECC is a public key cryptosystem. However, ECC security depends on interpreting logarithm problems, e.g., how to determine K given KP and P. Table I compares key differences between ECC and RSA in terms of computational effort for cryptanalysis [4]. Algorithms of key generation, encryption, and decryption are presented in Algorithm 1, Algorithm 2 and Algorithm 3, respectively [22].

Key generation step is the most important step in which an algorithm is used to generate both public and private keys. The parameters in the key generation algorithm defined as:

- E : elliptic curve defined over finite field Fp.
- P: point on the curve that has prime order n.
- Equation of elliptic curve, prime, point on the curve with its order n are all public domain parameters denoted by E, p, P respectively.
- d: private key selected randomly from interval [1, n-1].
- Q = dP : it corresponding public key.

And parameters in the encryption and decryption algorithms identified as:

- M: point that represent a message (plain text m).
- k: randomly selected integer between range [1,n-1].
- Q: public key's recipients.
- d: private key's recipients.
- B1, B2: two points that represent cipher text.

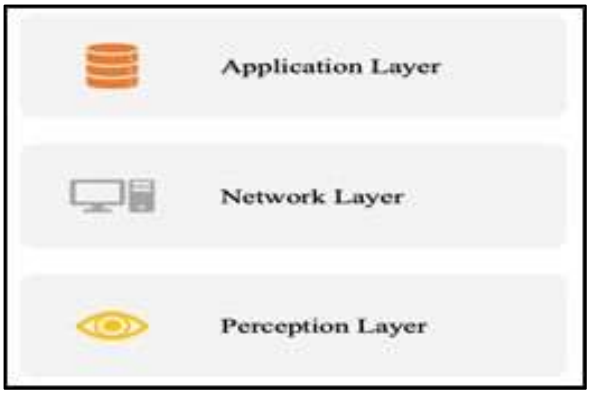


Fig. 1. IoT Architecture.

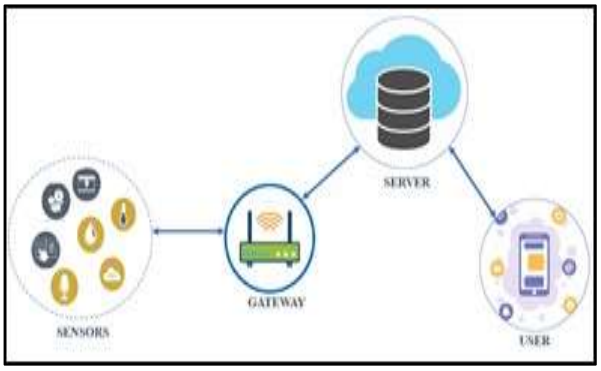


Fig. 2. Smart Home Environment.

TABLE I. Key Size Comparison RSA Vs. ECC

RSA	ECC
1024	160
2048	224
3072	256
7680	384
15360	521

Algorithm 1: Key Generation algorithm

1: begin

2: Input variables E, p, P

3: Set private Key $1 < d < n - 1$

4: Compute Public Key($Q = dp$)

5: Get (Q, d)

6: end

Algorithm 2: Encryption Algorithm

1: begin

2: Input variables E, P , p , n, m

3: Set $m \leftarrow M$

4: Set $1 < k < n - 1$

5: Compute Cipher text1(kP)

6: Compute Cipher text2(M + kQ)

7: Return (Cipher text1,Cipher text2)

8: end

Algorithm 3: Decryption Algorithm

1: begin

2: Input variables p, E, P, n, d, Cipher text1, Cipher text2

3: Compute m1(d ∗ Cipher text1)

4: Compute M (Cipher text2–m1)

5: Return M

6: end

D. Fast Scalar Multiplication

We applied the algorithm proposed by Faye et al. [13] to accelerate scalar multiplication KP over a finite field (Fp). This improvement concerns the negative of a point and a particular reduction of the scalar in a selected interval. A well-known cryptanalysis trick is to use negation, which is also used in cryptography, for scalar multiplication with addition-subtraction chains.

The algorithm of fast multiplication depends on the replacement of the KP point of the scalar multiplication operation with an equivalent representation point SP, where s and k are scalars and $k > s$. This technique is used in the interval $[\ln/2J + 1, n - 1]$, where ln/2J denotes the integer-part of n/2. The negative of a point can be used for fast computation because it is freely obtained. The point KP is replaced with an equivalent point representation SP by utilizing the negative point because, for each point P on an elliptic curve, the point - P is also on the curve. Given that point P = (xp, yp) in affine coordinates, to calculate the inverse of a point KP = (xkp, ykp), we can compute KP = (xkp, ykp) then we can change the sign on the y-coordinate (ykp).

Thus, we have the following equations for a secret key (integer number) by point KP to obtain an equivalent point representation of SP:

$$\text{If} \quad K > n, \quad Kp = Sp \quad \text{where} \quad S = (k - \lfloor k/2J.n) \tag{1}$$

$$\text{If} \quad K \in]\lfloor n/2J, n - 1], \quad kP = SP \quad \text{where } S = (k - n) \tag{2}$$

$$\text{If } K \in]0, \lfloor n/2J], \quad kP = SP \quad \text{where} \quad S = K \tag{3}$$

$$\text{If } K = n \quad \text{or } 0 \quad \text{or} - n\,, \; kP = \infty \tag{4}$$

$$\text{If} \quad k \in [-(n - 1), -\lfloor n/2J[, \; kP = SP \; \text{ where } \; S = (n + k) \tag{5}$$

$$\text{If } k \in [-\lfloor n/2J, 0[, \quad KP = SP \qquad \text{where} \quad S = K \tag{6}$$

$$\text{If } k < -n, \; KP = SP \qquad \text{where } \; S = k + n \; .\lfloor |k|/2J \tag{7}$$

IV. Experiment

An experiment was conducted to evaluate and compare the operational performance of point multiplication and fast multiplication in ECC. We explain the experimental setup and procedure.

A. Experimental Setup

The experiment was conducted in a laboratory environment, as shown in Fig. 3. The C programming language was used to develop a program that ran on a Raspberry Pi, and was used to simulate an IoT environment. The program performed a scalar multiplication KP for both a basic and a lightweight ECC. Subsequently, it measured the computational time of the scalar multiplication via a built-in library in C with a time() function. The consumed energy was computed using a digital voltage power meter. The time and energy data were recorded to evaluate and compare the performance of both the basic and fast scalar multiplication algorithms.

B. Raspberry pi

An IoT environment was created with a Raspberry Pi 4 Model B, which is the newest and fastest Raspberry product. It comes with various amounts of RAM (2, 4, or 8 GB), has a USB-C port for power, requires a MicroSD card to store all files and the operating system, has two micro HDMI ports, and offers the user the choice to connect to the internet via Ethernet cable or wirelessly. Table II shows the characteristics of the Raspberry Pi 4 Model B used in this experiment.

Fig. 3. Experimental Setup.

TABLE II. Raspberry Pi 4 Model B Specifications

Operating System	Raspbian OS
Card size	8 GB
Internet Connectivity	Wi-Fi

C. Experimental Procedure

A controlled laboratory environment was used as the setting of the experiment to avoid a disruption of the internet connection. Three main key sizes (K) were tested to evaluate the performance of the scalar multiplication KP. The key sizes were 32, 64, and 128 bits, representing the first, second, and third cases, respectively. For all three cases, the time was measured in seconds, and the energy consumption was computed in Watts.

V. Result and Discussion

The experimental analysis compared the performance of the standard scalar multiplication KP with that of the fast scalar multiplication SP based on the evaluation criteria of time and energy (in seconds and Watts, respectively). The evaluation was verified for three key sizes (32, 64, 128 bits). To compare basic and fast scalar multiplication in terms of time and energy consumption, the experiment was conducted multiple times for each key size using random numeric data, and the ratio for each key size was calculated.

The following subsections will discuss the result according to two aspects time and energy.

A. Time Consumption

Following Fig. 4, 5, and 6 present the time consumption for each key size. For the 32-bit key size, the times for both the basic and fast algorithms were similar at the beginning; however, at the end, the basic algorithm consumed more time, increasing by approximately 90 % compared with fast scalar algorithm, as shown in Fig.4. Second, for both the 64- and 128-bit key sizes, the fast algorithm consumed less than one second, while the basic algorithm consumed more time, as shown in Fig. 5 and Fig. 6, respectively.

Finally, as a result, at almost a half-second, fast scalar multiplication used less time than basic scalar multiplication for all three key sizes. The basic scalar multiplication needed more time for the three key sizes of 32, 64, and 128 bits, at 9.1, 98.7, and 216.1 seconds, respectively. As shown in Fig. 7, fast scalar multiplication running faster than basic scalar multiplication approximately 99%. In addition, the running time demonstrated a direct relationship with the received data, i.e., more time for more data.

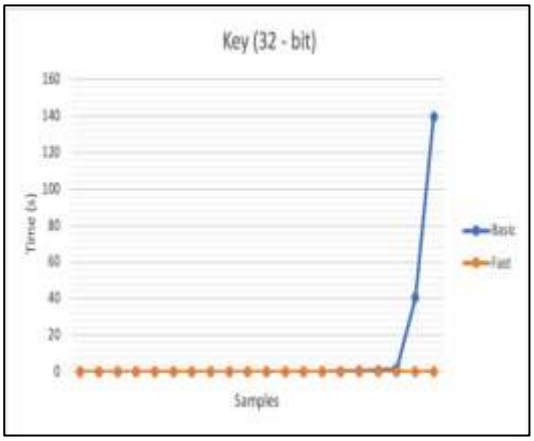


Fig. 4. Basic Running Time vs. Fast (Case1).

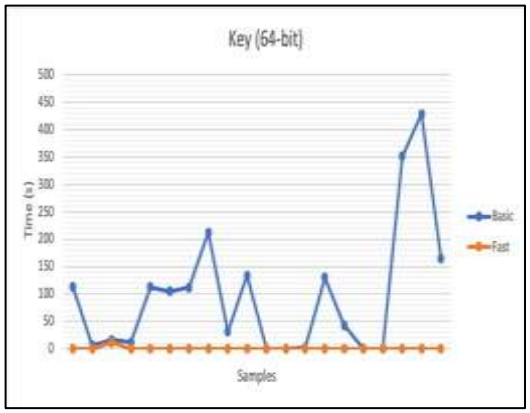


Fig. 5. Basic Running Time vs. Fast (Case2).

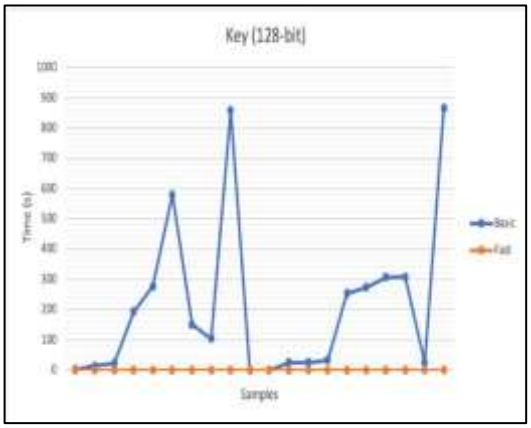


Fig. 6. Basic Running Time vs. Fast (Case3).

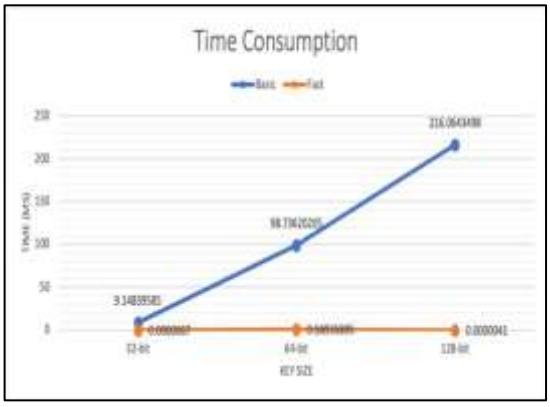


Fig. 7. Running Time Average for all Three Cases.

B. Energy Consumption

The energy consumption for each key size is shown in the following Fig. 8, 9, and 10. The basic algorithm consumed more energy than the fast algorithm when the key size was 32 bits, as shown in Fig. 8. For key sizes of 64 and 128 bits, as shown in Fig. 9 and Fig. 10, respectively, similar amounts of energy were consumed by fast and basic scalar multiplication.

Finally, as shown in Fig. 11, the fast algorithm uses less energy than the basic algorithm, by about 10 % in the case of the smaller key size. For the larger key sizes, the basic algorithm used about 0.5 % less energy than the fast algorithm.

To conclude, the fast scalar multiplication algorithm saves more time than the basic scalar multiplication algorithm. However, basic scalar multiplication and fast scalar multiplication consumed similar amounts of energy.

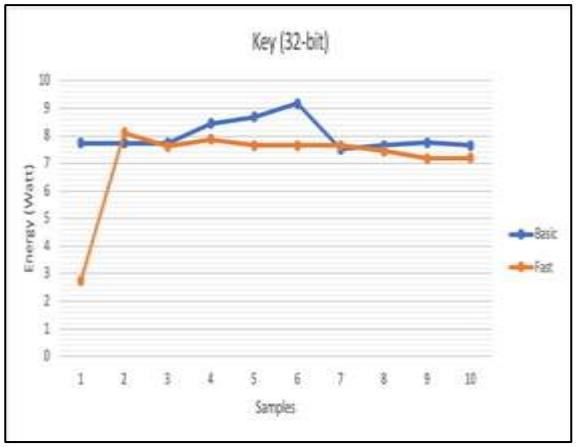


Fig. 8. Basic Energy Consumption vs. Fast (Case1).

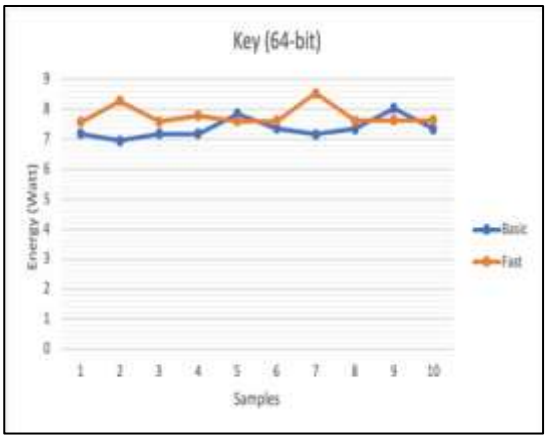


Fig. 9. Basic Energy Consumption vs. Fast (Case2).

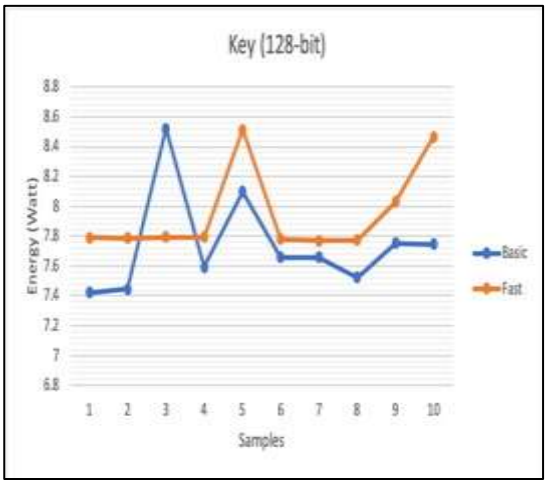


Fig. 10. Basic Energy Consumption vs. Fast (Case3).

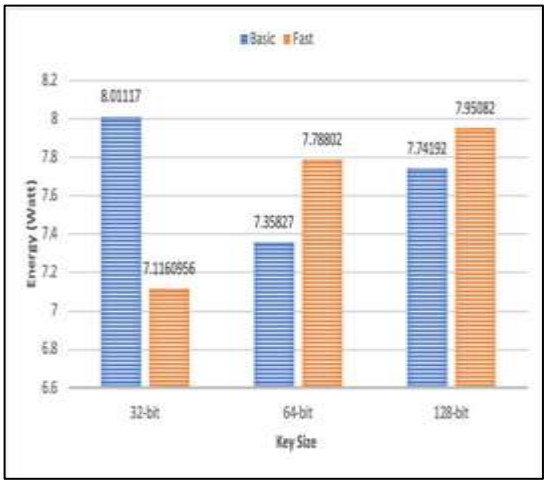


Fig. 11. Energy Consumption Average for all Three Cases.

VI. Conclusion

The increasing integration of IoT technology in various domains of daily life presents challenges regarding security, privacy, and cost of performance. Against this backdrop, this study focused on ECC in general, and on scalar multiplication methods in particulars. We presented various studies examining security solutions in the context of the IoT as well as enhancing the performance of ECC. As a time- and cost-effective security mechanism is crucial, our experiment examined and compared the performance of basic and fast scalar multiplication to show the efficiency and applicable of fast scalar multiplication in the embedded devices. The results revealed that fast scalar multiplication saves time for three key sizes with similar energy usage. The main contribution is that we demonstrated the fast scalar multiplication is faster than basic scalar multiplication around 99%. That indicates fast scalar multiplication is a suitable solution for embedded devices to reduce the cost of performance.

VII. Future Work

In future work, we plan to reduce the energy consumption by optimizing fast scalar multiplication. Furthermore, additional lightweight ECC can be examined and evaluated in terms of their performance. Moreover, other evaluation metrics can be adopted. Additionally, homomorphic encryption can be applied and compared with obtained result. Finally, studying and applying this technique to other devices offers an interesting research avenue.

References

[1] M. Abdur Razzaq, M. Ali Qureshi, S. Habib Gill, and S. Ullah, "Security Issues in the Internet of Things (IoT): A Comprehensive Study," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 6, pp. 383–388, 2017, doi: 10.14569/IJACSA.2017.080650.

[2] P. Sethi and S. Sarangi, "Internet of Things: Architectures, Protocols, and Applications," J. Electr. Comput. Eng., 2017, doi: 10.1155/2017/9324035.

[3] "Internet of Things for Telecom Engineers." [Online]. Available: http://forms1.ieee.org/rs/682-UPB-550/images/IEEE-IOT-White-Paper.pdf.

[4] W. Stallings, Cryptography and Network Security Prin- Ciples and Practice. Pearson Education, 2011.

[5] T. Daisy Premila Bai, K. Michael Raj, and S. Albert Rabara, "Elliptic Curve Cryptography Based Security Framework for Internet of Things (IoT) Enabled Smart Card," Proc. - 2nd World Congr. Comput. Commun. Technol. WCCCT 2017, pp. 43–46, 2017, doi: 10.1109/WCCCT.2016.20.

[6] A. Chhabra and S. Arora, "An Elliptic Curve Cryptography Based Encryption Scheme for Securing the Cloud against Eavesdropping Attacks," Proc. - 2017 IEEE 3rd Int. Conf. Collab. Internet Comput. CIC 2017, vol. 2017-Janua, pp. 243–246, 2017, doi: 10.1109/CIC.2017.00040.

[7] K. M. (Mat), "Capitalizing on the business value of the internet of things: The time to act is now," 2015.

[8] A. Amrani and N. A. Rafalia, "Lightweight Secure Scheme for IoT-Cloud Convergence based on Elliptic Curve," vol. 96, no. 1, pp. 144–155, 2019.

[9] K. Sha, W. Wei, T. Andrew Yang, Z. Wang, and W. Shi, "On security challenges and open issues in Internet of Things," Futur. Gener. Comput. Syst., vol. 83, pp. 326–337, 2018, doi: 10.1016/j.future.2018.01.059.

[10] S. M. M. R. M. Shamim Hossain, Ghulam Muhammad and A. A. Wadood Abdul, Abdulhameed Alelaiwi, "Toward End-to-End Biomet rics-Based Security for IoT Infrastructure," 2016.

[11] Y. Shou, H. Guyennet, and M. Lehsaini, "Parallel scalar multiplication on elliptic curves in wireless sensor networks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7730 LNCS, pp. 300–314, 2013, doi: 10.1007/978-3-642-35668-1_21.

[12] M. S. Albahri, M. Benaissa, and Z. U. A. Khan, "Parallel Implementation of ECC Point Multiplication on a Homogeneous Multi-Core Microcontroller," Proc. - 12th Int. Conf. Mob. Ad-Hoc Sens. Networks, MSN 2016, no. 1, pp. 386–389, 2017, doi: 10.1109/MSN.2016.070.

[13] Y. Faye, H. Guyennet, S. Yanbo, and I. Niang, "Accelerated precomputation points–based scalar reduction on elliptic curve cryptography for wireless sensor networks," Int. J. Commun. Syst., vol. 30, no. 16, pp. 1–11, 2017, doi: 10.1002/dac.3327.

[14] S. Thiraviya Regina Rajam and S. Britto Ramesh Kumar, "Enhanced elliptic curve cryptography," Indian J. Sci. Technol., vol. 8, no. 26, 2015, doi: 10.17485/ijst/2015/v8i26/80444.

[15] P. K. Dhillon and S. Kalra, "Elliptic Curve Cryptography for Real Time Embedded Systems in IoT Networks," 2016 5th Int. Conf. Wirel. Networks Embed. Syst., pp. 1–6, 2016.

[16] M. Q. Hong, P. Y. Wang, and W. B. Zhao, "Homomorphic Encryption Scheme Based on Elliptic Curve Cryptography for Privacy Protection of Cloud Computing," in Proceedings - 2nd IEEE International Conference on Big Data Security on Cloud, IEEE BigDataSecurity 2016, 2nd IEEE International Conference on High Performance and Smart Computing, IEEE HPSC 2016 and IEEE International Conference on Intelligent Data and S, 2016, no. September, pp. 152–157, doi: 10.1109/BigDataSecurity-HPSC-IDS.2016.51.

[17] J. R. Shaikh, M. Nenova, G. Iliev, and Z. Valkova-Jarvis, "Analysis of standard elliptic curves for the implementation of elliptic curve cryptography in resource-constrained E-commerce applications," 2017 IEEE Int. Conf. Microwaves, Antennas, Commun. Electron. Syst. COMCAS 2017, vol. 2017-Novem, pp. 1–4, 2017, doi: 10.1109/COMCAS.2017.8244805.

[18] H. Liu, Q. Dong, and Y. Li, "Efficient ECC scalar multiplication algorithm based on symmetric ternary in wireless sensor networks," Prog. Electromagn. Res. Symp., vol. 2017-Novem, pp. 879–885, 2017, doi: 10.1109/PIERS-FALL.2017.8293258.

[19] A. Mullai and K. Mani, "Enhancing the security in RSA and elliptic curve cryptography based on addition chain using simplified Swarm Optimization and Particle Swarm Optimization for mobile devices," Int. J. Inf. Technol., vol. 13, no. 2, pp. 551–564, 2021, doi: 10.1007/s41870-019-00413-8.

[20] K. Javeed, X. Wang, and M. Scott, "High performance hardware support for elliptic curve cryptography over general prime field," Microprocess. Microsyst., vol. 51, pp. 331–342, 2017, doi: 10.1016/j.micpro.2016.12.005.

[21] M. Hasan, P. Biswas, M. T. I. Bilash, and M. A. Z. Dipto, "Smart home systems: Overview and comparative analysis," Proc. - 2018 4th IEEE Int. Conf. Res. Comput. Intell. Commun. Networks, ICRCICN 2018, pp. 264–268, 2018, doi: 10.1109/ICRCICN.2018.8718722.

[22] N. York, B. Heidelberg, H. Kong, L. Milan, and P. Tokyo, Guide to Elliptic Curv e Cryptography Springer. Springer-Verlag New York, Inc., 2004.

A Comparative Analysis of Scalability Issues within Blockchain-based Solutions in the Internet of Things

Ahmed Alrehaili[1], Abdallah Namoun[2]
Faculty of Computer and Information Systems
Islamic University of Madinah
Madinah 42351, Saudi Arabia

Ali Tufail[3]
School of Digital Science, Faculty of Science
Universiti Brunei Darussalam
Brunei Darussalam

***Abstract*—Recently, enormous interest has been shown by both academia and industry around concepts and techniques related to connecting heterogeneous IoT devices. It is now considered a rapidly evolving technology with billions of IoT devices expected to be deployed in the upcoming years around the globe. These devices must be maintained, managed, traced, and secured in a timely and flexible manner. Previously, the centralized approaches constituted mainstream solutions to handle the ever-increasing number of connected IoT devices. However, these approaches may be inadequate to handle devices at a massive scale. Blockchain as a distributed approach that presents a promising solution to tackle the concerns of IoT devices connectivity. However, current Blockchain platforms face several scalability issues to accommodate diverse IoT devices without losing efficiency. This paper performs a comprehensive analysis of the recent blockchain-based scalability solutions applied to the Internet of Things domain. We propose an evaluation framework of scalability in IoT environments, encompassing critical criteria like throughput, latency, and block size. Moreover, we conduct an assessment of the notable scalability solutions and conclude the results by highlighting six overarching scalability issues of blockchain-based solutions in IoT that ought to be resolved by the industry and research community.**

***Keywords*—Blockchain; IoT; scalability; issues; distributed ledger; throughput; latency**

I. Introduction

The Internet of things (IoT) based solutions have evolved to cover every aspect of our daily lives. IoT technology has been deployed in various environments, including smart homes, healthcare, industrial etc. [1][2]. It is a collection of smart devices that are connected like a swarm of heterogeneous nodes. For decades, the centralized approach has been recognized as a widespread solution for such environments. However, the rapid increase in these nodes made it impractical to manage and maintain with the traditional centralized approach due to various scalability and speed challenges.

A decentralized approach seems to be a preferable candidate to address challenges within such complexed environments. It will assist in solving many challenges attached to IoT environments while reducing the significant costs related to the previously adopted centralized approach [80]. Blockchain technology is one of the most known decentralized approaches deployed to resolve concerns related to IoT devices [3]. It has demonstrated its efficiency and performance in the financial domain with applications, such as Bitcoin and Ethereum [4][5]. Blockchain is capable of keeping immutable records of every data generated and exchanged by IoT devices. Therefore, it can present a perfect solution in the following aspects:

- IoT environments need a layer to facilitate the interoperability of heterogeneous IoT devices. Blockchain can provide a composite layer above the peer-to-peer network with standard access for every IoT device.
- IoT environments require a tier to support the traceability of data among these IoT devices. Blockchain works as an immutable distributed ledger with a historic timestamp to ensure this feature for IoT devices.
- IoT environments are expected to provide security measures and improve trust aspects by deploying smart contacts and digital signatures.

While the deployment of blockchain technology in IoT-based environments offers various advantages, they still pose overarching scalability concerns due to the vast amount of data generated and the enormous number of IoT devices.

Traditional Blockchain platforms have inherited by design a challenge in their limited throughput. Throughput is determined by the number of transactions that can be appended and mined in the blockchain platform [77]. Various known blockchain platforms have different scalability rates, which is insufficient to handle the IoT environments [76][78]. For instance, Bitcoin has a limited number of transactions in a short period. The bitcoin network blocks are fixed in terms of size and frequency, which causes a scalability issue. The Bitcoin platform has even a lower throughput than Ethereum and other confidentiality issues [8]. However, the Ethereum platform is regarded to have a low throughput when deployed in IoT environments [6][7].

Researchers have carefully identified the so called scalability trilemma within the Blockchain environment [17], as depicted in Fig. 1. The concept, which Vitalik Buterin first coined, identifies the difficulty of finding a balance between three blockchain properties: decentralization, security, and scalability simultaneously [18]. Scalability trilemma means we can only achieve two out of the three properties at the same time. Furthermore, the scalability issue has some implications

related to the cost of the blockchain database. Practically, all transactions must be stored within a chain, so the chain size will increase as we append more transactions to the chain. This can increase the size of the chain, and maintaining and managing the chain becomes more difficult with time.

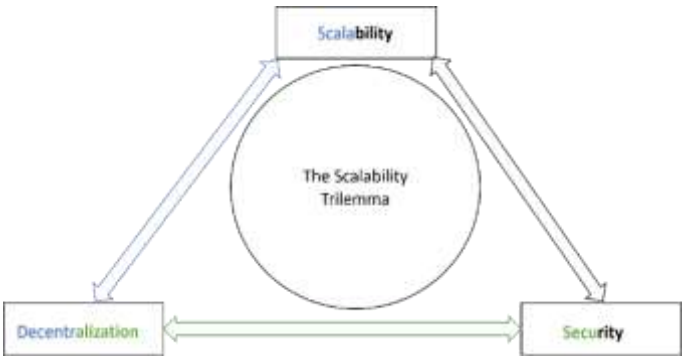


Fig. 1. The Blockchain Scalability Trilemma.

Currently, the blockchain size of Bitcoin and Ethereum are 354.419 GB and 870.37 GB, respectively [16][17]. Other blockchain platforms have been designed with high throughputs, such as IOTA, a commercial platform designed to be deployed in the IoT environment. However, it is regarded to have a long delay when addressing a massive amount of data [9]. Hyperledger Fabric and Ripple are two blockchain platforms that got high throughput [10][11]. Nevertheless, they suffer from the same issue of limited scalability, especially in terms of validating the nodes [12]. The following section will explain many solutions to tackle blockchain scalability issues.

In summary, we can summarize the contributions of our research as follows:

- Contribution one (theoretical): establish a fundamental understanding of the major scalability solutions using Blockchain in the IoT domain.
- Contribution two (theoretical): devise an evaluation framework for assessing the effectiveness of the current scalability solutions.
- Contribution three (empirical): evaluate existing scalability solutions with a focus on their strengths.

The remainder of this paper is divided into five sections. Section two reviews the Blockchain and IoT technologies. Section three compares various research scalability solutions that operate in different IoT layers. Section four proposes an evaluation framework and compares the Blockchain-based scalability solutions. Section five summarizes the key findings of our research.

II. Background Overview

A. Blockchain Technology

Blockchain, which is a distributed public ledger technology, was initially developed for cryptocurrencies such as Bitcoin. The concept of Blockchain was initially introduced by Nakamato [4] in 2008 but did not receive much attention initially. With the emergence of IoT in the past few years, Blockchain has started gaining the attention of researchers as a P2P technology for distributed and decentralized computation and data sharing. Blockchain can avert the possibility of intrusions by adopting cryptographic techniques in the absence of a centralized control environment. Interestingly, its unique security features, like transactional privacy, data immutability, authorization and integrity, fault tolerance, and transparency, allow Blockchain to be utilized in areas beyond cryptocurrency.

Blockchain technology has evolved around the idea that a single block, the fundamental component of Blockchain, stores certain types of information. The block is linked to similar blocks to form a chain where each block is associated with the previous block through a hash, as depicted in Fig. 2. The integrity of each block is assured by a hash function which is deployed to create a hash value of each block. The hash value is a digital fingerprint, which can be transformed to a different digital fingerprint by making minimal changes to the block, such as switching a bit value [52]. The hash value is the entity responsible for connecting every block with the previous block since each block possesses the block hash value behind it. Validation of the integrity by the system can easily be performed by running the hash function on every single block and then comparing the result with its prospective digital fingerprint.

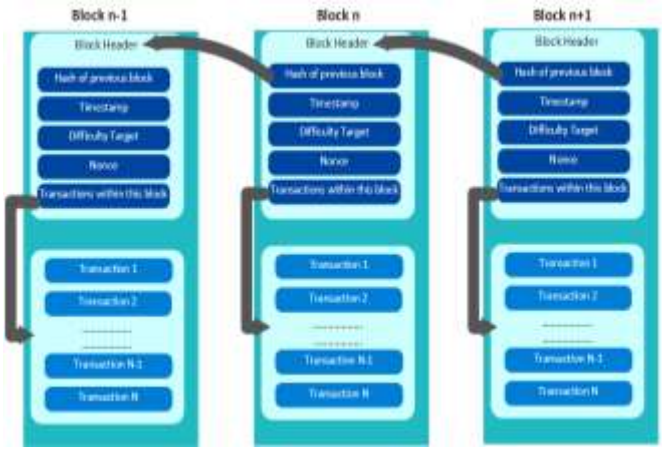


Fig. 2. The Blockchain Structure.

Blockchain technology is a decentralized ledger where each block is created and broadcasted to the connected peers. Therefore, each peer is guaranteed to have the identical most recent copy of the ledger. Thus, the forgery of a blockchain practically becomes very difficult. A blockchain environment has various characteristics, including decentralization, Tamper-proof, trustless, and anonymity.

- Decentralization: Blockchain is built around the idea of a distributed ledger with no central entity that controls the network. It means that the system is robust against a single point of failure. Therefore, if one node goes down, the system still functions properly.
- Tamper-Proof: The only way to take over the network is by launching a theoretical 51% attack [51]. In order to change the block content and make the validation process faster in comparison to all other peers within the network, the attacker roughly requires more than half of the computational power on Blockchain.
- Trustless: The blockchain environment depends on complete transparency. Thus, parties on the chain can trust each other.
- Anonymity: As mentioned above, there is no need for trust in the blockchain environment. Thus, parties on the chain remain anonymous with no need to reveal any party identity.

Furthermore, Blockchain can be arranged into three categories based on the participants' respective environment [27]. The categories can be summarized as follows:

1) Public blockchains: It is a permissionless blockchain that runs on a public network in a decentralized and distributed fashion. The environment is open, and any node can participate without any authorization [50].

2) Private blockchains: It is a permissioned blockchain that runs within a private network within an organization that governs and regulates all transactions.

3) Consortium blockchains: It is also a permissioned blockchain. However, it is initiated and controlled by related entities. A node must register ahead of their participation; then, they must adhere to rules and regulations.

Table I summarizes the key differences between the three blockchain categories.

TABLE I. COMPARISON OF MAJOR TYPES OF BLOCKCHAIN CATEGORIES

Characteristic	Public	Private	Consortium
Decentralization	Distributed Ledger	Centralized Ledger	Relatively Centralized Ledger
Immutability	Immutable	Not Immutable	Relatively Immutable
Transparency	Transpare nt	Not Transparent	Relatively Transparent
Scalability	Bad	Excellent	Good
Accessibility	Permissionless	Permissioned	Permissioned
Consensus Protocols	Proof of Work (PoW) & Proof of Stake (PoW)	Ripple	Practical Byzantine Fault Tolerance & Proof of Authority (PoA)
Example	Bitcoin & Ethereum	Ripple (XRP) & Multichain	Quorum & Hyperledger

B. Internet of Things (IoTs) Technology

Recently, the Internet of Things unleashed its power to deliver services across various domains from small businesses and social media to smart houses, smart cities, and industries. IoT connects resource-constrained heterogeneous devices with a broad range of functionalities in human and machine-centric communication networks. IoT has positively met the ever-evolving requirements of the above-mentioned sectors. However, the significant escalate in the number of such resource-constrained IoT devices and the massive information generated from them becomes a hurdle towards meeting the efficiency and security requirements.

The Internet of Things (IoT) is a network that attaches different devices to receive and transmit data over the Internet. The data is generated using various smart applications running on smart devices and sensors known as IoT devices. An estimated 50 billion IoT devices will be attached to the Internet worldwide in 2023 [49]. In Information Technology, IoT is undoubtedly a significant development connecting almost everything to the world wide web. Over the last few years, the increasing data rates and advancement in IoT paved the way for various concerns, with scalability being at the top of the list.

The IoT network consists of three layers, namely perception, communication, and industrial layer, as shown in Fig. 3. These sections can be briefly described as follows:

1) Perception layer: There are various IoT devices within this layer. These devices differ in function, which can include sensors, controllers, smart meters, etc. The primary function of these devices is sensing and collecting data from the physical environment. However, it might also react to actions in the physical environment.

2) Communication layer: There are several wireless/wired devices within this layer. These devices can be IoT gateway, Wi-Fi Access points, or small base stations. These devices deploy various communication protocols include Bluetooth, Near Field communication, etc. The primary function of these devices is to transfer data from the perception section to the industrial section.

3) Industrial layer: The industrial layer incorporates manufacturing, Airports, banks, supply chain etc. The decisions in these industrial organizations are build on the data gathered from the perception layer.

Previously, the centralized approach was the mainstream solution for handling complex structures of connected heterogeneous IoT devices. It was based on a traditional client server approach over the Internet. However, it suffered various challenges and is judged inadequate to handle data at this massive scale [80].

Fig. 3. Typical Three-Layer Internet of Things.

III. Analytical Comparison of Scalability Solutions

Due to the unraveled interest in deploying blockchain platforms in IoT systems, different approaches have been adopted to upgrade blockchain scalability. As mentioned, the challenge of enhancing blockchain scalability intensifies when more IoT devices/nodes are connected to each other and produce transactions at a higher rate. We identify and analyze the approaches published in recent literature to tackle the scalability issues. These approaches have been deployed in different layers of Blockchain and thereby can be classified as follows:

- Layer Zero "Approaches with the dissemination of Information": These proposed solutions focus on customizing the propagation protocol of information.
- Layer One "Approaches within the Blockchain": These proposed solutions focus on tackling the problem by changing the structure of blocks and consensus algorithms.
- Layer Two "Approaches off the Blockchain": These proposed solutions tackle the problem by executing some complex computational tasks off the Blockchain platform.

A. Layer Zero: Approaches with Propagation Protocol

Approaches dealing with the propagation protocol were classified recently by some researchers as a possible solution for scalability issues within Blockchain. Parties exchange and broadcast blocks of data/transactions inefficiently within the blockchain network, causing a high confirmation time. Enhancing and optimizing data transmission can result in improved throughput. Many studies have been published in layer zero, which can be explained in Table II.

B. Layer One: Approaches within the Blockchain

These proposed solutions on this approach focus on tackling the problem of scalability by different strategies, which can be viewed as follows:

- Redesign the structure of blocks.
- Implementing the DAG (Directed Acyclic Graph).
- Deploying Sharding techniques.
- Applying different consensus algorithms.

TABLE II. Comparison of Scalability Solutions within Layer Zero

Approach Name	How it Works	Advantages
BloXroute [70]	The design of the network is based on increasing the block size while decreasing the interval between blocks.	(+) Avoids forks (+) Enables fast propagations.
Velocity [71]	The structure of the protocol ensures an enhanced block propagation by deploying erasure code	(+) Increases throughput
Kadcast [72]	It is based on Kademlia Architecture, where it works similarly to the mechanism deployed for enhanced broadcasting with adjustable redundancy and overhead.	(+) Enables fast propagation (+) Enables secure transmission
Erlay [73]	The protocol improves the network connectivity while keeping the cost at a minimum level.	(+) Affordable cost (+) Private transmission

1) Redesigning the structure of blocks: The simplest approach to tackle the scalability concerns of Blockchain is redesigning the structure of blocks by increasing the block size. Practically, all transactions are appended within blocks in any blockchain platform. Since more transactions are recorded within a particular block, the throughput of transactions per block interval would consequently increase [13]. However, deploying such a simple approach comes with other direct and indirect challenges. One of these challenges is increasing the probability of hard forks in the blockchain platform. Consequently, a split of nodes within the Blockchain would happen as it happened in Bitcoin [14].

Traditionally, the Blockchain platform requires each node to record the complete history of all transactions to become a part of the network. An increase in block size means that each node must increase its storage requirements, making them more expensive to execute. Nodes that are not capable of securing such storage requirements would eventually be ruled out of the blockchain platform. As a consequence, a lesser number of centralized nodes would take control of the Blockchain. It leads Blockchain to lose its decentralized nature, so end users must have more trust in the protocol [15].

Redesigning the structured approach includes other techniques such a block compression. It can enhance the throughput of the Blockchain platform, where it reduces some unessential and redundant data of a block [22]. Compact block relay was designed and deployed according to the block compression approach [22]. It is based on changing the data structure of the original Bitcoin blocks along with shortening the transaction header data. Txilm is a technique based on the same concept of compression of blocks [22]. However, these kinds of techniques are prone to hash collisions.

2) Implementing Directed Acyclic Graph (DAG): The blockchain structure records transactions in chains that are arranged in a sole chain formation. Due to this type of liner formation, blocks are created one at a time with no concurrent operations. Consequently, Blockchain has a limited throughput with high latency. Allowing a concurrent operation would enhance throughput, so a new idea of blockchain structure build on DAG (Directed Acyclic Graph) is proposed [23].

The directed acyclic graph is a finite graph commonly deployed in a computer science major. The DAG-based blockchain Blockchain considers a block as a vertex in the DAG attached to other previous vertices. Moreover, The DAG-based Blockchain permits many vertices to be attached to a preceding vertex that creates simultaneous blocks. The IOTA foundation has designed its IoT-based Blockchain in the above-mentioned technique to address the scalability issues of Blockchain [28].

3) Deploying sharding techniques: The sharding technique was first developed within the database management field as an attempt to optimize large databases. It is based on partitioning a database into several physical fragments, where each fragment saves its distinct subset of the data. This divide of a large group across multiple servers permits the distributed management of operations of a single database, thus improving scalability [31]. Practically, it applies the concept of divide-and-conquer on the blockchain platform, so each platform will be divided into several smaller units called a shard. Fig. 4 shows the concept of the sharding technique on the blockchain platform. A pool of transaction is processed in multiple shards, that reduces the load on each node and makes it possible for nodes to process a small number of transactions. Recently, several studies have been published to tackle the scalability issues using the sharding technique to improve transaction throughput.

4) Applying different consensus algorithms: Various consensus algorithms have been used in different types of Blockchains. These consensus algorithms are used, so Blockchain becomes more resilient to malicious participants and message delays. Several algorithms are deployed in the research literature to solve security issues. However, each one of them has an overhead that affects blockchain throughput and scalability. Therefore, some optimizations are required to enhance the scalability of Blockchain. The essential consensus algorithms are as follows.

Proof of Work (PoW): To add blocks to the Blockchain, each node must perform some exclusive work known as Proof-of-Work (PoW) [36]. In Bitcoin, each node must compute a hash value less than a specific number, which is also known as the difficulty level set by the Blockchain. The difficulty level is changed periodically by the Bitcoin protocol, where it takes between five to ten minutes to produce a single block [36]. The procedure of finding a solution to the PoW puzzle (i.e., to find a winning hash value) is also called mining. Speed is critical in the the operation, so the mining prize is given to the first node that computes a winning hash. Furthermore, the node gets to include its proposed block in the Blockchain. Once a node finds a winning hash and broadcasts it to others. Next, other nodes have to confirm that the proposed hash value is correct and valid [37]. Since several nodes are computing the winning hash simultaneously, there is a possibility that several nodes compute the winning hash at the same time. Sequentially, each wining node includes its block, the Blockchain announces it over the peer-to-peer network. In such a scenario, there are temporary forks in the Blockchain due to some nodes including their block into the first branch of the Blockchain and others include in the second branch and so on. To fix this problem, the protocol will choose the longest branch and delete the other branches [36]. Due to the previous challenges in the original PoW algorithm, many optimization techniques were proposed to enhance the algorithm scalability [38][39][40].

Proof of Stake (PoS): It is deployed to avoid the PoW algorithm weaknesses. It replaces the mining process with an alternative idea where users can own a virtual currency in the blockchain platform. Practically, users can buy any amount of cryptocurrency and then utilize it in the form of the stake to purchase equivalent block creation chances in the blockchain platform by working as a validator. The validator cannot predict its turn ahead of time since the algorithm randomly chooses the validator node to create the block. At its original form, the algorithm has a problem called Nothing-at-Stake, where the algorithm does not provide incentives for nodes to vote for the accurate block. Nodes might vote for blocks supporting several forks and branches to maximize their chances of winning a reward as they do not consume anything from their resources [36]. There are other problems with the PoS where it assumes that the chances of an attack on the blockchain by the nodes having a higher amount of currencies are minimal. [37]. Therefore, several alternative solutions were proposed where [41] deploys randomization techniques to forecast the next validator. It utilizes a mechanism that finds the lowest hash number in combination with the length of the stake. Peercoin [42] selection is based on coin age-based selection, where older coins have a greater possibility of mining the next block. However, Ethereum is trying to switch from Ethash [43] to Casper [44].

Other Consensus Algorithms: Several consensus algorithms focus on tackling many problems where scalability is one of them. Other mentionable consensus algorithms available in the literature include Delegated Proof of stake [81], Practical Byzantine Fault Tolerance [82], Hybrid Consensus [83], Proof of Authority [84], Proof of Capacity [85], Proof of Participation [86]. Surveys comparing consensus algorithms are available in [19], [20],[61],[53], and [73]. We recommend analyzing these algorithms in future works.

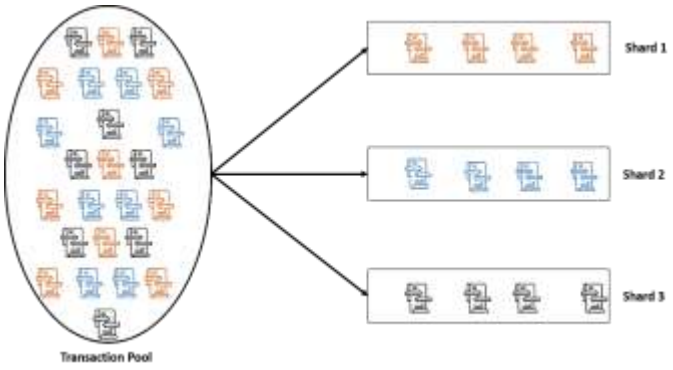


Fig. 4. An Exemplary Illustration of Sharding Techniques.

C. Layer Two: Approaches off the Blockchain

The proposed solutions in this approach focus on tackling scalability by executing some complicated computational work outside the Blockchain platform. These solutions apply different strategies, including payment channel, sidechain, off-chain computation, and cross-chain techniques. Below we provide an analysis of each approach.

1) Payment channels: The strategy of the payment channel is based on creating a temporary off-chain channel where some transactions can be executed off-chain so to reduce the volume on the main network and increase the transaction throughput of the whole Blockchain. Example approaches that employ payment techniques are described in Table III.

Fig. 5 demonstrates the concept of the lightning network technique, which includes three stages as described below:

- Establishing the channel by depositing some number of tokens in the channel (recorded on the main chain)
- Trading between two parties (recorded off the chain).
- Closing the channel where the number of tokens of both parties is recorded on the main chain.

2) Sidechain techniques: The Sidechain technique was first proposed at Pagged Sidechain [61]. Generally, it allows the assets in a specific blockchain to be moved between various sub-blockchains. It guarantees assets to be secure and saved. Several key sidechain algorithms are described in Table IV.

TABLE III. COMPARISON OF SCALABILITY SOLUTIONS USING PAYMENT CHANNEL TECHNIQUES

Approach Name	How it Works	Advantages
Lightning Network [45]	Uses two parties of Blockchain to establish their own off-chain private trading channel. The channel is dedicated to several low latency transactions.	(+) provides private communication
Raiden Network [46][47]	The technique is payment-based. The Raiden Network is deployed on the Ethereum network with support for all ERC20 [47].	(+) enables secure communication

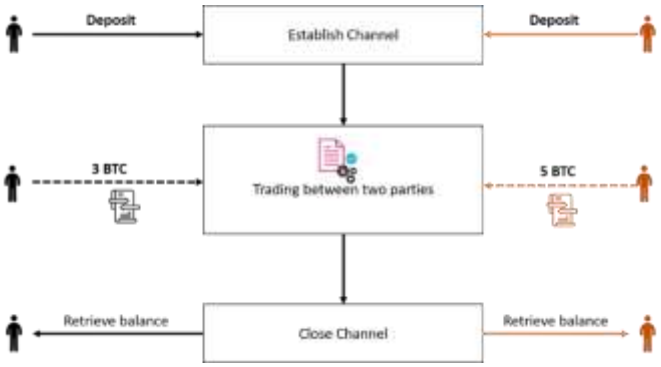


Fig. 5. An Exemplary Illustration of the Lightning Network

TABLE IV. COMPARISON OF SCALABILITY SOLUTIONS USING SIDECHAIN TECHNIQUES

Approach Name	How it Works	Advantages
Plasma [60]	The protocol permits a parent chain to create smaller copies as child chains. The created copy of a parent chain is designed and developed according to a specified use case. The parent chain delegates its work to child chains.	(+) improves the transaction throughput (+) delegates work to child chains
Pegged Side Chain [48]	The approach is based on a two-way peg, transferring the assets from the main chain to a child chain. It ensures that these assets are securely sent from the parent to a child by locking them until the pegged side chain obtains a simplified Payment Verification (SPV) proof. A confirmation period is enforced for security reasons. The newly transferred assets are halted on the sidechain to keep away from double-spending issues. The exact logic is applied once transferring the assets back to the main chain.	(+) provides secure communication
LiQuidity Network (NOCUST) [62]	The network is based on a data architecture named Merkleized Interval Tree. It is formed of a multi-layered tree which is deployed on NOCUST. It allows the party's' balances to be saved on private non-crossing interval space. Practically, all balances are verified against the amount registered in the smart contract on the main network.	(+) ensures the correctness of computations

3) Off-Chain computation: In Ethereum, miners must execute all contracts to validate their states. The operation is known to be costly and time-consuming. Therefore, many techniques help to build a scalable platform. Table V lists example off-chain computation techniques.

TABLE V. COMPARISON OF SCALABILITY SOLUTIONS USING OFF-CHAIN TECHNIQUES

Approach Name	How it Works	Advantages
Truebit [63]	It is designed based on outsourcing computations to trusted third parties known as solvers and challengers. Tokens are deposited to the smart contract by the solver. The challenger verifies the work done by the solver and gets compensation for its work.	(+) guarantees correctness of computations (+) adapts to computationally intensive applications.
Arbitrum[64]	Enables nodes to deploy smart contracts as virtual machines that include all rules of a contract. It has four types of roles: - Verifier: it acts as a global entity to validate transactions and publish accepted transactions. Key: it is a participant entity that can own currency and propose transactions. Virtual Machine: it is a virtual participant in the protocol which can own currency and exchange them. Manager: it manages the virtual machine and makes sure its correctness.	(+) enhances blockchain scalability

4) Cross-Chain techniques: Cross-chain techniques are considered to be potential solutions to improve scalability in Blockchain. Generally, these techniques are based on the interoperability among several separated chains. Therefore, the inter-connection between these chains can result in enhancing scalability. Fig. 6 depicts an example of the main cross-chain techniques. There are two main cross-chain algorithms which are listed in Table VI.

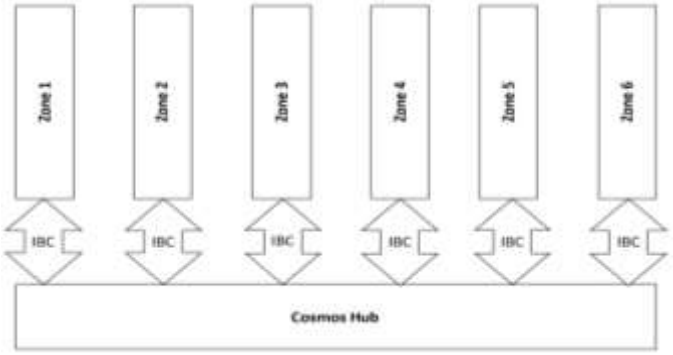


Fig. 6. An Exemplary Illustration of the Cosmos Network.

TABLE VI. COMPARISON OF SCALABILITY SOLUTIONS USING CROSS-CHAIN TECHNIQUES

Approach Name	How it Works	Advantages
COSMOS -+[65]	It is based on parallel independent Blockchain named as zones. The Tendermint BFT consensus algorithm supports each zone. The Cosmos Hub connects these zones.	(+) Increases throughput (+) Deploys Inter-blockchain communication protocol for privacy
POLKADOT [66]	It is based on a multi-chain protocol that attaches various blockchains with a relay chain. The relay chain allows separated blockchains to communicate with each other. The Polkadot acts as a mediator that connects to already functioning blockchains.	(+) Increases throughput (+) Secure communication

IV. SCALABILITY EVALUATION FRAMEWORK AND ANALYSIS RESULT

Scalability can be defined to incorporate some dimensions. The traditional definition stipulates from three perspectives, namely throughput, storage, and latency [74][75][79]. Blockchain is considered a network that can be measured by standard performance metrics like throughput and latency [77]. Since we are talking about Blockchain, throughput can be clearly associated with the number of committed valid transactions within the Blockchain per second [77]. Therefore, we can represent the throughput as follows:

Transaction Throughput=Number of Committed Transaction /Time in Seconds (1)

Latency is also associated with transaction latency which is defined as the proportion of the Blockchain to commit a transaction [77]. Therefore, we can represent the latency as follows:

Transaction Latency= (Confirmation Time * Blockchain Threshold)- Submission Time (2)

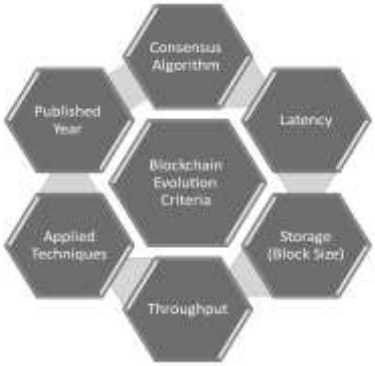


Fig. 7. Our Scalability Evaluation Framework.

Both performance metrics, throughput and latency, are closely related to the block size. Various blockchain networks suffer from issues about standing by for transactions to be committed within the block due to the fixed size of blocks [78]. Therefore, it is a critical parameter that must be included in blockchain network evaluations. Furthermore, Consensus algorithms and applied techniques are closely related to our scalability analysis, so we added them in our evaluation, as depicted in Fig. 7.

A. Comparison of the Scalability Blockchain-based Architectures

As mentioned above, this paper's main contribution is to analyze each Blockchain architecture and its main characteristics affecting IoT scalability. Our scalability evaluation framework incorporates various dimensions. Our selected dimensions include 1) throughput, 2) storage (block size), 3) latency, 4) deployed techniques, and 5) consensus algorithm. We will base our comparative evaluation on these criteria. Table VII details the findings of the comparison.

B. Summary of Scalability Issues

Our detailed analysis of state-of-the-art architectures aimed at resolving scalability challenges pertaining to blockchain solutions that could enhance the IoT domain. The significant challenges are summarized below.

- Challenge One: scalability is closely related to block the size. If the block size exceeds the network capacity, the block will not be attached to the chain. As a consequence, some solutions strive to increase the block size.
- Challenge Two: although increasing the block size enhances performance, it may increase the probability of blockchain forks. Therefore, other solutions enforce mechanisms to prevent the occurrence of forks.
- Challenge Three: scalability can be achieved by reducing some data within the block, so some solutions attempt to deploy compression techniques. However, it may affect valuable information about the block node states and records.

- Challenge Four: transactions are committed in the block only if all peers agree on its validity. As a result, the network suffers slow speed in appending transactions till it reaches consensus between the participating parties. Some solutions focused on implementing consensus algorithms to reduce the time required to achieve total agreement between peers.
- Challenge Five: more innovative solutions tried to redesign the structure of the Blockchain. Consequently, DAG and Sharding structures are deployed to avoid sequential execution of transactions which is adapted by the original blockchain structure. However, these structures inherite by design other issues. Data validity and availability are common issues within the Sharding structures, while computing power and cost are major concerns in the DAG structures.
- Challenge Six: scalability solutions are deployed outside the blockchain environments by outsourcing computationally intensive operations to a third party so the main chain can execute other light operations simultaneously. Accomplishing parallel execution of transactions enhance the prospect of scalability. However, the appeal of blockchain comes from the fact that we do not have to rely on third parties. By outsourcing the operations, we surrender an advantage and restrict the environment. Furthermore, concerns about third party's trustworthiness, security and privacy need to be resolved.

TABLE VII. A COMPARISON OF RECENT BLOCKCHAIN SOLUTIONS USING SCALABILITY DIMENSIONS

Blockchain Technology	**Distributed Technology**	**Throughput (Transactions per Seconds)**	**Latency (Secs)**	**Block Size (MB)**	**Consensus Algorithm**	**Year Originated**
Bitcoin [4]	List of blocks	7	600	1	Proof of Work (PoW)	2009
Segregated Witness [58]	Segregate digital sign	7	NA	4	Witnesses	2015
Inclusive block chain protocols [24]	Block DAG	65	NA	NA	Proof of work (PoW)	2015
IOTA [28]	Tx DAG	500	60	NA	Weight of transactions	2016
Byteball [29]	Tx DAG	20-30	60	NA	Witnesses	2016
Spectre [25]	Block DAG	NA	NA	NA	Proof of work (PoW)	2016
ByzCoin [67]	Apply different consensus algorithm (PBFT)	1000	20-15	NA	PBFT	2016
ELASTICO [32]	Sharding Technique	40	800	1	Proof of work (PoW) & PBFT	2016
ZILLIQA [57]	Sharding technique	2828	NA	NA	PoS	2017
Algorand [68]	Apply different consensus algorithms	875	22	NA	Byzantine Agreement	2017
Ouroboros [59]	Coin-flipping technique	257.6	120	NA	PoS Apply different consensus algorithm (PoS)	2017
Conflux [21]	DAG	6400	270-444	NA	PoS	2018
Phantom [26]	Block DAG	NA	NA	1	Proof of work (PoW)	2018
Nano [30]	Block-lattice	7000	1 to 10	NA	Weighted votes on transactions	2018
RapidChain [34]	Sharding technique	7380	7380	1	PBFT	2018
OmniLedger [33]	Sharding Technique & Block DAG	3500	800	1	PBFT	2018
DLattice [54]	Double DAG	1200	10		PANDA	2019
Monoxide [35]	Sharding technique	11694	13-21	1	Asynchronous consensus & Proof of work (PoW)	2019
CoDAG [55]	Block DAG	1151	NA	NA	NA	2020
Ostraka [56]	Sharding technique	400000	NA	1	Bitcoin-NG	2020
Meepo [69]	Sharding technique	120000	0.4-0.5	NA	consortium consensus	2021

V. Conclusion and Future Work

Enormous efforts have been made towards solving the scalability issues within Blockchain to adapt this promising solution to connecting heterogeneous IoT devices. In this paper, various scalability solutions were presented and compared according to their layer within the blockchain network. Next, the paper evaluated these solutions according to standard performance indicators such as throughput, latency, and storage. The paper attempted to summarize the existing blockchain solutions at different layers so to serve as a roadmap for more improvements by other researchers. In the future, we plan to extend our comparative analysis to investigate other issues impacting the blockchain-based networks, particularly those associated with security aspects.

References

[1] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, "A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications", IEEE Internet of Things Journal, Vol. 4 (5), pp. 1125-1142, (2017).

[2] Alrehaili, Ahmed, and Aabid Mir. "POSTER: Blockchain-based Key Management Protocol for Resource-Constrained IoT Devices." 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH). IEEE, (2020).

[3] A. Reyna, C. Mart´ın, J. Chen, E. Soler, and M. D´ıaz, "On Blockchain and its integration with IoT. challenges and opportunities," Future generation computer systems, vol. 88, pp. 173–190, (2018).

[4] S. Nakamoto, "Bitcoin: A peer-to-Peer Electronic Cash System" https://bitcoin.org/bitcoin.pdf (2009).

[5] G. Wood, "Ethereum: A Secure Decentralised Generalized Transaction Ledger", https://gavwood.com/paper.pdf

[6] Nejc Zupan, Kaiwen Zhang, and Hans-Arno Jacobsen. 2017. Hyperpubsub: A decentralized, permissioned, publish/subscribe service using blockchains: demo. In Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference: Posters and Demos. 15–16, (2017).

[7] Seyoung Huh, Sangrae Cho, and Soohyung Kim. Managing IoT devices using blockchain platform. In Proceedings of the 19th International Conference on Advanced Communication Technology. 464–467, (2017).

[8] Merve Can Kus Khalilov and Albert Levi. 2018. A survey on anonymity and privacy in bitcoin-like digital cash systems. IEEE Commun. Surv. Tutor, 1–44. https://ieeexplore.ieee.org/abstract/document/8325269, (2018).

[9] Bogdan Cristian Florea. Blockchain and internet of things data provider for smart applications. In Proceedings of the 7th Mediterranean Conference on Embedded Computing. 1–4, (2018).

[10] Hyperledger. Hyperledger-fabricdocs documentation release v0.6. Retrieved from https://buildmedia.readthedocs.org/media/pdf/ hyperledger-fabric/v0.6/hyperledger-fabric.pdf,(2017).

[11] Ripple. Solution overview. Retrieved from https://whitepaperdatabase. com/wp-content/uploads/2017/09/Ripple-XRP-Whitepaper.pdf, (2017).

[12] Runchao Han, Vincent Gramoli, and Xiwei Xu. 2018. Evaluating blockchains for IoT. In Proceedings of the 9th IFIP International Conference on New Technologies, Mobility and Security. 1–5, (2018).

[13] Jeff Coleman. State channels wiki, https://github.com/ledgerlabs/state-channels/wiki, (2016).

[14] Mattias Scherer. Performance and scalability of blockchain networks and smart contracts, 2 https://umu.diva-portal.org/smash/get/diva2: 1111497/ FULLTEXT01.pdf, (2017).

[15] Luke-jr. Block size limit controversy, https://en.bitcoin.it/wiki/Block_ size_limit_controversy, (2015).

[16] BitInfoCharts. URL: https://bitinfocharts.com/ethereum/, https://ycharts. com/indicators/ethereum_chain_full_sync_data_size Accessed 2021-07-14.

[17] Blockchain.com URL: https://www.blockchain.com/ko/charts/blocks size, https://ycharts.com/indicators/bitcoin_blockchain_size Accessed 2021-07-14.

[18] Ometoruwa, T. (2018), Solving the blockchain trilemma: Decentralization, security & scalability, www.coinbureau.com/ analysis/solving-blockchaintrilemma/, (2018).

[19] Nguyen, Giang-Truong, and Kyungbaek Kim. "A survey about consensus algorithms used in blockchain." Journal of Information processing systems 14.1, 101-128,(2018).

[20] Bamakan, Seyed Mojtaba Hosseini, Amirhossein Motavali, and Alireza Babaei Bondarti. "A survey of blockchain consensus algorithms performance evaluation criteria." Expert Systems with Applications 154 ,113385,(2020).

[21] C. Li, P. Li, D. Zhou,W. Xu, F. Long, and A. Yao, ``Scaling nakamoto consensus to thousands of transactions per second,", arXiv:1805.03870. [Online]. Available: https://arxiv.org/abs/1805.03870, (2018).

[22] Bip152. Comact Block Relay [Online]. Available: https://github.com/bitcoin/bips/blob/master/bip-0152.mediawiki Accessed on May 5, 2021.

[23] Shrier, Ian, and Robert W. Platt. "Reducing bias through directed acyclic graphs." BMC medical research methodology 8.1, 1-15,(2008).

[24] Y. Lewenberg, Y. Sompolinsky, and A. Zohar, ``Inclusive block chain protocols," in Proc. Int. Conf. Financial Cryptogr. Data Secur. San Juan, Puerto Rico: Springer, pp. 528_547, (2015).

[25] Y. Sompolinsky, Y. Lewenberg, and A. Zohar, ``Spectre: A fast and scalable cryptocurrency protocol," IACR Cryptol. ePrint Archive, vol. 2016, p. 1159, (2016).

[26] Y. Sompolinsky and A. Zohar, ``Phantom: A scalable blockdag protocol," IACR Cryptol. ePrint Archive, vol. 2018, p. 104, (2018).

[27] Lin, Iuon-Chang, and Tzu-Chun Liao. "A survey of blockchain security issues and challenges." Int. J. Netw. Secur. 19.5: 653-659,(2017).

[28] Iota. [Online]. Available: https://www.iota.org/ accessed on May 5, 2021.

[29] A. Churyumov. Byteball: A Decentralized System For Storage and Transfer of Value. [Online]. Available: https://byteball.org/Byteball.pdf accessed on May 5, 2021, (2016).

[30] C. LeMahieu. Nano: A Feeless Distributed Cryptocurrency Network. [Online]. Available: https://nano.org/en/whitepaper accessed on May 5, 2021.

[31] Weinan Wang, Joseph E Magerramov, Maxym Kharchenko, Min Zhu, Aaron D Kujat, Alessandro Gherardi, and Jason C Jenks. Facilitating data redistribution in database sharding, April 23, 2013. US Patent 8,429,162,(2013).

[32] L. Luu, V. Narayanan, C. Zheng, K. Baweja, S. Gilbert, and P. Saxena, "A secure sharding protocol for open blockchains," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur.-CCS, pp. 17_30,(2016).

[33] E. Kokoris-Kogias, P. Jovanovic, L. Gasser, N. Gailly, E. Syta, and B. Ford, ``OmniLedger: A secure, scale-out, decentralized ledger via sharding," in Proc. IEEE Symp. Secur. Privacy (SP), pp. 583_598, (2018).

[34] M. Zamani, M. Movahedi, and M. Raykova, ``RapidChain: Scaling blockchain via full sharding," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., pp. 931_948, (2018).

[35] J. Wang and H. Wang, ``Monoxide: Scale out blockchains with asynchronous consensus zones," in Proc. 16th USENIX Symp. Netw. Syst. Design Implement. (NSDI), pp. 95_112,(2019).

[36] A. Baliga, "Understanding blockchain consensus models," Tech. rep., Persistent Systems Ltd, Tech. Rep., (2017).

[37] Zheng, Z., Xie, S., Dai, H., Chen, X. and Wang, H. 'An overview of blockchain technology: Architecture, consensus, and future trends', Proceedings of the 2017 IEEE BigData Congress, Honolulu, Hawaii, USA, pp.557–564,(2017).

[38] I. Eyal, A. E. Gencer, E. G. Sirer, and R. Van Renesse, ``Bitcoin-NG: A scalable blockchain protocol," in Proc. 13th USENIX Symp. Netw. Syst. Design Implement. (NSDI), pp. 45_59,(2016).

[39] Y. Sompolinsky, Y. Lewenberg, and A. Zohar, ``Spectre: A fast and scalable cryptocurrency protocol," IACR Cryptol. ePrint Archive, vol. 2016, p. 1159,(2016).

[40] Y. Sompolinsky and A. Zohar, ``Secure high-rate transaction processing in bitcoin," in Proc. Int. Conf. Financial Cryptogr. Data Secur. San Juan, Puerto Rico: Springer, pp. 507_527, (2015).

[41] P. Vasin, "Blackcoins proof-of-stake protocol v2," [Online]. Available: https://blackcoin.co/blackcoin-pos-protocol-v2-whitepaper.pdf, (2014).

[42] S. King and S. Nadal, "Ppcoin: Peer-to-peer crypto-currency with proof-of-stake," Self-Published Paper, August, vol. 19, (2012).

[43] G. Wood, "Ethereum: A secure decentralised generalised transaction ledger," Ethereum Project Yellow Paper, (2014).

[44] V. Zamfir, "Introducing casper the friendly ghost," Ethereum Blog URL: https://blog. ethereum. org/2015/08/01/introducing-casperfriendly-ghost, (2015).

[45] J. Poon and T. Dryja. (2016). The Bitcoin Lightning Network: Scalable Off-Chain Instant Payments. [Online]. Available: https://www.bitcoinlightning.com,(2016).

[46] Raiden Network. [Online]. Available: https://raiden.network/ accessed on May 5, 2021.

[47] Erc20 Token Standard. [Online]. Available: https://theethereum.wiki/w/index.php/ERC20_Token_Standard accessed on May 6, 2021.

[48] A. Back, M. Corallo, L. Dashjr, M. Friedenbach, G. Maxwell, A. Miller, A. Poelstra, J. Timón, and P. Wuille. (2014). Enabling Blockchain Innovations with Pegged Sidechains. [Online]. Available: http://www.opensciencereview.com/papers/123/enablingblockchaininnovations-with-pegged-sidechains,(2014).

[49] S. Sorrel, "The Internet of Things: Consumer, Industrial & Public Services 2018-2023," Eds. Juniper, (2018).

[50] G. W. Peters and E. Panayi. Understanding modern banking ledgers through blockchain technologies: Future of transaction processing and smart contracts on the Internet of money. banking beyond banks and money. In Banking Beyond Banks and Money, pages 239–278. Springer, Cham, (September 2016).

[51] K. Gagneja, R. Kiefer, "Security Protocol for Internet of Things (IoT): Blockchain-based Implementation and Analysis" 2020 Sixth International Conference on Mobile and Secure Services (MobiSecServ), (March. 2020).

[52] A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, "Blockchain for iot security and privacy: The case study of a smart home," in 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 618–623, March 2017.

[53] Xiao, Y., Zhang, N., Lou, W., & Hou, Y. T.. A survey of distributed consensus protocols for blockchain networks. IEEE Communications Surveys & Tutorials, 22(2), 1432-1465,(2020).

[54] T. Zhou, X. Li, and H. Zhao, ''DLattice: DLattice: Permission-Less Blockchain Based on DPoS-BA-DAG Consensus for Data Tokenization,'' IEEE Access vol. 7, pp. 39273–39287, (2019).

[55] L. Cui, S. Yang, Z. Chen, Y. Pan, M. Xu, and K. Xu, ''An efficient and compacted DAG-based blockchain protocol for industrial Internet of Things,'' IEEE Trans. Ind. Informat., vol. 16, no. 6, pp. 4134-4145, (Jun. 2020).

[56] Manuskin, A., Mirkin, M., & Eyal, I. "Ostraka: Secure blockchain scaling by node sharding". In 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS & PW) (pp. 397-406). IEEE,(2020, September).

[57] Team, Z.. The ZILLIQA technical whitepaper. https://doi.org/10.2139/ssrn.3442330, (2017).

[58] L. Eric, L. Johnson, and W. Pieter. (2015). Segregated Witness Github Repository. Accessed: Jan. 15, 2021. [Online]. Available: https://github.com/bitcoin/bips/blob/master/bip-0141.mediawiki,(2015).

[59] A. Kiayias, A. Russell, B. David, and R. Oliynykov, ``Ouroboros: A provably secure proof-of-stake blockchain protocol," in Proc. Annu. Int. Cryptol. Conf. Santa Barbara, CA, USA: Springer, pp. 357_388,(2017).

[60] Poon, J., & Buterin, V. (2017). Plasma: Scalable autonomous smart contracts. White paper, 1-47,(2017).

[61] Sankar, L. S., Sindhu, M., & Sethumadhavan, M.. Survey of consensus protocols on blockchain applications. In 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 1-5). IEEE, (January 2017).

[62] "Liquidity Network" https://liquidity.network.

[63] Teutsch, J. & Reitwießner, C. (2019). A scalable verification solution for blockchains. arXiv preprint arXiv:1908.04756,(2019).

[64] Kalodner, H., Goldfeder, S., Chen, X., Weinberg, S. M., & Felten, E. W. (2018). Arbitrum: Scalable, private smart contracts. In 27th {USENIX} Security Symposium ({USENIX} Security 18) (pp. 1353-1370),(2018).

[65] https://v1.cosmos.network/resources/whitepaper.

[66] Wood, G. (2016). Polkadot: Vision for a heterogeneous multi-chain framework. White Paper,(2016).

[67] E. K. Kogias, P. Jovanovic, N. Gailly, I. Khof_, L. Gasser, and B. Ford,``Enhancing bitcoin security and performance with strong consistency via collective signing," in Proc. 25th USENIX Security Symp. USENIX Secur., pp. 279_296, (2016).

[68] Y. Gilad, R. Hemo, S. Micali, G. Vlachos, and N. Zeldovich, ``Algorand: Scaling byzantine agreements for cryptocurrencies," in Proc. 26th Symp. Operating Syst. Princ.-SOSP, pp. 51_68,(2017).

[69] P. Zheng, Q. Xu, Z. Zheng, Z. Zhou, Y. Yan and H. Zhang, "Meepo: Sharded Consortium Blockchain," IEEE 37th International Conference on Data Engineering (ICDE), 2021, pp. 1847-1852, doi: 10.1109/ICDE51399.2021.00165, (2021).

[70] Klarman, U., Basu, S., Kuzmanovic, A., & Sirer, E. G. (2018). bloxroute: A scalable trustless blockchain distribution network whitepaper. IEEE Internet Things J.,(2018).

[71] Chawla, N., Behrens, H. W., Tapp, D., Boscovic, D., & Candan, K. S. Velocity: Scalability improvements in block propagation through rateless erasure coding. In 2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC) (pp. 447-454). IEEE. https://doi.org/10.1109/BLOC.2019.8751427, (May 2019).

[72] Rohrer, E. & Tschorsch, F. Kadcast: A structured approach to broadcast in blockchain networks. In Proceedings of the 1st ACM Conference on Advances in Financial Technologies (pp. 199-213),(October 2019).

[73] Bouraga, S. A taxonomy of blockchain consensus protocols: A survey and classification framework. Expert Systems with Applications, 168, 114384 (2021).

[74] Naumenko, G., Maxwell, G., Wuille, P., Fedorova, A., & Beschastnikh, I. Erlay: Efficient Transaction Relay for Bitcoin. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (pp. 817-831). https://doi.org/10.1145/3319535.3354237, (November 2019).

[75] Zhou, Qiheng, Huawei Huang, Zibin Zheng, and Jing Bian. "Solutions to scalability of blockchain: A survey." IEEE Access 8: 16440-16455,(2020).

[76] Chauhan, A., Malviya, O. P., Verma, M., & Mor, T. S.. Blockchain and scalability. In 2018 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C) (pp. 122-128). IEEE,(July 2018).

[77] Luu, L., Narayanan, V., Zheng, C., Baweja, K., Gilbert, S., & Saxena, P.. A secure sharding protocol for open blockchains. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 17-30). https://doi.org/10.1145/2976749.2978389 , (October 2016).

[78] Hang, Lei, and Do-Hyeun Kim. "Optimal Blockchain Network Construction Methodology Based on Analysis of Configurable Components for Enhancing Hyperledger Fabric Performance." Blockchain: Research and Applications : 100009,(2021).

[79] Kim, Soohyeong, Yongseok Kwon, and Sunghyun Cho. "A survey of scalability solutions on blockchain." 2018 International Conference on Information and Communication Technology Convergence (ICTC). IEEE, (2018).

[80] Nartey, Clement, et al. "On blockchain and IoT integration platforms: current implementation challenges and future perspectives." Wireless Communications and Mobile Computing 2021 (2021).

[81] Larimer, Daniel. "Delegated proof-of-stake (dpos)." Bitshare whitepaper 81:85,(2014).

[82] Castro, Miguel, and Barbara Liskov. "Practical byzantine fault tolerance." OSDI. Vol. 99. No. 1999. (1999).

[83] Pass, Rafael, and Elaine Shi. "Hybrid consensus: Efficient consensus in the permissionless model." 31st International Symposium on Distributed Computing (DISC 2017). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017,(2017).

[84] De Angelis, Stefano, et al. "PBFT vs proof-of-authority: Applying the CAP theorem to permissioned blockchain." (2018).

[85] Sharma, Kapil, and Deepakshi Jain. "Consensus algorithms in blockchain technology: A survey." 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, (2019).

[86] Nandwani, Arpit, Mudit Gupta, and Narina Thakur. "Proof-of-participation: Implementation of proof-of-stake through proof-of-work." International Conference on Innovative Computing and Communications. Springer, Singapore, (2019).

An Internet of Things (IoT) Reference Model for an Infectious Disease Active Digital Surveillance System

Nur Hayati, Kalamullah Ramli, Muhammad Suryanegara, Muhammad Salman
Department of Electrical Engineering, Universitas Indonesia, Depok, Indonesia

***Abstract*—Internet of Things (IoT) technological assistance for infectious disease surveillance is urgently needed when outbreaks occur, especially during pandemics. The IoT has great potential as an active digital surveillance system, since it can provide meaningful time-critical data needed to design infectious disease surveillance. Many studies have developed the IoT for such surveillance; however, such designs have been developed based on authors' ideas or innovations, without consideration of a specific reference model. Therefore, it is essential to build such a model that could encompass end-to-end IoT-based surveillance system design. This paper proposes a reference model for the design of an active digital surveillance system of infectious diseases with IoT technology. It consists of 14 attributes with specific indicators to accommodate IoT characteristics and to meet the needs of infectious disease surveillance design. The proof of concept was conducted by adopting the reference model into an IoT system design for the active digital surveillance of the Covid-19 disease. The use-case of the design was a community-based surveillance (CBS) system utilizing the IoT to detect initial symptoms and prevent closed contacts of Covid-19 in a nursing home. We then elaborated its compliance with the 14 attributes of the reference model, reflecting how the IoT design should meet the criteria mandated by the model. The study finds that the proposed reference model could eventually benefit engineers who develop the complete IoT design, as well as epidemiologists, the government or the relevant policy makers who work in preventing infectious diseases from worsening.**

Keywords*—IoT; framework; digital surveillance; infectious disease; Covid-19

I. Introduction

The current Covid-19 pandemic era has taught us the importance of conducting surveillance of people who are carriers of infectious diseases. Such surveillance will provide the opportunity for medical personnel to monitor and avoid adverse events in the future. Infectious disease surveillance is a process that starts with data collection on the diseases and other relevant factors and is conducted continuously and systematically. It continues with a dynamic analysis of the disease spread from three different perspectives: temporal, spatial, and populational. This process aims to observe trends and current situations, and to provide data to help decide preventative measures and to control related diseases [1][2]. Therefore, the surveillance process is divided into the collection, analysis and interpretation, and dissemination of data [3][4].

The Internet of Things (IoT) is a technology with great scope for adoption as a tool in the digital surveillance of infectious diseases [5]. Functionally, IoT has the capacity to conduct an active surveillance process through the support of technological integration [2]. The IoT collects real data from sensors embedded in the end device, sends them to the data processing system, and shares the results in either real-time or through scheduling. Those activities demonstrate the IoT's ability to perform the surveillance process from start to end [6].

Many studies have been made of the development of the IoT for surveillance of infectious diseases [6][7][8][9]. However, such designs were developed based on the authors' own ideas or innovations, without taking into account a specific reference model. Therefore, it is important to build such a model which encompasses end-to-end system design.

This paper proposes a reference model for the design of an active digital surveillance system of infectious diseases employing IoT technology. We utilized the framework of the Center for Disease Control (CDC) surveillance evaluation [10][11][12], which consists of 10 parameters ranging from usefulness to representativeness, to which we added two modified parameters of security and standard [13][14], together with two others related to the essential concerns of mobility and sustainability. These framework constructs comprise the 14 new attributes of the reference model. Each of the attributes has specific indicators, which we customized to accommodate IoT characteristics and to meet the needs of infectious disease surveillance design.

To verify our proposal, we tested the reference model on a conceptual IoT design as a Covid-19 digital surveillance system. We analyzed the compliance of such a system design with the 14 attributes of our proposed reference model. The design of the Covid-19 digital surveillance system consisted of IoT end devices, local server, IoT gateway, internet connection, cloud server, and users' end-devices.

This work contributes to the work of designers or engineers on developing a complete IoT system for active surveillance of any infectious disease. The model can be used as a reference to evaluate whether their end-to-end design meets essential engineering parameters, as reflected by the 14 attributes in the reference model. The work should also eventually benefit epidemiologists, governments, or relevant policy makers. For example, the data quality attribute reflects epidemiologists' concerns over the completeness and validity of data. Compliance with this attribute will ensure that the system provides valid data, which will help epidemiologists advising the government or policy makers in making appropriate interventions to prevent infectious diseases from worsening.

Hibah Penelitian Disertasi Doktor, Kementerian Riset Dan Teknologi/Badan Riset Dan Inovasi Nasional, Indonesia

The remainder of the paper is organized into sections addressing the theories of the IoT as a surveillance system (Section II); the IoT as a form of active digital surveillance of infectious diseases (Section III); the reference model that we are proposing (Section IV); discussion of the proof of the concept (Section V); and the conclusion to the study (Section VI).

II. Theoretical Approach: The IoT as a Surveillance System

A. An Infectious Disease Surveillance System based on Digital Technology

The development of digital technology has reached the health-related sector, including the digital surveillance of infectious diseases [15]. Such surveillance can be conducted with either an active or passive approach, according to the data collection method [16]. Passive surveillance is generally performed by monitoring query logs from the users' search engines, keywords on the web, and hashtags on social media, to obtain the most widely circulated data related to infectious diseases, followed by analysis of the data [17]. Google Flue Trends is an example of an internet-based surveillance application which provides information about early detection systems for epidemics. The application was built based on trend data from the processing of users' keywords to search for disease information. This was then validated by matching the trend with confirmed case data from the laboratory [17]. Although passive surveillance obtains the data more quickly, it does not fully represent detailed geographic and demographic information [14]. Therefore, J.K. Harris et al. [18] concluded that technology is needed that allows the surveillance process to involve active participation from the community.

An example of active digital surveillance is reporting from the public through an online system. This means that the public's active participation using technological platforms is needed to obtain more accurate information [19]. However, such a reporting mechanism requires intense public intervention. The IoT is an alternative technological solution, which is suitable for active digital surveillance by utilizing sensors embedded in the IoT end devices (wearable medical devices) to collect specific data related to the disease. This solution made IoT minimizing users' efforts and concentrated intervention in the data reporting process.

B. The IoT as an Active Digital Surveillance System

The IoT has advantages in terms of processing speed and system automation [20]. Compared to traditional surveillance, internet-based technology has features which can automatically detect infectious diseases more quickly [17]. For such diseases, computing speed is valuable for shortening their detection time, allowing preventive measures to be taken immediately [21] [22]. System automation increases surveillance efficiency when disease transmission incidents occur on a large scale [7][14], and also effectively reduces the workload of medical personnel in pandemic situations [13].

Another advantage of the IoT is its technological ability to provide quality data [14], which are derived from the aggregated sources. The aggregated source completes the data to facilitate analysis, early detection, surveillance, and monitoring of the emergence of infectious diseases [23]. This advantage overcomes the shortcomings of traditional surveillance systems which result in incomplete reporting data on the emergence of infectious diseases due to limited resources, time, and reporting systems [17]. Because of these advantages, we are optimistic that the IoT can provide the meaningful time-critical data needed by infectious disease surveillance systems.

Several studies have developed the IoT as an active digital surveillance system for infectious diseases. In the case of Ebola, research [6] proposed as a framework to assist in detecting and monitoring patients suspected of being infected with the disease. The framework integrated RFID technology, wearable sensors, 3G/4G wi-fi internet connections, and cloud computing. In another example, in the respiratory infection SARI, the IoT was employed to prevent and treat the disease, to improve patient management, and to provide effective consultation [7]. According to Y. Song et al., in their research [7], the IoT infrastructure was used to develop long-distance communication between patient devices, hospitals, and medical equipment to manage SARI.

Research on the IoT for an infectious disease surveillance system can be implemented on a small and large scale. Research on the IoT as a small-scale surveillance system was conducted by Lundrigan et al. [9], who built an IoT-based monitoring system on a household scale to help epidemiologists observe disease exposure through the sensor data. In addition, research on larger scale implementation of the IoT was conducted by Nsoesie et al. [15], which found that mass gatherings with high densities, with residents close to each other, increased the risk of spreading diseases. Consequently, the IoT was utilized as early detection technology for controlling disease spread at mass gatherings [15]. Additionally, on a global scale, Zhu et al. [8] demonstrated IoT's ability to quickly perform detection and monitoring of infectious disease using cost-effective point-of-care (POC) diagnostic devices connected to the internet.

III. IoT Infrastructure for an Active Digital Surveillance System of Infectious Diseases

A. Illustration and Interpretation of an IoT Infrastructure for the Infectious Disease Surveillance Process

IoT is a technology whose infrastructure consists of four-layers of protocols with different technologies and functions on each layer. However, they complement each other in constructing an IoT-based end-to-end technology solution. The protocols comprise the sensing and identification layer, network infrastructure layer, data processing layer, and integrated application layer. This section maps each IoT layer onto each surveillance process to adjust the layer functionality of the IoT infrastructure in surveillance operations. In the IoT infrastructure design, infectious disease surveillance consists of four processes: collection, transmission, analysis and interpretation, and dissemination of data.

A visualization of the interpretation of the infectious disease surveillance process by the IoT infrastructure is shown in Fig. 1. According to the illustration, the surveillance process can be interpreted as follows. The data collection process

reflects the functionality of the sensing and identification layer. In the IoT design, the data transmission process realizes the tasks of the network infrastructure layer, while the dynamic analysis and interpretation realize the function of the data processing layer. Dissemination of data to provide evidence for the surveillance process performs the integrated IoT application layer function.

Active digital infectious disease surveillance begins with primary data collection from the community utilizing sensor technology. Data gathered from the sensor is then sent automatically to the data processing center over the transmission media. This automatic data transmission process is the fundamental difference between traditional and IoT surveillance systems. In traditional surveillance, officers collect data manually, input them into the system, and then distribute them. The following surveillance process is dynamic analysis and interpretation, which can be performed more quickly in IoT due to the supporting intelligent system features in its data processing center. The output of data processed by the IoT intelligent system is used to provide surveillance data ready to be disseminated as evidence.

B. Supporting Technology to the Design of an IoT Infrastructure Used for the Active Digital Surveillance System of Infectious Diseases

Each process in the active digital surveillance system requires supporting technology. For the collection of data in the IoT design, this is supported by sensors, which are usually embedded in the end device. Fig. 2 presents a technical diagram of the functions and types of sensor employed as data collection technology in IoT. The sensors can be categorized into three types:

- A sensor-type actuator, which operates by receiving stimuli then converting them into electrical signals.
- A network-based sensor type, which generates data via wireless communication between the transmitter and its receiver, and is characterized by integration of the transceiver with the IoT system.
- The Tags type sensor, which uses identifiers/codes that are scanned/tagged to obtain the data.

Both wired and wireless technologies support IoT in the process of transmitting surveillance data. In the IoT design, wired technology is generally used to provide a high-speed network connection of the aggregated systems in a data center. On the other hand, wireless technology is used from IoT end devices to the gateway due to its design flexibility, which supports user mobility and installation simplicity during implementation. Fig. 3 presents considerations when choosing data transmission media. Internet service providers (ISPs) normally handle high-speed data transmission to data centers, with either wired or wireless connections, while the connection from the IoT end device to its gateway is generally an option selected by the developer when building a surveillance system.

Supporting technology for analyzing and interpreting surveillance data on the IoT include cloud computing, fog computing, edge computing, and cloudlets. Fig. 4 shows a diagram of the data processing technology function and options. Several factors need to be considered when selecting the technology, including the amount of data being processed, network connections, and the capacity of the data processing resources. Centralized processing utilizes cloud computing technology equipped with an intelligent system. In contrast, distributed data processing employs fog computing and edge computing technology. In addition, cloudlets are an option if the surveillance requires a distributed data processing system in many locations with light processing loads. These employ resource sharing to ease the burden on the data processing system. Distributed processing shortens the data path taken and the data transmission time.

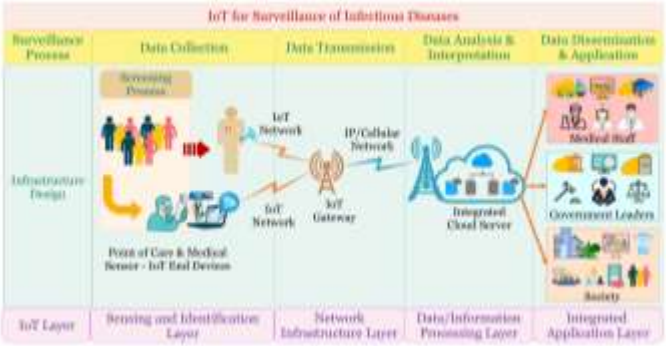


Fig. 1. Interpretation of the IoT Infrastructure as an Infectious Disease Surveillance System.

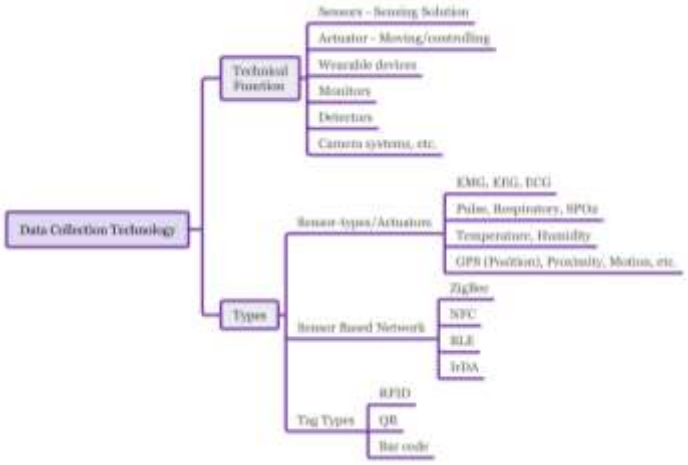


Fig. 2. Diagram Functions and Types of Data Collection Technology.

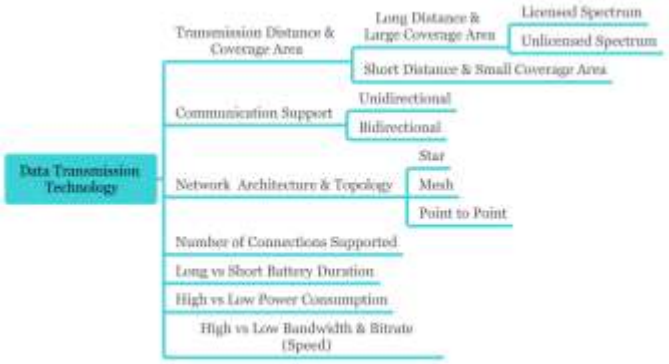


Fig. 3. Diagram of the Factors for Consideration when Selecting Data Transmission Technology.

Fig. 4. Diagram Functions and Types of Data Collection Technology.

In IoT design, surveillance data are disseminated online via web-based or mobile-based applications. Web-based ones are more suitable with respect to the data processing center. On the end-user side, mobile applications employing the CoAP (Constrained Application Protocol) or MQTT (Message Queue Telemetry Transport) protocols are more suitable because of the IoT environment.

IV. A Reference Model for the Design of an IoT-based Active Digital Surveillance System of Infectious Diseases

This section describes our proposed reference model for the design of an IoT-based active digital surveillance system of infectious diseases. The model consists of 14 attributes divided into several technical detail indicators. We customized the indicators of each attribute to accommodate IoT characteristics and meet the needs of infectious disease surveillance design. We derived ten of the 14 attributes from the CDC's surveillance evaluation with their modified indicators [10][11][12]. We adapted two attributes from the papers by R. P. Singh, M. Javaid, A. Haleem, and R. Suman in [13] and by S. L. Groseclose and D. L. Buckeridge in [14]. We then added the remaining two attributes, namely mobility and sustainability. Table I presents the proposed reference model, comprising the 14 attributes together with their respective indicators.

Unlike the case of noninfectious diseases, infectious disease surveillance requires control of the disease spread through contact tracing mechanisms. Therefore, the mobility attribute plays an important role in complying with the basic need for infectious disease surveillance. In addition, IoT technology has a relatively long-life span, so maintaining sustainability in its implementation is necessary, especially when the IoT is employed in the health sector. Several indicators, such as maintenance, inspection, evaluation, and auditing, are essential for maintaining sustainability and preventing any technical issues from arising in the IoT surveillance system.

1) ATTRIBUTE #1: USEFULNESS: Attribute #1 refers to the usefulness needed to align the IoT design with the surveillance objectives and to ensure that the IoT infrastructure and service benefit the IoT-based surveillance. This attribute consists of two indicators, applicability and customizability. The applicability indicator refers to the adjustment of IoT technological capabilities to the design needs for epidemiological surveillance, while the customizability indicator refers to a customizable IoT design for a single or multi-purpose surveillance system.

2) ATTRIBUTE #2: DATA QUALITY: Attribute #2 refers to the data quality demanded by epidemiologists to determine the most appropriate interventions for infectious disease. This attribute is realized by two indicators: data completeness and data validity. Data completeness is generated through the IoT infrastructure, which acts as global healthcare and monitoring technology to increase data gathering in the system [5]. In addition, the IoT is designed to run a verification program to validate all the data entering the system.

TABLE I. Proposed Reference Model for the Design of an IoT-based Active Digital Surveillance System of Infectious Diseases

No	Reference Model Attribute	Reference Model Indicator
1	Usefulness	Applicability
		Customizability
2	Data Quality	Data completeness
		Data validity
3	Timeliness	Rapidity
		Timeliness
4	Flexibility	Adaptability
		Scalability and extensibility
5	Simplicity	Manageability
		Automatic system
6	Acceptability	Interoperability
		Compatibility
		User friendly
7	Stability	Reliability and availability
		Accessibility
8	Sensitivity	Precision
		Threshold configuration
		Calibration
9	Positive Predictive Value	Accuracy
10	Representativeness	Data Synchronization
		Data traceability
11	Security	Infrastructure and management security
		Data security and privacy
12	Standards	Standardization
		Regulation
		Certification
13	Mobility	Tracking system
		Handover and roaming
		Portability
14	Sustainability	Maintenance
		Inspection
		Evaluation and audit

3) ATTRIBUTE #3: TIMELINESS: Attribute #3 refers to the timeliness of the surveillance system. This attribute plays an important role in reporting the emergence of infectious diseases, finding suspects of the infectious disease, and preventing the disease from spreading in outbreak situations. In an IoT-based system designed for infectious disease surveillance, the timeliness attribute is divided into two indicators, rapidity and timeliness.

4) ATTRIBUTE #4: FLEXIBILITY: Attribute #4 refers to the flexibility of design related to surveillance needs. A surveillance system should be flexible, able to follow the changes in epidemiological patterns, information needs, clinical practice, reporting sources, and other required changes [4][11] [14]. The attribute of flexibility consists of two indicators: adaptability and scalability and extensibility. Adaptability refers to IoT infrastructure and platforms that are able to keep up with changes in the surveillance environment. Scalability and extensibility refer to changes in the levels of hardware, software, and services that can be added or subtracted physically or logically.

5) ATTRIBUTE #5: SIMPLICITY: Attribute #5 refers to simplicity, meaning the ease of operating the surveillance system. In IoT design, the simplicity attribute is realized through two indicators: manageability and automatic systems. Manageability is concerned with managing the complex IoT systems to obtain full visibility from a series of processes that work automatically [24]. Technically, the IoT can be managed in a centralized or distributed manner. An automatic system allows the IoT infrastructure operational design to work automatically, thus reducing human intervention.

6) ATTRIBUTE #6: ACCEPTABILITY: Attribute #6 refers to acceptability, which has three indicators: interoperability, compatibility, and user friendly. The acceptability attribute in IoT-based surveillance is defined as the system's ability to involve technology as a surveillance provider, and the community as surveillance users. Involvement of technology in systems is realized through the interoperability and compatibility indicators, while community engagement in accessing the system is facilitated through a user-friendly design.

7) ATTRIBUTE #7: STABILITY: Attribute # 7 refers to stability in the performance of IoT-based surveillance systems. This attribute is supported by indicators of accessibility, availability and reliability. The accessibility indicator ensures that the available resources are accessible by users, while reliability and availability are related to determination of the Service Level Agreement (SLA) value of the built surveillance systems, which is affected by agreements with supporting parties as service providers.

8) ATTRIBUTE #8: SENSITIVITY: Attribute #8 refers to sensitivity and indicates the ability of IoT surveillance to identify every case. There are three indicators for realizing sensitivity in IoT design: precision, threshold configuration, and calibration. Precision in the IoT-based surveillance system setting helps to obtain data precisely. Accurate threshold configuration of each required parameter also ensures that the data are precise. Calibration is used to maintain system accuracy; the calibration process is performed periodically.

9) ATTRIBUTE #9: POSITIVE PREDICTIVE VALUE: Attribute #9 refers to the positive predictive value, which is defined as the IoT surveillance system's ability to report data according to the case definition. The indicator of this attribute is data accuracy. Therefore, IoT-based system design should produce accurate infectious disease case data, which should be validated with other relevant data.

10) ATTRIBUTE #10: REPRESENTATIVENESS: Attribute #10 refers to representativeness. In designing IoT-based surveillance system, representativeness refers to the system's ability to present data which depict the number of cases in the population under surveillance. This is achieved through two indicators: data synchronization and data traceability. Both of these are essential for ensuring data correctness and portraying the data's originality consistent with the source [25][26].

11) ATTRIBUTE #11: SECURITY: Attribute #11 refers to security. This attribute plays an important role in maintaining public trust in the surveillance process using the IoT. It is supported by two indicators: infrastructure and management security and data security and privacy. Infrastructure and management security is a series of processes to secure the system end to end, from sensors to applications. It involves various types of action and policy to secure the infrastructure and to ensure that the security arrangement works properly and efficiently. Security and data privacy is an effort to preserve data at rest, in motion, and in use, which involves process and transmission media.

12) ATTRIBUTE #12: STANDARDS: Attribute #12 refers to standards, which consist of three indicators: standardization, regulation, and certification. Standardization is a global reference for building an integrated IoT-based surveillance system, while regulation refers to national or regional standards, and the rules derived from global standards. It is used to regulate IoT implementation design in each country or region. Subsequently, certification is applied to guarantee the system's operation and its security and safety. This should be done before mass implementation.

13) ATTRIBUTE #13: MOBILITY: Attribute #13 refers to mobility. This attribute is related to three indicators: tracking system, portability, and handover and roaming. A tracking system in IoT-based surveillance is applied to track the infectious disease suspects, and also to prevent and find any close contact in community interactions. Portability of devices effectively supports the functionality of IoT-based systems, which operate at rest and in motion. Handover and roaming support IoT mobile devices to operate in the moving system.

14) ATTRIBUTE #14: SUSTAINABILITY: Attribute #14 refers to sustainability, which has three indicators: maintenance, inspection, and evaluation and audit. Sustainable IoT design for surveillance system is needed, since they are planned to be implemented over a relatively long period. The IoT system requires maintenance to optimize its performance during its lifetime, and also requires inspection to avoid malfunctions and maintain its operational quality. Each change or improvement to the IoT system should be documented in a reporting system. The report can then be used as a supporting document for evaluation and audit in order to accomplish management requirements [27][28].

V. PROOF OF CONCEPT

In this section, we test our proposed reference model by designing an IoT-based community Covid-19 surveillance system. We illustrate the IoT infrastructure design supporting a Covid-19 active digital surveillance system. We then analyze the compliance of the system design by utilizing the 14 proposed reference model.

A. IoT Infrastructure Design for Community-based Covid-19 Surveillance

Covid-19 is an infectious disease spreading all over the world which has attained pandemic status [29] and has a high rate of spread and transmission. According to data published by WHO [30], the number of confirmed cases around the world had reached 57,882,183 including 1,377,395 deaths, as of November 22, 2020.

This study tested the proposed reference model by designing a community-based surveillance (CBS) system for Covid-19 employing the IoT. The community was represented by a nursing home for elderly people with a high risk of contracting the disease, as they tend to socialize with others in the home, which is contrary to the efforts to prevent Covid-19, namely minimizing gatherings and maintaining a distance. Eradication of the virus in such a setting will increasingly become a challenge considering that Covid-19 spreads through droplets, and as some elderly people have poor hearing and tend to interact at close range. We designed the infrastructure for the Covid-19 surveillance in the nursing home as depicted in Fig. 5. The end-to-end IoT infrastructure design consisted of wearable IoT medical devices, communication media, an integrated processing system, and user end devices. Hereafter, wearable IoT medical devices are referred to as IoT end devices.

In designing the Covid-19 surveillance in the nursing home, IoT end devices are employed to detect initial symptoms and prevent closed contact. The devices were built from an ESP32 development board equipped with a wireless system to manage the connection. Each board was configured with a unique identification code indicating the identity of the device. The code was then synchronized with a unique identifier for those using the device and being monitored. Therefore, medical staff could conduct the individual monitoring of the registered users under surveillance via the attached IoT end device. Initial Covid-19 symptoms were detected from body temperature through a DS18B20 sensor and breathing patterns through a heart rate and oxygen saturation sensor, namely MAX30102. In addition, close contact prevention was performed using a real-time location sensor, DWM1000.

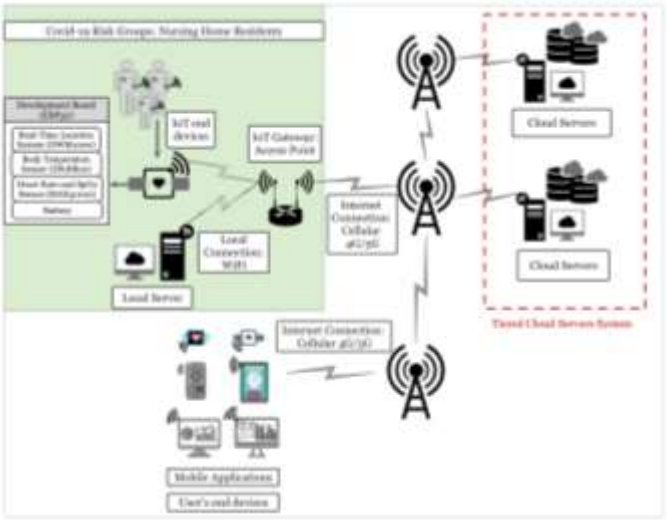


Fig. 5. IoT Infrastructure Design as a Covid-19 Active Digital Surveillance System.

Data collected from the sensor were encrypted and transmitted through Wi-Fi using the ESP32 module to the access point. In the design, the access point acts as an IoT gateway connected to the local and internet systems. Local systems were represented by IoT end devices worn by the elderly and a local server managed by surveillance staff in the nursing home. The local server was supported by cloudlet technology and served as a distributed data recording and processing system in the local community. The local server was connected to the cloud server with a tiered level. It means that the local server had a connection to cloud servers in the nearest cloud server, such as in community health centers or referral hospital, and then to cloud servers at a regional or national level. The local server's database was then synchronized with other databases on a larger scale in a cloud server, supported by cloud computing technology. The WHO cloud servers acting as the global data centers are used to aggregate the Covid-19 related data from each country.

Data from many sources were collected and integrated in the cloud servers for further processing. Complete, accurate, and up-to-date data from various sources enrich our knowledge of the incidence of Covid-19 from temporal, spatial, and populational aspects. The data can also be used to determine trends and to predict the distribution of Covid-19 cases at the regional level or aggregated at the national level.

The users' end devices, as shown in Fig. 5, included a mobile phone, gadget, laptop, or computer belonging to the community members or their guardians, stakeholders, or public. For guardians and medical staff, the user's end device had an additional function, allowing it to be used as a

medium for receiving alerts when Covid-19 symptoms or close contact interaction were detected in the elderly or community members. The device's general function was as a means to access and report the Covid-19 data according to the user's role. Public participation in reporting cases via the online system significantly and concurrently increases the completeness of the surveillance data [2].

In term of security, a lightweight security program was configured to secure the IoT end device entity and encrypt the data. The design of lightweight IoT security was based on the work of the authors of the reference [31]. IoT end devices must not be duplicated and must be secure from any types of attack. In addition to securing the devices, firmware and local storage were configured so that they could only be accessed and modified by the authorized parties. End-to-end encryption was designed to be configured in the transmission media. Wireless security standard was applied to secure data during transmission. The security applications were applied on local and remote servers of the processing system's infrastructure to secure the Covid-19 related data. Appropriate security implementation was needed in both the hardware and software to strengthen the security system. The security design should also include regulations to manage access to the data processing system securely by setting up access control and strict authentication. Additionally, update and patch mechanisms should be scheduled regularly. The mechanism is applied following the software security requirements to avoid system vulnerability [32]. As part of solution in preserving data security and privacy in the users' side, they were requested to regularly update their mobile application password installed on the end device. In terms of data access and reporting, privacy protection must be applied to protect the distribution and utilization of data [33].

In this designed scenario, the surveillance operation was performed in a tiered system with various supporting technologies. Hence, we designed the IoT operation in a real-time and scheduled system. The real-time system operation was applied to detect the Covid-19 symptoms and prevent closed contact. It was applied between the IoT end device to the nearest cloud server in a community health center, or referral hospital with more medical specialists and adequate facilities. In contrast, the scheduled system was applied to report and synchronize Covid-19-related data between cloud servers at a regional or national level.

As tools for active digital Covid-19 surveillance, IoT-based systems require public participation. Therefore, user-side systems should be designed with a user-friendly approach to attract the public. Such systems contribute to determining the interest of the public and organizations in collecting surveillance data [34]. User-friendly systems that are straightforward and intuitive are essential for the community, especially those who are not familiar with technology applications [35]. Fig. 5 shows the user-side systems representing front-end systems consisting of IoT end devices worn by the elderly or other community members and the users' end devices. IoT end devices require a user-friendly interface in order to be used by community members with varying levels of technological understanding. The user end-device applications also require a user-friendly interface to facilitate access and reporting of Covid-19-related data or to receive alerts and proactive notifications in the case of anomalies [36]. Furthermore, the user side representing the backend system is part of the IoT-based surveillance, apart from the frontend. This is operated by the surveillance staff, epidemiologists, or stakeholders. A user-friendly interface system in the backend system is also required to assist the surveillance staff in operating and efficiently managing the surveillance systems.

B. Analysis of Design Compliance with the Proposed Reference Model

The Covid-19 surveillance system implemented in the nursing home started with data collection from a screening process targeted at the elderly with signs or symptoms of the disease and all suspected cases. Fig. 6 illustrates the initial Covid-19 surveillance and screening process in the community. As shown in the figure, the elderly as community members are considered as mobile users. The nursing home health facilities located in each community are illustrated as local systems. The remote system represents the systems installed outside the nursing homes, which is a centralized aggregation system supported by cloud computing technology for the Covid-19 management center.

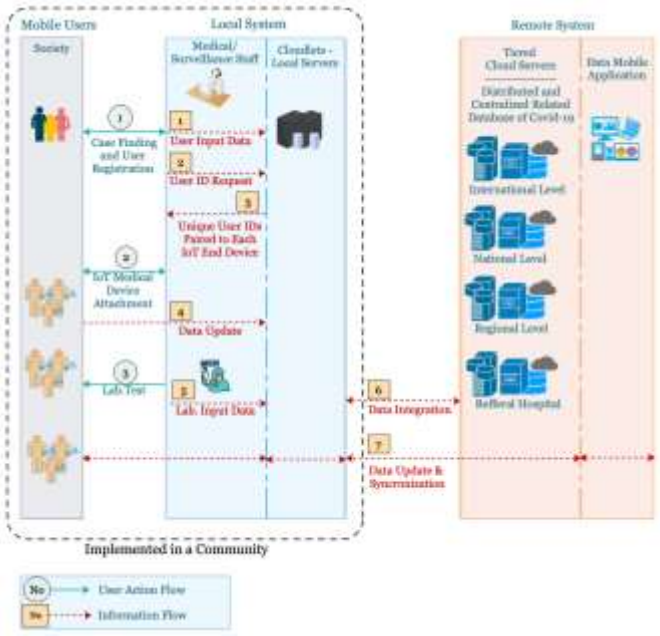


Fig. 6. Design of the Screening Mechanism and Initialization of the Covid-19 Surveillance Process.

Three steps were involved in implementing the Covid-19 surveillance system with the IoT. The first step was to find a case at the community level by examining the members. Initial laboratory tests were performed in this step to maximize the prevention spread of asymptomatic Covid-19. The surveillance staff recorded each community member on a registration form and input their medical data records into the Covid-19 database in the local system. The second step was to attach a wearable IoT medical device (IoT end device) to each community

member, which continuously monitored three parameters: users' body temperatures, and whether they had reached 38°C; respiratory rates; and real-time user location. The first two parameters were used to check the users' possibility of respiratory tract infection symptoms and to monitor the condition of those who had already had congenital SARI disease. The last parameter was used to prevent close contact interaction. The last step was Covid-19 laboratory testing. For community-based surveillance, it is better to utilize point-of-care laboratory test devices that are portable, support mobility and are connected to a data center via an internet network. Example portable laboratory devices are GeNose [37] [38] or in the case of this research point-of-care RT-PCR devices [8]. For the design, as an IoT point of care device, RT-PCR was designed to support mobility and to be connected to a data center via an internet network.

The IoT end devices were attached until the quarantine or treatment period was over. Following WHO procedures, IoT end devices were taken off suspect users whose results from two RT-PCR tests on two consecutive days with an interval of> 24 hours were negative. The devices were unattached from close contact users who had completed a 14-day quarantine period. However, for suspect and probable users whose RT-PCR test results were confirmed positive, the monitoring and examination process was conducted continuously according to the Covid-19 handling procedures, until the patient recovered.

After data collection, the surveillance process continued with data transmission, analysis and interpretation. In the community-based Covid-19 surveillance utilizing the IoT, data gathered from the IoT end devices were recorded on the local server and then updated and synchronized with the remote system on a broader scope to be integrated with other databases for analysis and interpretation. This was so that the IoT-based Covid-19 data were complete, detailed, valid, and continuously updated data. The data processing system implemented at either cloudlet or cloud computing level was equipped with intelligent systems to support the data analysis and interpretation. This intelligent IoT system played an essential role in accelerating and aggregating the data computation process in both the cloudlets and cloud computing. The last surveillance process was data dissemination and application. In this design, the process was propagated via lightweight mobile applications towards the users' end device.

Following the illustrations in Fig. 5 and 6, we now analyze how the IoT design for an active digital community Covid-19 surveillance system complies with the 14 attributes proposed in the reference model (Table I).

1) DESIGN COMPLIANCE WITH ATTRIBUTE #1 (USEFULNESS): Attribute #1 has two indicators, applicability and customizability. The applicability of IoT in Covid-19 surveillance relates to rapid case detection and close contact prevention. The community-based application is part of national and global surveillance systems to find large numbers of cases and prevent contact quickly. As shown in Fig. 6, the applicability of IoT-based surveillance in detecting cases follows three steps: user data collection, medical screening, and monitoring of users' condition through the IoT end devices. These actively check users' temperature and breathing patterns in the form of initial symptom data. The device is attached for approximately 14 days as long as there is no interaction with Covid-19 patients, or for more than 14 days as a result of growth in local transmission cases in each region.

The customizability of IoT-based system in Covid-19 surveillance is needed to adjust the design to different surveillance scenarios. Following WHO guidelines, each region or community could have a different surveillance scenario according to the Covid-19 transmission level. For example, in a nursing home located in an area with community transmission, the surveillance is applied to all community members, with standard sensors in their IoT end devices. On the other hand, in hospital-based surveillance the IoT end device design can be customized to have more functions in a complex system. Aside from detecting early symptoms, use of the IoT for Covid-19 can be in the form of a multi-purpose system, such as an integrated application for tracking suspect cases and detecting clusters, or a system for online reporting and trend monitoring. The IoT system can also provide additional health services, such as preparing up-to-date patient and family medical data. This helps medical personnel with predictions, diagnoses, treatments, and decisions related to infectious disease cases [39].

2) DESIGN COMPLIANCE WITH ATTRIBUTE #2 (DATA QUALITY): Attribute #2 has two indicators, data completeness and data validity. Data completeness for Covid-19 surveillance is obtained by integrating primary and secondary IoT data. Primary data are collected from the initial screening process, online reporting and the IoT end devices used by the elderly in nursing homes, or by subjects of community-based surveillance during the observation period. The primary data encompass personal data, telephone numbers, residence address during the last 14 days, travel history, date of the appearance of symptoms related to Covid-19, and associated conditions. The data are recorded on the local server or cloudlet, and labeled with a unique identifier to represent each user's identity. This identifier also acts as a synchronization code to identify its IoT end device pair. The auto-updated system also supports data completeness in the IoT-based surveillance. On the other hand, secondary surveillance data are obtained from the database. These are used for complementing the dynamic analysis and validating the primary data. Examples of secondary data are related to population and GIS; hospital databases; laboratory, pharmaceutical and EHR data; and other related data.

Data validity of the IoT-based Covid-19 surveillance is checked from the screening process in nursing homes and other communities. All personal data registered on the system is validated using an identity card or passport. Symptom data from the initial medical examination are validated by integrating them with each patient's EHR data. This integration is continued by incorporating data collected from the IoT end devices and advanced laboratory RT-PCR tests. Additional

symptom data beyond the basic examination and IoT end device capability can be added or reported via a mobile application system. Medical personnel need to validate all the additional reported data.

The Covid-19 data quality generated by the IoT-based surveillance systems should be more complete and valid since the data are integrated thoroughly. Data integration enriches the Covid-19 analysis related to patients' comorbidities and their advanced symptoms. Such integrated data also helps professionals provide much more appropriate treatment and follow up on emergency health events more accurately [2].

3) DESIGN COMPLIANCE WITH ATTRIBUTE #3 (TIMELINESS): Attribute #3 has two indicators, rapidity and timeliness. The rapidity of IoT-based Covid-19 surveillance in a community is needed to address the increasing spread of the disease. The pandemic has put the global health situation at very high risk; a significant number of Covid-19 positive cases leads to many casualties and the exhaustion of medical personnel. The design of rapidity in IoT-based Covid-19 surveillance in each community is achieved by the implementation of a partial real-time system. The system that connects the IoT end device to the nearest cloud server or Covid-19 management center is set to operate on a real-time basis. It is designed to assist medical personnel in detecting and monitoring Covid-19 suspects and their probable condition directly. Real-time surveillance design also influences Covid-19 patients' rapidity in receiving medical treatment or other follow up care according to public health management procedures.

The timeliness of IoT-based Covid-19 surveillance is achieved by combining real-time and scheduled system design. A real-time system is operated for Covid-19 case management in each community, while a scheduled system is operated to provide aggregated case data reporting. According to WHO provisions, all suspected, probable, and confirmed Covid-19 cases must be reported to the health office and the Public Health Emergency Operation Center. Immediate reporting should be made within 24 hours of the epidemiological investigations, and the timeliness of the IoT system can be designed to satisfy these provisions. Therefore, the IoT system accelerates the process of reporting data, which is useful for advanced Covid-19 analysis and cluster identification [13]. The IoT also speeds up the updating of infectious disease-related data to stakeholders and the public for surveillance dissemination [23]. Overall, the end-to-end IoT integration system is key to the rapidity and timeliness of IoT-based Covid-19 surveillance held in each community. However, the rapidity and timeliness of IoT integration are influenced by data processing and transmission speed, and IoT end device response time.

4) DESIGN COMPLIANCE WITH ATTRIBUTE #4 (FLEXIBILITY): Attribute #4 has two indicators, adaptability and scalability and extensibility. The adaptability of IoT-based Covid-19 surveillance is required since each community has different scenarios and technology support. Since Covid-19 transmission has reached the global community in many countries, adaptability design should comply with many different conditions; those which may differentiate the design include the unequal distribution of IoT infrastructure, differences in geographic area, and the population density in each community. The adaptability of IoT ensures that the surveillance process can work properly under various conditions. In designing the system, the IoT network from the end device to the gateway should be adapted to the local infrastructure, while maintaining the quality of network services. The local server's processing resources should consider the computational load. Specification of the processing device implemented in a region with community transmission scenarios will differ from sporadic case scenarios. The adaptability of IoT design also includes flexibility in the selection of technology for mass community implementation. However, the selected technology must comply with the specified standards, pass medical device test standards, and be certified.

The scalability and extensibility design of IoT-based Covid-19 surveillance is correlated with the expansion of the system implementation. It is required when communities or regions change their Covid-19 surveillance due to growth in case numbers. The scalability and extensibility of IoT design includes both physical and logical systems. A scenario change in the community can be scalable by upgrading the device and adjusting its specifications so that it can be applied to the new scenario. It also can be extensible by optimizing the functionality, performance and efficiency of the existing system. Optimization of the existing network can be made by increasing bandwidth capacity and giving priority to the QoS of the existing network. Additionally, optimization of the IoT mobile application can also be made by supplementing its features to support hospital management. This could reduce medical personnel's workload in the current Covid-19 pandemic situation. With specific regard to patient management, IoT saves patients' time with the readmission feature for those who have been confirmed Covid-19 positive, thereby reducing the density of service centers and speeding up the patient handling process [13].

5) DESIGN COMPLIANCE WITH ATTRIBUTE #5 (SIMPLICITY): Attribute #5 has two indicators, manageability and automatic system. The manageability of IoT-based Covid-19 surveillance applied in communities is paramount. This is because surveillance involves many resources and users and consists of many tiers. The ease of manageability of IoT simplifies the varying Covid-19 surveillance activities. In community-based surveillance, as illustrated in Fig. 5, it can be seen that the number of IoT end devices is directly proportional to the number of users under surveillance. The devices can be easily managed because each of them has a device identity, namely, a unique identification code configured on the ESP32 development board. In addition, the IoT infrastructure that supports the surveillance can be managed by using

hardware, software, or system management tools. These are configured to send an alert and notification when part of the system fails, or when anomalous activities occur. Overall, IoT design manageability should be compatible with heterogenous IoT protocols and platforms in its integrated system [40].

The automatic system of the IoT applied in each community simplifies operation and minimizes the manual process of Covid-19 surveillance. The IoT end device for detecting symptoms and preventing close contact is designed to be worn by all community members, including those with disabilities. Therefore, it is ideally set to minimize user intervention. The device's sensors operate automatically in detecting Covid-19 symptoms and the devices can also be configured to automatically send an alert. In the design shown in Fig. 5, an automatic alert is sent out when a user is detected as having a temperature of ≥38°C, a high respiratory rate or unstable breathing pattern. Simultaneously, an automatic alert is also sent when a community member is detected to have interacted with Covid-19 patients within a radius of 1 meter for >15 minutes [41]. In larger community-based surveillance, the IoT end device can be equipped with GPS technology to accommodate an automatic system for tracing contacts over a wide area. The device is configured to monitor users' history of locations visited based on GPS coordinates, from two days before the onset of the disease up to 14 days afterwards [41]. All IoT end device alerts are sent to community members' guardians and medical personnel at the local level. Furthermore, the automatic system of IoT-based Covid-19 surveillance can be seen in its data updating and synchronization, which impacts on the rapidity of Covid-19 data dissemination to stakeholders. We can thus state that an automatic system of IoT integration greatly helps Covid-19 surveillance in high-risk communities with fewer medical staff.

6) DESIGN COMPLIANCE WITH ATTRIBUTE #6 (ACCEPTABILITY): Attribute #6 has three indicators, interoperability, compatibility, and user friendly. An interoperable system is required since IoT-based Covid-19 surveillance is constructed from different protocols, devices, and communication media. The various technologies comprising the IoT system should operate seamlessly. IoT system interoperability can minimize the technical constraints that might appear with global implementation. It also can minimize the problem of differences in technological backgrounds between developed and developing countries which adopt IoT-based surveillance to fight Covid-19. Consequently, communities with different settings and technological backgrounds can synergize the end-to-end process of Covid-19 surveillance.

Compatibility of devices, platforms, and services is required to build an integrated IoT-based surveillance system. Hardware and software must be compatible with the main technology infrastructure in each community or region. Hence each community can implement the IoT-based Covid-19 surveillance system efficiently. The compatibility of mobile applications with various device platforms and network conditions is also required to facilitate user access. Another type of compatibility related to IoT-based surveillance is that of data format. The system implemented in each community should generate a standardized data format that can be interchangeable purely for Covid-19 surveillance purposes.

A user friendly system is necessary to facilitate the end-to-end operation of IoT-based Covid-19 surveillance. As illustrated in Fig. 5, the IoT infrastructure system for such community-based surveillance design is divided into backend and frontend systems. User friendly design dashboards assist stakeholders in managing the IoT infrastructure and all the systems, while user friendly application interfaces assist surveillance staff in managing surveillance-related data. In general, both designs help staff to simplify operations on the backend system. With regard to the frontend system, a user-friendly system encompasses hardware and software interfaces. An example of a hardware interface is the IoT end device worn by the elderly. The user friendly hardware design must consider user comfort across all age categories, including the elderly, and the supporting technology infrastructure. User friendly software should be designed to be an easy-to-use application for all communities with different technological understanding and language skills. Such user friendly software encompasses the design of the mobile application interface installed on user end devices. Community members or societies require a user friendly frontend application to quickly understand the agreement related to data privacy [42]. Moreover, those with special needs require a simpler application interface for accessing and reporting Covid-19 data; for example, by simply pressing a specific button on their mobile application. We therefore hope that information technology can play a significant role in involving the community in improving health surveillance [14].

7) DESIGN COMPLIANCE WITH ATTRIBUTE #7 (STABILITY): The attribute #7 indicators are accessibility, availability and reliability. Accessibility design refers to convenient access to a globally integrated IoT-based Covid-19 surveillance information system. IoT-based surveillance is designed to provide easy access to the integrated system and its data from various technological platforms. Access can be made anytime, anywhere, and by anyone, according to the users' role. Accessibility design for medical professionals is used to access and monitor those suspected of having Covid-19 and patients' updated condition in real-time. Accessibility design for community members and the public is used to access and report Covid-19 related information to the official and trusted system.

The availability and reliability of IoT design for Covid-19 surveillance are correlated with the service delivery of the system in each community. They are influenced by the approved Service Level Agreement (SLA), which determines its uptime and downtime. Internet Service Providers (ISPs) and application service providers are the parties involved in determining SLAs to support the overall IoT infrastructure. If the community-based surveillance system commits to providing integrated services with an SLA value of 99.999%, this means that the IoT infrastructure built to support the surveillance system should be ready to operate 24 hours a day

and 7 days a week, with a downtime of 5.26 minutes per year. To monitor the system's availability and reliability, we can apply an internet-based management system application. This will generate and send a notification to the administration when the system is inaccessible so that the problem can be followed up immediately.

Regarding technical detail, the design of IoT infrastructure availability and reliability can be realized by implementing redundancy topology and configuring high availability in crucial devices. Redundancy applies to infrastructure lines and devices so that there is no single point of failure. The high availability configured in the crucial devices indicates the system's capability to perform a failover function. Next, the Quality of Services (QoS) feature can be applied to the IoT infrastructure [43] to guarantee excellent service, namely, one that can be accessed concurrently and is seamlessly connected to the system. Simultaneously, the availability of surveillance data can be maintained by preparing a data recovery and backup system. The IoT-based surveillance system should be supported by appropriate storage capacity. Furthermore, higher IoT stability design can be achieved by considering a resilient system, which refers to the IoT's ability to defend against various types of interference, restabilize changing conditions, and adapt its behavior and structure to constant changes [44][45].

8) DESIGN COMPLIANCE WITH ATTRIBUTE #8 (SENSITIVITY): Attribute #8 has three indicators: precision, threshold configuration, and calibration. The precision of the ongoing IoT-based Covid-19 surveillance system is primarily determined by the measurement sensitivity of the IoT end device sensors. In general, the sensors' precision affects the handling procedure of Covid-19, since the sensor detection results contribute to determining individual cases. This precision greatly impacts the measurement accuracy of users' body temperature, respiratory rate patterns, and the distance between interacting users. The DS18B20 sensor should precisely check users' body temperature; if this reaches 38°C, it is the first sign that an initial Covid-19 symptom has been detected. The second sign is the user's breathing pattern detected by the MAX30102 sensor, whose detection precision is used to check whether the breathing pattern is irregular. In addition, to prevent close interaction, the DWM1000 sensor is configured to monitor each user's position in real-time. All the sensor measurements then become a reference for detecting initial Covid-19 symptoms and preventing close contacts and also for activating alerts sent to community members' guardians and medical staff. In contrast, any imprecision of IoT end devices will worsen the Covid-19 control measures.

Threshold configuration is applied to the IoT end device sensor. The device is designed to send out an alert when the user's body temperature is ≥38°C, and their respiratory rate patterns change. This threshold configuration is applied following interim WHO guidance for determining initial Covid-19 symptoms. A real-time position sensor is configured to ensure a safe distance is maintained between users when they interact. The threshold configuration is set to a 1-meter distance, with less than 1 meter assumed to be physical contact, and it is combined with a device timing system with a threshold of 15 minutes. An alert is sent out when the sensor detects an interaction between users and Covid-19 patients within a radius of 1 meter or less for > 15 minutes [41]. The precision of setting the threshold determines the accuracy of the alerts. An incorrect threshold setting will lead to inaccurate and late alerts and alarms sent to users and medical personnel. This situation also means that case detection will become less accurate, resulting in medical treatment delays and endangering patients. Therefore, precision and thresholds must be set appropriately and accurately.

The calibration process of the IoT end devices is performed before the devices are distributed. After a certain period, device re-calibration is needed to maintain measurement precision. Device calibration keeps the IoT system's operational capabilities standardized, thus minimizing the tendency to generate anomalous data. Therefore, the calibration process is performed to ensure that IoT devices can be employed anywhere and will produce the same data quality.

9) DESIGN COMPLIANCE WITH ATTRIBUTE #9 (POSITIVE PREDICTIVE VALUE): The attribute #9 indicator is accuracy. The accuracy of reporting Covid-19 data following the WHO case definition guidelines is essential for presenting factual data. The latest version of the WHO interim guidance currently gives three terms to define Covid-19 cases: suspect, probable, and confirmed. These three case definitions are used as the standard for ongoing Covid-19 surveillance. Therefore, IoT-based Covid-19 surveillance needs to translate each case definition of the clinical symptom criteria proposed by WHO into the IoT system's technical functionality. WHO also differentiates the details of Covid-19 symptom criteria for children and adults into mild, moderate, and severe illnesses, including asymptomatic patients. To achieve the accuracy of the system for finding case data in the field, IoT end devices and their supporting systems must be designed and configured strictly, following the WHO clinical criteria provisions. The IoT system's accuracy should ensure that all reported Covid-19 case data meet with the case definition accurately.

The accuracy design of IoT-based surveillance is illustrated in Fig. 6. There are three steps to finding Covid-19 suspects in community-based surveillance, such as in a nursing home. These are initial laboratory tests, monitoring of each community by IoT end devices, and advanced laboratory tests. After passing the screening process, namely the initial laboratory test, the ongoing surveillance is performed with the assistance of the IoT end devices. Initial Covid-19 symptoms are detected according to data parameters gathered from the attached IoT end device. To confirm the symptoms, advanced laboratory tests, such as RT-PCR, are conducted on each suspect and probable Covid-19 case. In this way, medical staff will have accurate data regarding the case, positive or negative. For further examination, the RT-PCR test results that confirm

positive cases are aggregated with EHR data to complement the comorbid analysis and other required medical data to achieve more accurate Covid-19 analysis. Consequently, epidemiologists will have more in-depth and comprehensive analysis data related to Covid-19 occurrence, trends, predictions, and other health-related data.

10) DESIGN COMPLIANCE WITH ATTRIBUTE #10 (REPRESENTATIVENESS): Attribute #10 has two indicators, data synchronization and data traceability. Data synchronization in community-based Covid-19 surveillance, as illustrated in Fig. 6, occurs between IoT end devices and the local and remote servers. IoT-based surveillance can provide representative data by employing individual devices to monitor the condition of each community member under surveillance. Through direct user involvement, the data collection process can represent the actual number of suspects, probable, and confirmed positive cases of Covid-19. Employing IoT end devices as part of an active digital surveillance system provides the advantage of real field data. Each user's data are securely synchronized with the attached IoT end device, employing the user's identity code for individual monitoring. The IoT end device data are complemented through the reporting mechanism via the users' mobile applications. Automatic data synchronization is then applied to individual medical data and the real field data collected from the IoT end devices and reporting applications. These data are then integrated with the Geographical Information System (GIS) data and population distribution data. Eventually, IoT-based surveillance is expected to help epidemiologists to obtain more representative Covid-19 data which portrays the number of cases with the updated and accurate demographic data in some geographic regions under surveillance.

The data traceability of the IoT-based Covid-19 surveillance system is used to maintain data consistency. An integrated surveillance system from an end-to-end IoT infrastructure makes it easier to realize this traceability. According to the data activity logs, data cloud computing in the remote system can be traced to its source. The tiered system design of the IoT infrastructure supports data tracing down to its original source. Any data entering the IoT information system must be traceable transparently, because it is difficult to use inconsistent data to represent the number of cases in a particular population. An alternative solution to accommodating data traceability is the data standardization format and data security application. The standardization of data reporting formats simplifies data tracing in the IoT system. The security application can be applied to check the data manipulation. Data changes can be easily traced when data integrity checking is applied. Additionally, the non-repudiation feature of security could also be used to trace data sources.

11) DESIGN COMPLIANCE WITH ATTRIBUTE #11 (SECURITY): Attribute #11 has two indicators: infrastructure and management security and data security and privacy. The infrastructure and management security design of IoT-based Covid-19 surveillance is divided based on a layering approach. The selection of security technology and methodology applied to the Covid-19 surveillance system should be suitable for each IoT layer. The IoT infrastructure and management security' layers include the security applied in IoT end devices, transmission media, data processing systems, and applications. In the IoT end devices, security involves various techniques that ensure the devices are authenticated and the data are encrypted to global standards. A lightweight security application is more suitable for application in IoT end devices. In the transmission media, security should cover certain mechanisms to maintain infrastructure resilience against various attacks, vulnerabilities, and threats that commonly occur on the networks, both accidental and incidental. In the data processing systems, security generally applies layered designs, ranging from physical location security to computing process applications. In IoT applications, security covers possible mechanisms to ensure that application services are secure from any threats of identity theft or of other important data while continuing to deliver valid information to the appropriate users. Users are requested to change the application system passwords regularly, and a notification will be sent to remind them. If the application is updated, then the users will also be reminded to install the updated version so that application vulnerabilities can be minimized. The overall security applied to the end-to-end IoT system for the Covid-19 surveillance design should be managed in a decentralized model, yet also be connected to a central security management system. Additionally, the implementation of security management technologies such as Syslog and SNMP is required to prepare a forensic analysis if an attack occurs.

Data security and privacy are the main concern when involving technology in the medical sector, such as the IoT-based Covid-19 surveillance system. Most Covid-19 surveillance data consist of personal and medical data, which are classified as confidential. Therefore, the IoT system design should guarantee users' data security and privacy. These can be maintained by implementing the security parameter in the IoT system software, hardware, and services. According to security standards for maintaining data security and privacy, the IoT security system should fulfill four parameters: confidentiality, integrity, availability, and non-repudiation. The confidentiality parameter protects against data abuse and leakage. If the IoT system fails to maintain data confidentiality, it might have fatal consequences for society's trust in Covid-19 surveillance participation. Simultaneously, the integrity parameter ensures that data related to Covid-19, especially data from the IoT end devices; do not change from source to destination. If there is a change, this might lead to patient misdiagnosis due to invalid data, which could endanger their safety. Data availability is also a vital parameter in the Covid-19 pandemic. Unavailability of data could interfere with the situation, since controlling the

disease without data support would be a hard task. In addition, a non-repudiation parameter in the IoT design for Covid-19 surveillance is associated with the certainty of data collected from users' IoT end devices. Non-repudiation ensures that the IoT end devices truly monitor the users whom data synchronized to the surveillance system. The synchronized users' data is beneficial for individual case handling which they can by getting more specific treatment based on their accurate medical data.

12) DESIGN COMPLIANCE WITH ATTRIBUTE #12 (STANDARDS): Attribute #12 has three indicators: standardization, regulation, and certification. Standardization in IoT-based Covid-19 surveillance systems is necessary to maintain system implementation at ideal levels. The implementation of community-based Covid-19 surveillance has various scenarios and conditions which may vary across countries. Therefore, standardization can be used as a global reference in building IoT-based surveillance systems, with the aim to stabilize and measure system performance. Global standardization involves various technical aspects, including standards for the software, hardware, protocols, and technology services used to build the surveillance systems. In practice, standardization facilitates the integration process between devices, media, and services to provide globally formatted surveillance data. Standardization of the thresholds and operations of the IoT end devices helps medical personnel to find cases precisely. Parameter standardization of network and data processing maintains the quality of the surveillance process. Standard parameters allow the system to operate in real-time or to be set following the WHO schedule. The processing system should be standardized to generate data in a standard format set by WHO, and should have the additional ability to process data in a non-standard format. In addition, standardization of system security is also needed to assure users' confidential data and preserve public trust in participating in the active surveillance of Covid-19.

Regulation of the technical details for adopting an IoT-based Covid-19 surveillance system in each country and its district could be derived from global standards. National- scale regulations need to be developed as more specific technical guidelines for designing systems following each country's applicable regulations and conditions. Moreover, regulation could overcome language barriers to the use of specific applications by individual users in different countries. Regulations are applied to the entire IoT system, starting from attached devices on the end-user side, transmission, the processing system, and data delivery. Regulations on medical devices ensure their safety, with the level of regulation required depending on the level of risk associated with the device [46]. Regulations regarding the security of both data and infrastructure also need to be devised so that the system can guarantee users' trust to be actively involved in surveillance. Furthermore, the government also needs to develop a regulation to ensure that each party involved in supplying hardware, software, and applications follows standard guidelines. This kind of regulation is needed in mass field implementation. For example, a country may have a specific regulation related to the radio frequency used that differs from those of other countries. Hence, device specifications need to be adjusted to that radio frequency. Another example is that there are many technology brands and technical details of programming application in field implementation. Consequently, chips and sensors installed in the IoT end device in one region could be slightly different from those of other regions, as may be the applications.

Certification is a way of ensuring that the quality of system performance is maintained [47]. It is needed in every implementation of IoT-based Covid-19 surveillance to keep the system quality standardized. Certification in IoT implementation is required for both software [48] and hardware [49]. Especially in the case of hardware that interacts directly with users, it is necessary to ensure that it has passed technical and safety testing and certification before use. The certification system maintains the quality standard of the surveillance in each community facing different Covid-19 transmission scenarios.

13) DESIGN COMPLIANCE WITH ATTRIBUTE #13 (MOBILITY): Attribute #13 has three indicators: tracking system, portability, handover and roaming. In IoT-based surveillance, these are applied to support the mobile design of active digital surveillance for Covid-19. In community-based surveillance, such as nursing homes, the tracking system is designed to prevent close contact interaction. A real-time position sensor, DWM1000, is applied to track community members' movements at a close distance. In the previous section, we stated that the IoT end device is configured to send an alert when close interaction occurs, thus helping to prevent the spread of Covid-19 in a community building. This real-time position sensor is part of a tracking system for indoor environments or buildings, in which short distances can be detected. However, for tracking in an outdoor and a large-scale environment, an additional sensor such as a GPS sensor is needed. It should be noted that GPS has limitations in detecting very short distances accurately; therefore, we chose a DWM1000 sensor that can detect short distances much more accurately, which is appropriate for a design employed in a nursing home. A GPS sensor is an example of a tracking system used in large-scale Covid-19 surveillance systems to monitor participants' movements during the process. Participants or users can continue performing their activities following health protocols while being continuously monitored by the system.

In IoT-based Covid-19 surveillance, besides preventing close contact, the tracking system also facilitates tracking contacts. Contact tracing using technology has been supported by WHO, which has issued guideline on such use of digital devices [50]. Contact tracing is conducted continuously, considering that Covid-19 transmission is high-risk, taking place person to person through contact and droplets. Transmission could occur to residents who interact with

probable or confirmed positive Covid-19 cases. The higher the number of interactions is, violating health protocols, the greater the need for tracing contacts. Therefore, manual contact tracing is not recommended, since it requires intensive resources and the risk of medical personnel becoming infected when performing their duties. For this reason, the IoT serves mobility design to ease medical personnel's burden through digital contact tracing with less effort, especially in community transmission areas. According to WHO provisions, contact tracing in different location settings has different criteria. The most general criteria are having a history of face-to-face interaction contact with Covid-19 patients at a minimum distance of 1 meter for more than 15 minutes, or direct physical contact with Covid-19 patients. Both criteria are valid for tracing contacts in community location settings, closed places, healthcare facilities, public transportation, and community gathering locations. Other detailed criteria are adjusted to the location setting. The contact tracing period for Covid-19 exposure determination in the community starts from 2 days before the development of symptoms, up to 14 days after such development (symptomatic cases), or 2 days before and 14 days after the date of a confirmed laboratory check (asymptomatic cases) [41].

The tracking system design in the IoT for Covid-19 surveillance helps digital contact tracing through a combination of the real-time position sensor and GPS sensors installed in the IoT end devices. The real-time position sensor DWM1000 acts as a proximity sensor that detects the interaction distance between individuals and Covid-19 patients. Simultaneously, the GPS sensor generates position coordinates to determine each suspect's travel history in the contact tracing period. Data from the proximity and GPS sensors are combined to simplify the finding of close contacts and to increase the awareness of those who have visited the same location as Covid-19 patients and have interacted with them. This data combination is also useful in a large-scale and real-time tracking system of Covid-19 patients. In areas which apply the zonation system, the combination of proximity data and users' actual locations could be used to notify the, when they are entering Covid-19 danger zones.

The portability of IoT-based Covid-19 surveillance design is applied in the form of IoT end devices worn by community members and point-of-care (POC) devices used for laboratory testing. These portable devices assist the mobility of IoT-based surveillance and accommodate users' and medical personnel's mobilization during Covid-19 surveillance. Therefore, user monitoring can be conducted continuously, and laboratory testing can be conducted flexibly, either in a fixed or mobile situation. In the nursing home, as illustrated in Fig. 5, IoT end device portability is designed to be worn on community members' wrists. On the other hand, point-of-care device portability is exemplified by the utilization of the point-of-care RT-PCR device [8] or GeNose [37][38]. Point-of-care diagnostics tools utilize a Polymerase Chain Reaction (PCR) system equipped with a chip to connect to Bluetooth on a mobile phone and then to the internet [8]. Meanwhile, Genose is a portable Covid-19 detection tool that only takes around 2-3 minutes to test and obtain a result. GeNose is beneficial for medical personnel to detect massive cases of Covid-19 using a simple procedure without reagents or other chemicals. The test is conducted by taking a breath exhalation sample, which is more comfortable than a PCR or swab test [37][38]. Overall, the portable system can detect real-time samples and has a user friendly design which simplifies the detection process and supports medical personnel mobility.

Handover and roaming in community-based Covid-19 surveillance are needed to support user movement in areas covered by the IoT network infrastructure. IoT end device operation is supported by wireless networks, which have limitations in terms of coverage area. Hence, a soft handover configuration is applied in the local wireless system, consisting of several access points or other wireless technology. The local wireless system could be in a nursing home or community member's building, and it can be integrated into a regional or national wireless system. A soft handover configuration maintains the device's connection while moving in certain wireless network coverage area. In comparison, a roaming configuration is applied in an international wireless system. In global surveillance design, roaming configuration maintains a device's movement across countries' geographical boundaries, and generally occurs in border areas. Therefore, IoT end devices and point of care devices can connect to the IoT back-end system over the IoT network infrastructure, and surveillance can take place continuously at a local and global scale. Ultimately, this IoT design supports the generation of spatial-based surveillance data without the limitations of geographical boundaries.

14) DESIGN COMPLIANCE WITH ATTRIBUTE #14 (SUSTAINABILITY): Attribute #14 indicators are maintenance, inspection, evaluation and audit. Maintenance is used to ensure the sustainability of the IoT-based Covid - 19 surveillance system. Such a process has been enforced in many countries for more than a year, since the outbreak was first detected. Sustainable surveillance has been conducted to date, following the dramatic rise in confirmed cases and the incremental number of deaths due to COVID-19 exposure. This situation indicates that a maintenance system is needed to maintain IoT performance in Covid-19 surveillance. Maintenance systems can be implemented by applying appropriate preventive and corrective maintenance measures to the IoT infrastructure and service systems in each community and back-end system. An undeniable fact is that IoT design implementation utilizes numerous devices and many types of technology that might differ in each tier. Device and service failures are unavoidable in massive technology utilization. However, if the system is well organized and monitored, failure in one part would be easier to overcome if other parts remain operating optimally [51].

Inspection of IoT-based Covid-19 surveillance can minimize system malfunctions, especially IoT end devices that directly interact with users in the community. IoT end-device applications must have no impact on public safety. Inspection of devices is made to maintain the precision of devices in detecting initial symptoms and tracing contacts accurately. Inspection in network transmission is also necessary to ensure

that it is seamless, while inspection in the data computation system can ensure the processing is powerful and trivial mistakes are avoided. In general, inspection measures can speed up the process of replacing malfunctioning devices and upgrading resources beyond scheduled maintenance. Immediate inspection of anomalous systems prevents adverse impacts on IoT implementation.

Evaluation and audit should be implemented at the community scale and in larger-scale situations running IoT-based Covid-19 surveillance systems. Evaluation and audit are conducted to preserve the operational quality and stability of the IoT-based surveillance. Proper documentation in IoT-based surveillance system development complements the evaluation and audit process. Internal and external audits should be conducted to meet management needs [27][28]. Audit helps discover and minimize any impacts and risks arising from technological and non-technological aspects. During system maintenance and inspection, changes should be administratively reported, including the replacement of malfunctioning devices, resource upgrading, software updating, and patching to fix built system gaps. Data from regular and incidental reporting can be used as a guideline for improving the surveillance system performance. Data parameters for each device configuration and service must be correctly set and evaluated to avoid instability in the surveillance system. Every action to maintain stability and improve the IoT-based surveillance system quality is recorded in the online reporting management system.

VI. Conclusion

This paper has proposed a reference model for designing an IoT-based active digital surveillance system of infectious diseases. The proposed reference model consists of 14 attributes with respective indicators. The attributes were developed by utilizing the basic 10 parameters from CDC, adapting two attributes of security and standards from the other studies, and adding the new attributes of mobility and sustainability. Each of the parameters has specific indicators which need to be considered to complement design requirements when adopting IoT technology as an infectious disease surveillance tool.

In the reference model, we highlight the mobility attribute, which plays a special role in infectious disease surveillance; namely, to control disease spread. The attribute helps medical personnel to prevent close contact, conduct large-scale contact tracing and to monitor suspects and patients on a real-time basis. The mobility attribute also facilitates the community to remain active in line with health protocols while participating in the surveillance process. Hence, we hope that the attribute can indirectly increase the community's active participation in surveillance. Another attribute that we added was sustainability. This was included as an IoT-based surveillance system is designed to be implemented over a relatively long time, and its infrastructure can be reused for different surveillance requirements. The sustainability of IoT-based surveillance is realized by implementing maintenance, inspection, evaluation, and audit of the system.

Testing of the concept was performed by adopting the proposed reference model in an IoT-based active digital surveillance system for the Covid-19 disease. We elaborated the compliance of such end-to-end design in line with the 14 attributes of our proposed reference model. The compliance was analyzed on the basis of the specific indicators of each of the 14 attributes, reflecting how the IoT design should meet the criteria of our proposed reference model. As future work, a testbed should be conducted to examine further the field implementation of the framework and its design.

Acknowledgment

This research is supported by Kementerian Riset Dan Teknologi/Badan Riset Dan Inovasi Nasional – Republik Indonesia through Hibah Penelitian Disertasi Doktor Scheme under contract number NKB-332/UN2.RST/HKP.05.00/2021, in which Prof. Dr-Ing. Kalamullah Ramli is the corresponding author. Ms Hayati is in PhD study supported by Beasiswa Unggulan Dosen Indonesia Dalam Negeri (BUDI-DN), Lembaga Pengelola Dana Pendidikan (LPDP), and a cooperation of the Ministry of Research and Higher Education and the Ministry of Finance of the Republic of Indonesia.

References

[1] S.B. Thacker, R.L. Berkelman, PUBLIC HEALTH SURVEILLANCE IN THE UNITED STATES, Epidemiologic Reviews. 10 (1988) 164–190. https://doi.org/10.1093/oxfordjournals.epirev.a036021.

[2] L. Wang, L. Jin, W. Xiong, W. Tu, C. Ye, Infectious Disease Surveillance in China, in: Early Warning for Infectious Disease Outbreak, Elsevier, 2017: pp. 23–33. https://doi.org/10.1016/B978-0-12-812343-0.00002-3.

[3] Jamison DT, Breman JG, Measham AR, et al., editors., Chapter 53. Public Health Surveillance: A Tool for Targeting and Monitoring Interventions, in: Disease Control Priorities in Developing Countries. 2nd Edition, The International Bank for Reconstruction and Development / The World Bank, Washington (DC); Oxford University Press, New York, 2006: pp. 997–1016. https://www.ncbi.nlm.nih.gov/books/NBK11770/?report=reader.

[4] U.S. Department of Health and Human Services, CDC, Public Health 101 Series: Introduction to Public Health Surveillance, (2014). https://www.cdc.gov/publichealth101/surveillance.html (accessed July 20, 2020).

[5] Md.S. Rahman, N.C. Peeri, N. Shrestha, R. Zaki, U. Haque, S.H.A. Hamid, Defending against the Novel Coronavirus (COVID-19) outbreak: How can the Internet of Things (IoT) help to save the world?, Health Policy and Technology. 9 (2020) 136–138. https://doi.org/10.1016/j.hlpt.2020.04.005.

[6] S. Sareen, S.K. Sood, S.K. Gupta, IoT-based cloud framework to control Ebola virus outbreak, Journal of Ambient Intelligence and Humanized Computing. 9 (2018) 459–476. https://doi.org/10.1007/s12652-016-0427-7.

[7] Y. Song, J. Jiang, X. Wang, D. Yang, C. Bai, Prospect and application of Internet of Things technology for prevention of SARIs, Clinical EHealth. 3 (2020) 1–4. https://doi.org/10.1016/j.ceh.2020.02.001.

[8] H. Zhu, P. Podesva, X. Liu, H. Zhang, T. Teply, Y. Xu, H. Chang, A. Qian, Y. Lei, Y. Li, A. Niculescu, C. Iliescu, P. Neuzil, IoT PCR for pandemic disease detection and its spread monitoring, Sensors and Actuators B: Chemical. 303 (2020) 127098. https://doi.org/10.1016/j.snb.2019.127098.

[9] P. Lundrigan, K.T. Min, N. Patwari, S.K. Kasera, K. Kelly, J. Moore, M. Meyer, S.C. Collingwood, F. Nkoy, B. Stone, K. Sward, EpiFi: An in-Home IoT Architecture for Epidemiological Deployments, in: 2018 IEEE 43rd Conference on Local Computer Networks Workshops (LCN Workshops), IEEE, Chicago, IL, USA, 2018: pp. 30–37. https://doi.org/10.1109/LCNW.2018.8628482.

[10] European Centre for Disease Prevention and Control., Data quality monitoring and surveillance system evaluation: a handbook of methods and applications., Publications Office, LU, 2014. https://data.europa.eu/doi/10.2900/35329 (accessed November 13, 2020).

[11] W. Ahrens, I. Pigeot, eds., Handbook of Epidemiology, Springer New York, New York, NY, 2014. https://doi.org/10.1007/978-0-387-09834-0.

[12] Updated Guidelines for Evaluating Public Health Surveillance Systems: Recommendations from the Guidelines Working Group: (548222006-001), (2001). https://doi.org/10.1037/e548222006-001.

[13] R.P. Singh, M. Javaid, A. Haleem, R. Suman, Internet of things (IoT) applications to fight against COVID-19 pandemic, Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 14 (2020) 521–524. https://doi.org/10.1016/j.dsx.2020.04.041.

[14] S.L. Groseclose, D.L. Buckeridge, Public Health Surveillance Systems: Recent Advances in Their Use and Evaluation, Annual Review of Public Health. 38 (2017) 57–79. https://doi.org/10.1146/annurev-publhealth-031816-044348.

[15] E.O. Nsoesie, S.A. Kluberg, S.R. Mekaru, M.S. Majumder, K. Khan, S.I. Hay, J.S. Brownstein, New digital technologies for the surveillance of infectious diseases at mass gathering events, Clinical Microbiology and Infection. 21 (2015) 134–140. https://doi.org/10.1016/j.cmi.2014.12.017.

[16] O.B. Leal-Neto, G.S. Dimech, M. Libel, W. Oliveira, J.P. Ferreira, Digital disease detection and participatory surveillance: overview and perspectives for Brazil, Revista de Saúde Pública. 50 (2016). https://doi.org/10.1590/S1518-8787.2016050006201.

[17] G.J. Milinovich, G.M. Williams, A.C.A. Clements, W. Hu, Internet-based surveillance systems for monitoring emerging infectious diseases, The Lancet Infectious Diseases. 14 (2014) 160–168. https://doi.org/10.1016/S1473-3099(13)70244-5.

[18] J.K. Harris, R. Mansour, B. Choucair, J. Olson, C. Nissen, J. Bhatt, Health Department Use of Social Media to Identify Foodborne Illness — Chicago, Illinois, 2013–2014, 63 (2014) 40.

[19] S.P. van Noort, C.T. Codeço, C.E. Koppeschaar, M. van Ranst, D. Paolotti, M.G.M. Gomes, Ten-year performance of Influenzanet: ILI time series, risks, vaccine effects, and care-seeking behaviour, Epidemics. 13 (2015) 28–36. https://doi.org/10.1016/j.epidem.2015.05.001.

[20] H.D. Park, O.-G. Min, Y.-J. Lee, Scalable architecture for an automated surveillance system using edge computing, The Journal of Supercomputing. 73 (2017) 926–939. https://doi.org/10.1007/s11227-016-1750-7.

[21] E. Christaki, New technologies in predicting, preventing and controlling emerging infectious diseases, Virulence. 6 (2015) 558–565. https://doi.org/10.1080/21505594.2015.1040975.

[22] P. Rattanaumpawan, A. Boonyasiri, S. Vong, V. Thamlikitkul, Systematic review of electronic surveillance of infectious diseases with emphasis on antimicrobial resistance surveillance in resource-limited settings, American Journal of Infection Control. 46 (2018) 139–146. https://doi.org/10.1016/j.ajic.2017.08.006.

[23] J. Murray, A.L. Cohen, Infectious Disease Surveillance, in: International Encyclopedia of Public Health, Elsevier, 2017: pp. 222–229. https://doi.org/10.1016/B978-0-12-803678-5.00517-8.

[24] Cisco Industrial Network Director, (2018). https://www.cisco.com/c/en/us/products/cloud-systems-management/industrial-network-director/index.html (accessed October 2, 2020).

[25] A. Demuth, R. Kretschmer, A. Egyed, D. Maes, Introducing Traceability and Consistency Checking for Change Impact Analysis across Engineering Tools in an Automation Solution Company: An Experience Report, in: 2016 IEEE International Conference on Software Maintenance and Evolution (ICSME), IEEE, Raleigh, NC, USA, 2016: pp. 529–538. https://doi.org/10.1109/ICSME.2016.50.

[26] W. Tibboel, M. Barnasconi, "Advancing traceability and consistency in Verification and Validation," in 2014 Design and Verification DVCon Conference and Exhibition Europe, Munich, Germany, Oct. 2014.

[27] Protiviti Team, "The Internet of Things: What Is It and Why Should Internal Audit Care?," 2016, https://www.protiviti.com/.

[28] I. Cooke and R. V. Raghu, "IS Audit Basics: Auditing the IoT," ISACA Journal, vol. 5, p. 5, Sep. 2018. https://www.isaca.org/resources/isaca-journal/issues/2018/volume-5/is-audit-basics-auditing-the-iot.

[29] D. Atri, H.K. Siddiqi, J.P. Lang, V. Nauffal, D.A. Morrow, E.A. Bohula, COVID-19 for the Cardiologist, JACC: Basic to Translational Science. 5 (2020) 518–536. https://doi.org/10.1016/j.jacbts.2020.04.002.

[30] World Health Organization, WHO Coronavirus Disease (COVID-19) Dashboard, Coronavirus Disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update. (2020). https://covid19.who.int.

[31] N. Hayati, K. Ramli, M. Suryanegara, Y. Suryanto, Potential Development of AES 128-bit Key Generation for LoRaWAN Security, in: 2019 2nd International Conference on Communication Engineering and Technology (ICCET), IEEE, Nagoya, Japan, 2019: pp. 57–61. https://doi.org/10.1109/ICCET.2019.8726884.

[32] M. Suryanegara, N. Hayati, An Integrated Model of Technical and Non-Technical Perspectives on Managing IoT Security, in: Proceedings of the 9th International Conference on Information Communication and Management - ICICM 2019, ACM Press, Prague, Czech Republic, 2019: pp. 142–146. https://doi.org/10.1145/3357419.3357450.

[33] M. Suryanegara, A.S. Mirfananda, M. Asvial, N. Hayati, 5G as Intelligent System: Model and Regulatory Consequences, in: Y. Bi, S. Kapoor, R. Bhatia (Eds.), Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016, Springer International Publishing, Cham, 2018: pp. 893–902. https://doi.org/10.1007/978-3-319-56994-9_61.

[34] E. Fadda, D. Mana, G. Perboli, R. Tadei, Multi Period Assignment Problem for Social Engagement and Opportunistic IoT, in: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), IEEE, Turin, 2017: pp. 760–765. https://doi.org/10.1109/COMPSAC.2017.173.

[35] G. Jonsdottir, D. Wood, R. Doshi, IoT network monitor, in: 2017 IEEE MIT Undergraduate Research Technology Conference (URTC), IEEE, Cambridge, MA, 2017: pp. 1–5. https://doi.org/10.1109/URTC.2017.8284179.

[36] S.-H. Chang, R.-D. Chiang, S.-J. Wu, W.-T. Chang, A Context-Aware, Interactive M-Health System for Diabetics, IT Professional. 18 (2016) 14–22. https://doi.org/10.1109/MITP.2016.48.

[37] Natasa Adelayanti, UGM Innovation: GeNose Can Detect Covid-19 Less Than 2 Minutes, (2020). https://www.ugm.ac.id/en/news/20122-ugm-innovation-genose-can-detect-covid-19-less-than-2-minutes (accessed December 20, 2020).

[38] Natasa Adelayanti, UGM GeNose Receives Distribution and Marketing Permission, (2020). https://ugm.ac.id/en/news/20557-ugm-genose-receives-distribution-and-marketing-permission (accessed January 2, 2021).

[39] R. Madurai Elavarasan, R. Pugazhendhi, Restructured society and environment: A review on potential technological strategies to control the COVID-19 pandemic, Science of The Total Environment. 725 (2020) 138858. https://doi.org/10.1016/j.scitotenv.2020.138858.

[40] S. Sinche, D. Raposo, N. Armando, A. Rodrigues, F. Boavida, V. Pereira, J.S. Silva, A Survey of IoT Management Protocols and Frameworks, IEEE Communications Surveys & Tutorials. 22 (2020) 1168–1190. https://doi.org/10.1109/COMST.2019.2943087.

[41] WHO Team, Contact tracing in the context of COVID-19: interim guidance, 10 May 2020, (2020). https://apps.who.int/iris/handle/10665/332049.

[42] S.-C. Cha, M.-S. Chuang, K.-H. Yeh, Z.-J. Huang, C. Su, A User-Friendly Privacy Framework for Users to Achieve Consents With Nearby BLE Devices, IEEE Access. 6 (2018) 20779–20787. https://doi.org/10.1109/ACCESS.2018.2820716.

[43] A. Khalil, N. Mbarek, O. Togni, Self-Configuring IoT Service QoS Guarantee Using QBAIoT, Computers. 7 (2018) 64. https://doi.org/10.3390/computers7040064.

[44] K.A. Delic, On Resilience of IoT Systems: The Internet of Things (Ubiquity symposium), Ubiquity. 2016 (2016) 1–7. https://doi.org/10.1145/2822885.

[45] K. A. Delic and D. M. Penkler, "Architecting Resilient IoT Systems," Project: Exascale Computing Systems, Mar. 2018. Accessed: Jan. 07, 2021. [Online]. Available: https://www.researchgate.net/publication/331072091_Architecting_Resilient_IoT_Systems.

[46] M.J. McGrath, C.N. Scanaill, Regulations and Standards: Considerations for Sensor Technologies, in: Sensor Technologies, Apress, Berkeley, CA, 2013: pp. 115–135. https://doi.org/10.1007/978-1-4302-6014-1_6.

[47] IoT Device Certification Landscape, (2019). https://www.gsma.com/iot/resources/iot-device-certification-landscape/ (accessed September 20, 2020).

[48] G. Ferreira, Software Certification in Practice: How Are Standards Being Applied?, in: 2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C), IEEE, Buenos Aires, Argentina, 2017: pp. 100–102. https://doi.org/10.1109/ICSE-C.2017.156.

[49] A. Muñoz, A. Maña, Software and Hardware Certification Techniques in a Combined Certification Model:, in: Proceedings of the 11th International Conference on Security and Cryptography, SCITEPRESS - Science and and Technology Publications, Vienna, Austria, 2014: pp. 405–410. https://doi.org/10.5220/0005098204050410.

[50] WHO Headquarters (HQ), Digital tools for COVID-19 contact tracing, Digital Tools for COVID-19 Contact Tracing. (2020). https://www.who.int/publications/i/item/WHO-2019-nCoV-Contact_Tracing-Tools_Annex-2020.1 (accessed September 25, 2020).

[51] J.V.L. do Monte, V.M. da S. Fraga, A.M.N.C. Ribeiro, D. Sadok, J. Kelner, IMMS: IoT Management and Monitoring System, in: 2018 IEEE Symposium on Computers and Communications (ISCC), IEEE, Natal, 2018: pp. 00422–00425. https://doi.org/10.1109/ISCC.2018.8538755.

Personally Identifiable Information (PII) Detection in the Unstructured Large Text Corpus using Natural Language Processing and Unsupervised Learning Technique

Poornima Kulkarni[1]
Assistant Professor, Department of ISE
RV College of Engineering, Bengaluru, India

Cauvery N K[2]
Professor, Department of ISE
RV College of Engineering, Bengaluru, India

***Abstract*—Personally Identifiable Information (PII) has gained much attention with the rapid development of technologies and the exploitation of information relating to an individual. The corporates and other organizations store a large amount of information that is primarily disseminated in the form of emails that include personnel information of the user, employee, and customers. The security aspects of PII storage have been ignored, raising serious security concerns onindividual privacy. A significant concern arises about comprehending the responsibilities regarding the uses of PII. However, in real-time scenarios, email data is regarded as unstructured text data, detecting PII from such an unstructured large text corpus is quite challenging. This paper presents an intelligent clustering approach for automatically detecting personally identifiable information (PII) from a large text corpus. The focus of the proposed study is to design a model that receives text content and detects possible PII attributes. Therefore, this paper presents a clustering-based PII Model (C-PPIM) based on NLP and unsupervised learning to address detection of PII in the unstructured large text corpus. NLP is used to perform topic modeling, and Byte mLSTM, a different approach of sequence model, is implemented to address clustering problems in PII detection. The performance analysis of the proposed model is carried out existing hierarchical clustering concerning silhouette and cohesion score. The outcome indicatedthe effectiveness of the proposed system that highlights significant PII attributes, with significant scope in real-time implementation. In contrast, existing techniques are too expensive to function and fit in real-time environments.**

Keywords*—PII; natural language processing; word2vec machine learning; PII detection; security

I. Introduction

The progressive digitization of functional domains of various processes in individual human and business contexts produces various data types. The data are generated in text format, audio format, video format, image format, and many more custom formats. The business objectives often demand to store or archive these data for longer, making it voluminous. The analogy of various formats of data and larger size of it is popularized in the recent past as verity and volume, respectively [1]. The ever-evolving business models at the pace of technological advancements have provided possibilities of innovative business applications in the healthcare industry, banking & finance, education, aviation, Défense etc. There are many contexts where certain information in these data is very private to a user or a system [2]. Providing necessary security to this private information becomes essential to mitigate associated risk due to system design vulnerabilities, potential threats, and attacks. The study of security towards this private information is popular as privacy preservation. The complexities and challenges of designing effective privacy preservation methods depend solely on the type of the data format, its size, and the data flow into the application. However, another popular term appears in the context of designing security models to preserve privacy – "Personally Identifiable Information" (PII). The data which can be used to identify a person is the higher layer meaning of PII [3]. Fig. 1 shows the relationship between PII and private information.

In the last decade, the increasing popularity of the Internet and PII collection has raised serious concerns about privacy policy. A surge in personal data breaches for an in-depth understanding of preferences, authenticating customers and employees has become a common occurrence in data-driven organizations. Even government organizations heavily rely on the collection of PII to carry out an important decision. Therefore, people are required to share their personal data. However, a significant concern arises towards comprehending the responsibilities regarding the uses of PII and its protection [4]. More specifically, the people are not knowing, how the government organization and corporates are handling individuals' PII. PII collection over the Internet is highly prone to identity theft, social engineering attacks and is most vulnerable to fraudulent criminals [5]. The evidence in the report provided by the Data Breach Level Index shows that 76.20% of data breaches in 2018 belonged to the social media industry, most of which were related to identity theft [6]. It is not surprising that PII has already turned out to be the new resource, and threat modeling for privacy is at the peak of the industry 4.0 revolution [7]. In addition, the cases of data breaches in the past recent years fascinated a compulsive concern towards PII in the research community. In response, significant efforts have been devoted in the existing literature to developing solutions against PII disclosure. The literature presented many solutions considering different contexts, such

as PII on social networking sites, mobile applications, mobile network traffic, healthcare, corporate internal communications, and many more [8-10]. However, PII identification schemes in the existing literature do not provide comprehensive insight. Most of the existing schemes are based on rule-based approaches, limiting the scope and applicability of existing solutions in the real context. The rule-based approaches are based on the set of procedures and principles to represent knowledge from the structured data. Since, in real-world cases, the organization mostly maintains a large corpus which stores PII in the textual data format such as emails, contracts, IPv4 and MAC addresses, and telephone numbers. Among these textual data, emails contain more entropy compared to the other textual data. Apart from this, the textual corpus, especially email, is mostly unstructured since the information is presented in the native format, especially the email contents written with the different writing styles, contains the short subjective textual body. Therefore, the rule-based approaches are not much suitable for identifying PII from the unstructured large text corpus because they mainly deal with structured data formats. However, with the advent of machine learning (ML) models and advancement in natural language processing (NLP), PII of individuals from large unstructured text corpus can be efficiently identified, which cannot be addressed by applying the existing rule-based solution discussed so far [11-13]. In the existing literature, many efforts have been put forward by the researchers. But, the research work on detecting PII in unstructured text corpus is minimal. The existing literature does not focus on detecting full PII and its diversity that reflects user privacy and identity differently. A particular type of PII is easy to be identified, but it is crucial to consider the sources or categories of the content that reflects highly vulnerable PII. Therefore, the existing study lacks topic modeling and also suffers from huge computational overhead.

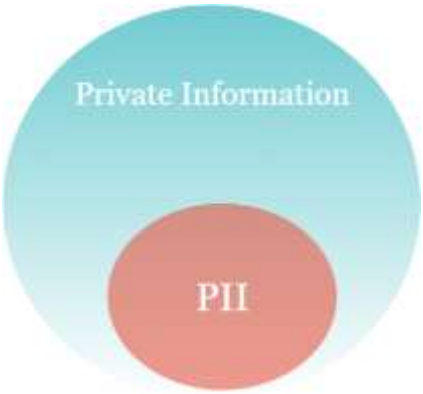


Fig. 1. Privacy Information and Personally Identifiable Information (PII ⊂ PI).

In the current study, the email dataset is considered an unstructured text corpus for the model evaluation. The email is mainly composed of low-quality text and one of the potential sources of the PII disclosure. The proposed research considers PII detection from the unstructured text corpus as a text mining and clustering problem. It introduces a clustering-based PII detection model (C-PIIM) using natural language processing (NLP) and an unsupervised learning algorithm. NLP is employed to perform topic modeling using Word-to-Vec and Bag of Word models. On the other hand, Byte-mLSTM as an unsupervised learning algorithm is used to detect full-PII from the text corpus. The performance validation of the proposed system C-PIIM is carried out concerning clustering performance metrics and PII probability in the text corpus. The significant contribution of the proposed work can be summarized as follows:

- The proposed system addresses the problem of automatically detecting and classifying the possibly included PII attributes from the large text data.
- The study also addresses the problem of precise feature extraction from low-quality textdata by introducing an effective data modeling and processing mechanism.
- A topic modeling is done to determine a set of contexts that show which category of text document has the most probable vulnerable PII.
- The hybrid nature of the deep learning technique is employed to achievehigher accuracy inclassifying the PPI from the email or text data.

The design and development of the proposed system are carried out in such a manner that it can meet corporate production requirements by automatingmonitoring and detecting PII and ensuringagreement.

The remaining sections of this paper are arranged in the following manner. Section II discusses the related work on personnel data leakage and PII detection; Section III presents the proposed system C-PIIM design and its implementation, followed by systematic data modeling and discussion; Section IV presents the result and discussion; finally, Section V concludes the entire work of this paper.

II. Related Work

The incidents of information breaches and openings attracted the wide attention of security agencies and government, particularly if it encompasses PII. Corporate companies, healthcare industries, and social networking sites are the most attractive targets for the attackers looking for PII, which can be used for identity theft, and even unauthorized access to sensitive data. Over the years, personal data leaking has been widely studied in the existing literature. A recent study by Go et al. [14] raises serious concern on the opening and un-intended disclosure of information on social media platforms. The authors suggest an intelligent and flexibly centralized model based on software-defined networking with virtualization to counteract unintended PII opening on the network traffic. This work improves the previous work done by Liu et al. [15], where automatic detection of PII is carried out by employing a set of systematic expressions and dictionary-based methods. Another work in this context by Ren et al. [16] proposes a model, namely ReCon, to address the discloser of PII in the mobile network traffic using the supervised learning model. ReCon leverages the decision tree to identify and block the open PII in the mobile network traffic.

The application of the learning technique, as in the work of Noever [17], proposes a method to identify the Person of interest (PoI) and PII from the email dataset of the corporate

company. This method uses a mechanism of ensembled learning that considers decision trees, support vector classifier, random forest, and neural network. The PoI and PII are identified from the text analysis, such as financial records and emails. This study also analyses the sentiments of several employees regarding corporate crises. Zaeem and Barber [18] focus on the increasing PII misuses over the Internet. This study reveals an interesting statistic concerning lacking user privacy preservation incorporate companies in North America. The author collected data from the stock exchange platform and labeled the collected information with different rating scores. This study provides an effective direction so that corporate industries can improvise their privacy law.

In critical applications such as healthcare, researches are carried out to find useful information from the medical data or records to aid emergency management. However, the healthcare records usually consist of patient information, and the researchers often encounter PII that needs to be protected. In this regard, Michael et al. [19] discussed securing PII for the human participant and health research using a big data tool. Similarly, Alnemari et al. [20] focus on protecting PII available in healthcare data. The authors have presented an interesting discussion on the existing techniques for protecting PII. The study findings showed that multiple attribute workload distribution is more effective than the traditional anonymization and differential privacy approaches to obscure PII while allowing the researchers to conduct analysis efficiently.

The work of Onik et al. [21] presented an intelligent risk classification model based on accumulated mobile application permission data to classify vulnerable PII associated with the mobile owners. The authors have developed a google-play API to collect permission data of several android applications. They adopted different ML classifiers to detect the most significant PII such as contact number, social graph, email, location, biometric ID, and a unique ID. This study has presented significant work, but it has achieved less accuracy than similar existing research works due to training the model with fewer data.The work of Majeed et al. [22] suggested an improved vulnerability-aware PII anonymization scheme to ensure user privacy. The author used random forest mechanisms to identify the identity of the most vulnerable PII and used the Simpson index to calculate the diversity to reduce the risk of PII disclosure.

The work of Venkatanathan et al. [23] examined the impact of public and private data opening on social media toward disclosure of the PII to strangers. This study has demonstrated that the wall posts and extensive descriptions of individuals on their profile pages in social networks trade significant privacy leakage to the world. The authors have also presented an analytical design for privacy and PII masking. The work of Tesfay et al. [24] studied the challenges and issues in discovery of PII from textual data using ontologies, NLP and learning approaches. This study has considered the both PII and privacy sensitive information with all relevant definitions, and terminologies. A systematic approach is carried out to explore different types of the problem considering the information which needs to be regarded. Liu et al. [25] provided significant work on the identification of PII based on the information related to the user behavior in the network traffic. Firstly, the authors have discussed on the role of application and internet service provider that collects data user information to enhance their quality of experience, traffic control and improve security services. Further, the authors have discussed the challenges in identifying PII from the massive network traffic. The authors then presented an efficient algorithm, namely TPII, to address the detection of PII in the massive and complex nature of traffic data. A concept of a decision tree with an optimization approach is used to perform the classification of the PII. This study uses a dataset of real data gathered from a university network with more than 10k users. The work of Vishwamitra et al. [26] suggested a collaboratively controlling mechanism for protecting PII for photo sharing over the social network. The presented scheme is designed based on the multiparty access control mechanism and policy specification-based detection process.

III. Research Problem

In literature, a number of research works have been presented for privacy security concerning PII. This section highlights some significant issues associated with the existing approaches as follows:

- It has been analyzed that research works on detecting PII in unstructured text corpus using ML and NL is quite limited.
- Most of the existing works do not providea comprehensive insight into data modeling and its processing regarding PII discovery.
- However, they have achieved a good result for structured data, but the existing techniques may show poor performance for low-quality data, like email,which is often unstructured as it contains many acronyms, short-text, and errors.
- The existing studies lack topic modeling, which is important when dealing with private information discovery.
- It has also been noticed that the previous prediction-based approaches based on the traditional machine learning and statistical techniqueslack efficiencyandhence, suffer from huge computational overhead to achieve higher accuracy in the privacy information detection and classification process.

Hence, the problems mentioned above are addressed in the proposed study. The next section details the proposed system and strategy adopted in its implementation.

IV. System Design and Implementation

The proposed research study focuses on detecting PII from the large unstructured text corpus to advance privacy practices and create better regulations in the communication processes. The study believes that the chain of custody of the data, i.e., information in the internal communication, must not be broken and openly disclosed. The PII in the data must be retrieved only when it is necessary.Therefore, an effective and automated system is developed to identify PII from the large text corpus that consists of emails and text messages. The

current work does not intend to detect quasi PII. Instead, the method is devised to identify full PII to check if there is a direct identifier such as name, email address and social security numbers etc., based on the subject and contents of the text corpus. The introduced model adopts the application of ML and NLP for clustering the full-PII identifiers and non-PII identifiers from a large unstructured text corpus. The schematic architecture of the proposed model is described in Fig. 2.

The implementation of the proposed model follows a systematic data modeling to achieve a suitable and precise feature vector for the clustering process using unsupervised learning for PIIidentification from the unstructured text corpus. The system design consists of a total of five core modules, namely, i) Dataset and its importance, ii) Dataset visualization, iii) pre-processing, iv) Topic modeling using NLP, and v) mLSTM implementation for PII identification.

A. Dataset and its Importance

This section presents a brief description of the dataset and its importance in the context of PII detection in corporate companies. A collaborative consortium of the 22 institutions under the flagship of the Artificial intelligence center of SRI, international has initiated a project, namely, "for Cognitive Agent that Learns and Organizes (CALO)". The "Enron email dataset" was collected during the project CALO. There is an interesting background behind the creation of this dataset an investigation study by the Federal Energy Regulatory Commission (FERC) behind the malpractices followed within the eco-system of the Enron corporation – a flagship company that has revenue up to 101 billion USD till the year 2000 and became bankrupt in the year 2001. The FERC made the Enron-email dataset public for the first time, containing email communication data from 150 senior management with approximately half a million messages. The various acquisitions and transitions in the dataset have taken place from time to time. In this paper, a version of the dataset published by Carnegie Mellon University in 2015 is used, obtained from the Kaggle. The top contributors are the researchers majorly from the two countries, including the United States and India. The statistics in Table I show its current popularity status.

B. Dataset Visualization

In this section, exploratory analysis is carried to visualize the characteristics and attributes of the data set to comprehend the need for pre-processing over the text dataset for further processing in the topic modeling. Since this is the initial phase of the proposed system implementation, the first step is to import the dataset (DS) available in the form of .csv to the data frame (DF) of the computing environment, which represents the entire structure of DS in tabular representation with rows (R) and columns (C). The structure of DF is illustrated with text samples in Table II.

In Table II, visualization of DF is carried out that represents a total of 517401 categorical data $\in$ R corresponding to 2 headers fields such that {file, message} $\in$ C, where the file (F) refers to the actual placewhere all the mail is stored that contains user information (U_I), type of inbox (T_{inb}) and subfolder (S_F) on the mail, such that $F \in C \supseteq \{U_I, T_{inb}, S_F\}$. On the other hand, message (M_{sg}) consists of the email content (body) and header (H) of the email. A sample view of $M_{sg}$is shown in Fig. 3.

Further, the analysis on the 'file'$\in$ C is carried out towards identifying its uniqueness, and it is found that it does not associate with any repetitive and duplicate entries. The next section presents pre-processing operation to split the header and body of the M_{sg} for the PII detection.

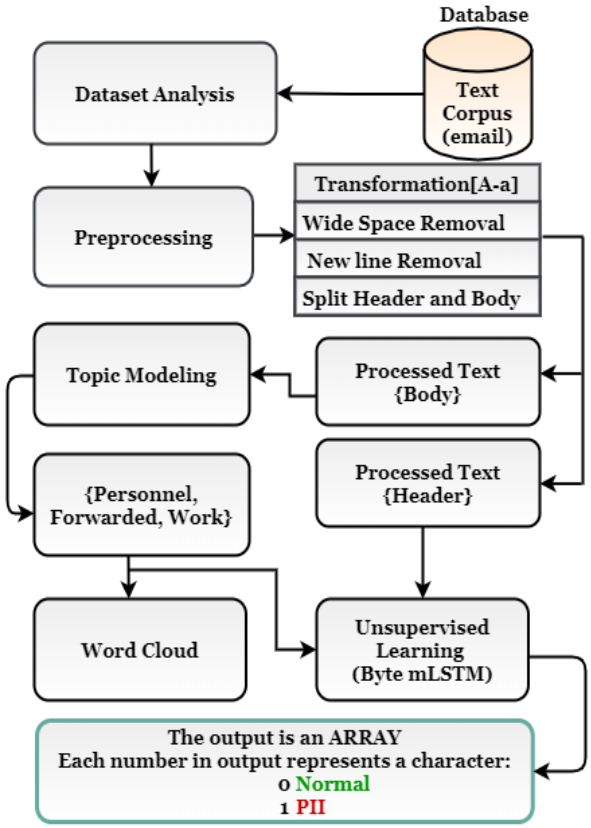


Fig. 2. Schematic Architecture of the Proposed Model.

TABLE I. POPULARITY INDICATOR OF THE ENRON-EMAIL DATASET

Views	Downloads	Download per view ratio	Total unique contribution
245000	29300	0.12	180

TABLE II. VISUALIZATION OF DF

SI. No	file	message
0	allen-p/_sent_mail/1.	Message-ID: <18782981. 1075855378110.JavaMail. e
1	allen-p/_sent_mail/10.	Message-ID: <15464986. 1075855378456.JavaMail. e
2	allen-p/_sent_mail/100.	Message-ID: <24216240. 1075855687451.JavaMail. e
3	allen-p/_sent_mail/1000.	Message-ID: <13505866.1075863688222.JavaMail.e
⋮	⋮	⋮
517400	zufferli-j/sent_items/99	Message-ID:<28618979. 1075842030037.JavaMail. e

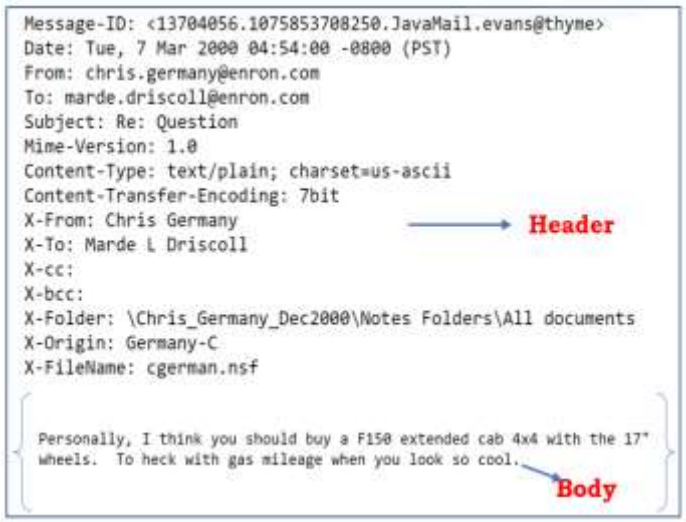


Fig. 3. Illustration of the Message Field in the Dataset.

C. Dataset Preprocessing

Based on the exploratory analysis, it has been analyzed that the email dataset is associated with a different form of text representation; some are small, and some texts are capital in between the sentences. It has also been identified that most email bodies have wide spaces and new line characters, which may introduce uncertainty and ambiguity in the learning model. Therefore, the study considers removing wide spaces, newline characters, and transforming capital letters into small letters. Further, the algorithm performs a data splitting operation to separate the header and body of the messages filed of the email dataset. The above-mentioned algorithm demonstrates pre-processing operation carried out for removing irrelevancy in the dataset, which after processing provides Header and Body of message field as a separate vector. However, the data in M_{sg} also consists of stop words and punctuations with low entropy, but in the current study, it is considered significant in topic modeling.

Algorithm-1email Data Pre-processing

Input: M_{sg}
Output: Header (H) and Body (B)
Start:
Init H →[ ], B →[ ]

$$\text{Load} \rightarrow \text{DF}[M_{sg}]$$

def function: email_prepros(M_{sg})

text_lower = re.findall(M_{sg}, '(.|||[a − z]|||[A − Z])))**do**

$$M_{sg} = \text{text.lower}(M_{sg})$$

$$M_{sg} = \text{text.replace\_arg}(M_{sg}, :\backslash n + ')$$

[M_{sg}]←Text.split(M_{sg})

return = H, B

For each R **from** M_{sg} **do**

$$\text{DF}_p = \text{email\_prepros}(M_{sg})$$

H←append.Header
B←append.Body
End

D. Topic Modelling

The PII has different nature thatreflects user privacy and identity differently. The identification of a certain type of PII is not a much challenging task. However, identifying the most vulnerable PII in unstructured text corpus is a challenging task. If such vulnerable PII is identified or exposed in the data, their presence may increase privacy risks.Once compromised, attackers can use the data to obtain PII to harm users by social engineering attacks or identity theft. Therefore, organizations or individuals must assess vulnerable PII to reduce the risk of privacy leakage. Therefore, the proposed study performs topic modeling to examine a set of documents and extract the categories and groups that reflect probable vulnerable PII. The proposed study considers topic modeling as an NLP problem and adopts the Word2vec model [27], an unsupervised text clustering model. The model takes the document text or email body (B) as input and gives a topic vector as an output. The document is converted to a vector with the help of the bag of words (BoW) technique, which is a word matrix that represents the count of each word in B. The sample representation of the word matrix for text data is illustrated in Fig. 4 where each row (R) of the matrix represents a B and the column represents tokenized text contained in the B of the email.

The word matrix obtained from the BoW is then fed to the word2vec algorithm to group emails into a cluster based on the topics needed for the study. However, the number of clusters can be limited, and in the current study, it has been limited to three clusters: personnel emails, work emails, and forward emails. Fig. 5 illustrates the working procedure of the word2vec algorithm m.

	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6

Fig. 4. A Bag of Words Algorithm

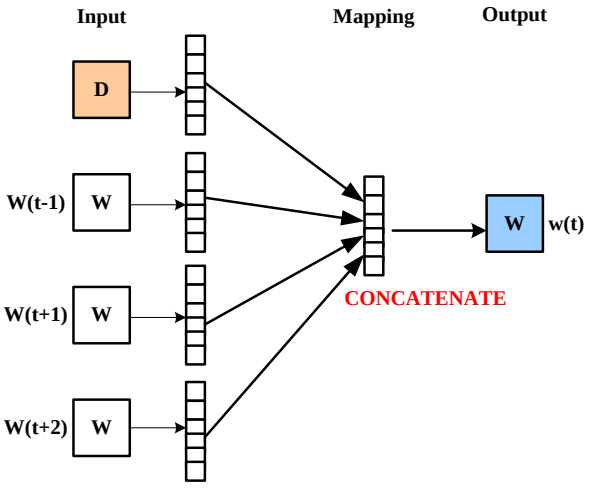


Fig. 5. Architecture of Word2vec Algorithm.

Once the model is trained, it can be used for text partitioning. Text partitioning is a collective term used to assign a topic to the text corpus and separate the corpus based on the topics extracted. In this case, the text is separated based on whether the email is personal or work or forwarded. This classification is preferred since, logically speaking, personal mails may reveal more PII than work email and forwarded. It should be noted that the work email contains more sensitive and confidential data than personal email. However, the scope of this study is to protect individual privacy rather than protecting the company's data. Other modules will be in a position to prevent data leakage. With this, the topics will be assigned to respective clusters, and new data is obtained that willbe further provided as an input to the next algorithm for PII detection. The output of the word2vec for topic modeling is demonstrated using word cloud as follows.

The word cloud of personal email is shown in Fig. 6, which reveals that the words are mostly related to personal topics like vacation planning or ranting about the employers etc. This is the place where most of the quasi and full PII can be exposed since people may speak about their family and friends in such emails. The next word cloud is shown for the work email.

In Fig. 7, the word cloud is shown for work emails that describe most of the words repeated here are related to the company's fuel and energy, which is the company's main domain. Since most of the discussion is about the company work, the chances of exposure of PII are significantly less. However, the people speak about their colleagues and employers, and hence there will be PII exposure in these emails as well. For example, the word 'Pallen'shown in the word cloud is an important person in the company.

Fig. 6. Word Cloud for Personal Email.

Fig. 7. Word Cloud for Work Email.

Fig. 8. Word Cloud for a Forwarded Email.

The word cloud for forwarded email isdepicted in Fig. 8 that showsdiscussions were more about the general context, such as jokes or some commonly shared information. These types of emails may contain very little PII. The next section presents an unsupervised learning model, namely byte mLSTMalgorithm, for detecting PII in the text data Corpus. Byte mLSTM is an unsupervised approach for NLP or text processing [2]. In the present study, it is being used to detect PII.

E. Byte-mLSTM

Long-short-termmemory (LSTM) is an improved version of Recurrent Neural Network (RNN) model, suitable for sequence classification problems. In the proposed study, a different form of LSTM, i.e., byte mLSTM algorithm, is used to perform the clustering of text into normal text and PII. The mLSTM is given by Krause et al. [28] designed based on the joint approach of LSTM and multiplicative RNN architecture. The detection of PII in the proposed study using Byte m-LSTM can be described as clustering problem of n samples such that $X = \{x_1, x_2 x_2 \cdots L\}$ into K groups described by c_i using a nonlinear mapping function $f_\alpha: X \rightarrow Y$, where x denotes input sequences of length L where $X \in R^{d\times 1}$ and $c_i$denotes centroid and $i = \{1,2,3 \cdots K\}$, α is the training parameters, and $Y \in R^K$is the embedded feature space from the model. For a given input sequence X to the learning model Byte m-LSTM, the encoder $E_i \epsilon R^c$ at time t_i and LSTM cell c and Decoder combinedly trained to reconstruct input feature vector space X in inverse order by reducing the following objective as follows:

$$\sum_{X\epsilon N_t} \sum_{i=1}^{L} ||x_i - x_i'||^2 .... \quad (1)$$

Where, N_t refers to the training set sequences, and L output state of the encoder is used as the initial input x_i to the decoder part of the the-LSTM model. The decoder part then gets a state $S \epsilon D^{i-1}$ and maps to the expected output space x'_{i-1} corresponding to the target x_{1-1} vector space. In the training phase, the model employed CAHL (Clustering assignment hardening loss) function, which is explicit and most suitable for the clustering mechanism. The proposed study considers stabilized resemblances between data points and c_i as a soft assignment (q_{ik}) of k with statistical estimation technique to measure the resemblance between learned features y_i and c_i. The probability of assigning sample i to k can be numerically represented as follows:

$$q_{ik} = \frac{\left(1+||y_i - c_k||^2/\varphi\right)^{-\varphi+1/2}}{\Sigma_k\left(1+||y_i - c_k||^2/\varphi\right)^{-\varphi+1/2}} \quad (2)$$

Where y_i corresponding to $x_1 \epsilon X$ after feature representation and φ denotes the degree of freedom. In the proposed study,φ is constant, which is considered equal to 1. In the next step of the resemblance between the distributions is evaluated using KLD by reducing the distance between q_{ik} and the supplementary distribution P and probability distribution Q as follows:

$$KL(P||Q) = \sum_i \sum_j p_{ij} \log {}^{p_{ij}}/_{q_{ij}} \quad (3)$$

Where KL denotes relative entropy, p_{ij} and q_{ij} ϵP and Q, respectively, are enhanced and adjusted through the backpropagation mechanism of the neural network, resulting inthe reduced distance between the data samples corresponding assigned k towards a better quality of clusters. However, the data points of a cluster closer to the average value of another cluster may adversely impact the loss. Therefore, q_{ik} is further normalized based on the frequency for each cluster K and raised by power factor 2 numerically expressed as follows:

$$p_{ij} = \frac{{}^{q_{ij}^2}/_{f_i}}{\Sigma_j {}^{q_{ij'}^2}/_{f_{i'}}} \quad (4)$$

where, f_i denotes q_{ik} frequency

The proposed architecture of mLSTM for text clustering is given in Fig. 9, consisting of two blocks, i.e., encoder and decoder LSTM. The encoder parts consist of multiple LSTM cells in parallel to the decoder part.

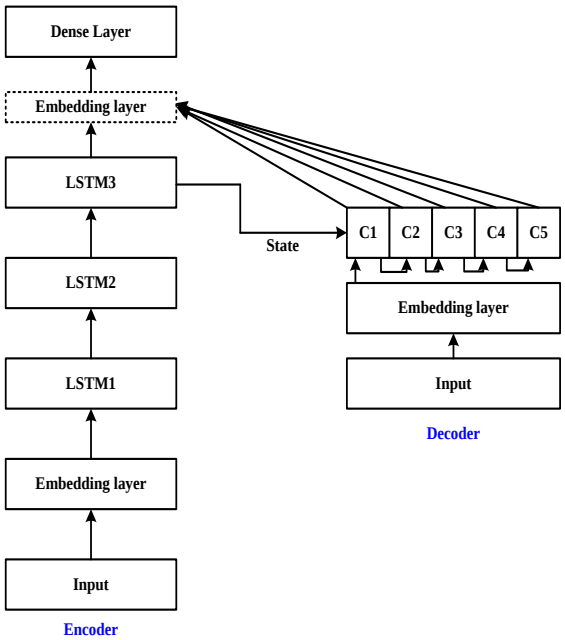


Fig. 9. The Schematic Architecture of mLSTM.

It can be observed that the LSTM-1 layer of the decoder network is placed parallel to LSTM 3 of the encoder layer. In the proposed study, the input to the learning model is carried out in the sequence of ASCII code of each letter in the input text. The encoder is an LSTM layer that accepts delayed input in the form of a vector forwards it to LSTM one by one in a sequence. The training of implemented learning model is carried out by loss function, namely, CAHL.The model takes input as H, clusters obtained from previous Algorithm word2vec, i.e., personnel (P), work (W), and forwarded (F) with the help of an encoder at the input layer. The encoder takes the input {H,P,W, F}and converts every character into ASCII code, which is fed one by one in a sequence. The model considers a sequence encoder instead of a window function, as the sequence encoder always works better in NLP problems. At the same time, the Decoder is takingthe output of the previous algorithm, i.e., word2vec, and also header information which is non-sequential data. In this process, layer 3 of the encoder and layer 1 of the Decoder functions parallelly. The output of the Decoder LSTM-1 layer is multiplied by the output of encoder LSTM-3. The multiplication operation in m-LSTM is carried out using an embedding layer that acts as a multiplier, letting the encoder input pass or stop based on the output of the decoder layer. The decoder LSTM also changes its output according to the input given in the encoder, and due to this reason,the LSTM-3 layer of the encoder is made aware ofthe LSTM1 layer of the Decoder.

V. Discussion and Perspective

The management of private information and its legislation is one of the important issues for organizations today. For instance, email is one of the main sources of communication in most corporate companies. However, this is also one of the potential sources of privacy leaks. In this regard, companies need an effective technology that can automate the process of managing personal information, assisting in monitoring PII leaks in the workflow process.Although the previous works have shown considerable efforts in private information discovery, to date, none of the existing techniques are more effective at identifying potentially vulnerable PII and its source. The proposed study attempted to address this issue by designing an effective mechanism of topic modeling that shows different categories of text documents containing sources and different aspects of the vulnerable PII. The proposed work focuses on detecting full PII followed by clustering operations that allow to determine and derive valuable patterns between text elements that help to offers important meaning to distinguish structures of PII. Unlike previous techniques, the proposed study uses hybrid nature of deep learning mechanism that detects and highlights the presence of PII in the given text data without depending on other external resources and rule-based approaches. The advantage of the adopted learning mechanism is that it achieves better generalization by taking advantage of both LSTM and multi-placative RNN (m-RNN) algorithms. The LSTM is good at handling natural language as sequence classification problems. With the implementation of a suitable embedding and encoding layer, LSTM can generalize the actual meaning in the text data. The advantage of m-RNN is that it generalizes long-term dependencies between the input

text feeds and achieves stability in the overall learning process.Therefore, the learning model of unsupervised nature employed in the proposed system is efficient in addressing the problem of modeling text elements under a variable context. Another significance of the proposed model is that it can support and handle large and unstructured text corpus and faster responsive features. The model's design is user interactive, flexible to different lengths of data usage, and offers a better discovery of private information than the existing system.

VI. Result and Performance Analysis

The design and implementation of the proposed system are carried out using the Python programming language. The current study's entire work followed systematic data modeling and a clustering approach to identify PII from the unstructured text corpus.The outcome obtained for the proposed C-PIIM based on NLP and unsupervised learning employed in the study yielded promising results. This section discusses the results and performance analysis of the proposed C-PIIM concerning the type of mails, message head, and body. Also, the model's effectiveness is justified based on the comparative analysis of the proposed clustering method with the existing hierarchal clustering method [29] regarding two clustering performance metrics such as silhouette and cohesion score. The performance metric cohesion can be defined as a performance indicator that shows how fit the class methods are identical to each other. The performance metric silhouette can be described as a performance indicator that measures the quality of a clustering technique, numerically expressed as follows:

$$\text{silhouette} = \frac{(q-p)}{\max(p,q)} \quad (5)$$

The analysis from Fig. 10 shows thata maximum number of texts are within 10,000 – 20,000 characters. The total number of characters includes the characters in the header as well as the body and whitespaces. There are some mails whose length goes up to 2,00,000 characters upon further inspection; such lengthy mails are regarding knowledge transfer. However, a significant observation can be made from the analysis that most PII is being exposed in shorter mails. The following analysis in Fig. 11 is carried out regarding PII identification (%) in various mails determined from the topic modeling.

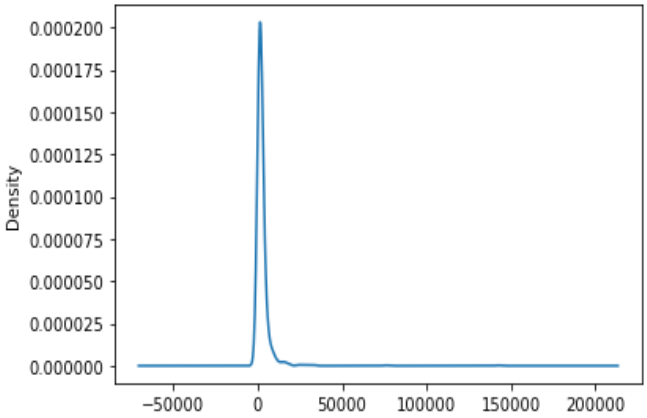


Fig. 10. Histogram for the Length of Emails in a Number of Characters.

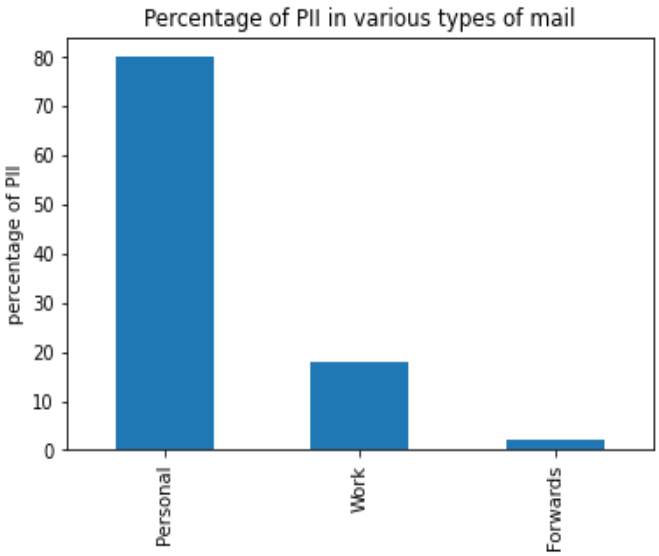


Fig. 11. Percentage of PII in Various Types of Mail.

The graph trend from Fig. 11 exhibits that personnel mail has a higher percentage of PII exposer than the work category and forward category of mails. The outcome shows 80% of the PII is exposed in personal mail, 18% is exposed in work emails, and only around 2% is exposed in the forwarded emails. Therefore, from the analysis, most of the PII is exposed in the personal mails.

In Fig. 12, analysis is carried out concerning message contents. The graph trend exhibits, the headers contain more PII compared to the body contents. This comes as no surprise as the header always contains the email and name of both sender and receiver. However, this will be prevented by the system for exposure of PII. The main focus is the body of the mail, where 20% of the PII is exposed. It exposes vital data which may be used to cause a financial loss to the individual. The next analysis presents the proposed system's comparative analysis with the existing hierarchical method regarding culturing performance metrics.

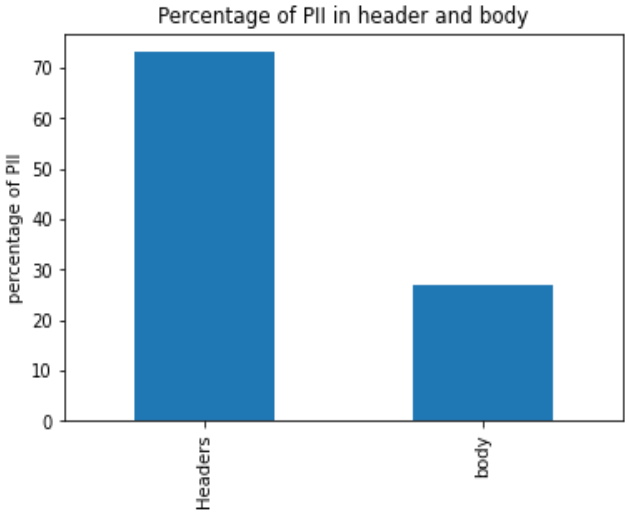


Fig. 12. Percentage of PII in Header and Body.

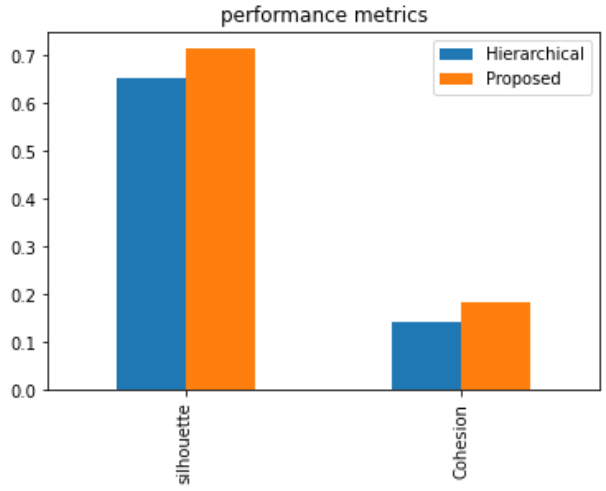


Fig. 13. Comparative Analysis.

Fig. 13 exhibits that the proposed C-PIIM outperforms the existing method in terms of silhouette and cohesion metric. From the analysis, the proposed system achieved 0.7156 and 0.1832 silhouette and cohesion scores, respectively. In contrast, the existing method has achieved 0.1421 and 0.1832 silhouette and cohesion scores, respectively, proving the proposed model'seffectiveness and applicability for real-time implementation. The next Fig. 14 presents the outcome achieved by the proposed model C-PIIM.

Fig. 14. Detection of PII.

Fig. 14 exhibits the outcome of the proposed C-PIIM that highlights the presence of PII in the email (text corpus) of the user, that contains addresses unique to sender and receiver as well as it also reflects the email address of some senior person of the company, which may be vulnerable to social engineering threat or identity theft.

VII. Conclusion

As PII collection continues to increase, the costs of data breaches also increase, ranging from economic losses to reputation losses. For understanding the risks associated with privacy opening, several efforts have been made in the literature to detect PII disclosure and leaks. However, there are limited works regarding PII detection in the large unstructured text corpus. In this paper, NLP and Byte-mLSTM mechanisms are used to design an effective model for the purpose of PII leak detection.The proposed system comprises topic modeling for segmenting and grouping data storage categories that account for disclosure of potential or most vulnerable PII. Byte mLSTM as unsupervised learning is employed as an effective clustering mechanism to detect vulnerable PII in the text data. The study outcome proved the effectiveness of the proposed clustering-oriented PII detection compared to the existing hierarchical clustering approach. The proposed model can be used in real-time scenarios like in the corporates to warn their employees against sending PII included in the email. However, the model is limited to detection of full PII, cannot detect Quasi PII. The proposed study will be extended in future work considering the quasi PII and de-identification process.

References

[1] S. Chenthara, H. Wang, K. Ahmed, “Security and Privacy in Big Data Environment”, In: Sakr S., Zomaya A.Y. (eds) Encyclopedia of Big Data Technologies. Springer, Cham, 2019.

[2] M. Petrescu, A.S. Krishen, “Analyzing the analytics: data privacy concerns”, J Market, vol. 6, pp. 41–43, 2018.

[3] F. Alizadeh, T. Jakobi, J. Boldt, and G. Stevens, “Gdpr-reality check on the right to access data: Claiming and investigating personally identifiable data from companies”, In Proceedings of Mensch und Computer, pp. 811-814, 2019.

[4] Boyd JH, Randall SM, Ferrante AM. Application of privacy-preserving techniques in operational record linkage centres. Medical data privacy handbook. 2015:267-87.

[5] Pawlicka A, Choraś M, Pawlicki M. Cyberspace threats: not only hackers and criminals. Raising the awareness of selected unusual cyberspace actors-cybersecurity researchers' perspective. InProceedings of the 15th International Conference on Availability, Reliability and Security 2020 Aug 25 (pp. 1-11).

[6] “Data breach statistics.” https:// breachlevelindex.com/, Retrieved on 25th September 2021.

[7] M. M. H. ONIK, C. KIM and J. YANG, "Personal Data Privacy Challenges of the Fourth Industrial Revolution," 21st International Conference on Advanced Communication Technology (ICACT), pp. 635-638, 2019.

[8] A. Iyengar, A. Kundu, G. Pallis, “Healthcare informatics and privacy”, IEEE Internet Computing, vol. 22(2), pp. 29-31, 2018.

[9] D. Hiatt, Y.B. Choi, “Role of security in social networking”, International Journal of Advanced Computer Science and Applications, vol.7(2), 2016.

[10] D.R. Pope, Y.H. Hu, M.A. Hoppa, “A Survey on Securing Personally Identifiable Information on Smartphones”, Virginia Journal of Science, vol.71(3), 2020.

[11] P. Silva, C. Gonçalves, C. Godinho, N. Antunes, M. Curado, “Using natural language processing to detect privacy violations in online contracts”, InProceedings of the 35th Annual ACM Symposium on Applied Computing, pp. 1305-1307, 2020.

[12] R.N. Zaeem, M. Manoharan, Y. Yang, K.S. Barber, “Modeling and analysis of identity threat behaviors through text mining of identity theft stories”, Computers & Security, vol. 1, 65, pp. 50-63, 2017.

[13] S. Applebaum, T. Gaber, A. Ahmed, “Signature-based and Machine-Learning-based Web Application Firewalls: A Short Survey”, Procedia Computer Science, Jan 1;189:359-67, 2021.

[14] S.J. Go, R. Guinto, C.A. Festin, I. Austria, R. Ocampo, W.M. Tan, “An SDN/NFV-enabled architecture for detecting personally identifiable information leaks on network traffic”, In Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), pp. 306-311, 2019.

[15] Y. Liu, H. H. Song, I. Bermudez, A. Mislove, M. Baldi, and A. Tongaonkar, "Identifying personal information in internet traffic," in Proceedings of the 2015 ACM on Conference on Online Social Networks, COSN '15, (New York, NY, USA), pp. 59–70, ACM, 2015.

[16] J. Ren, A. Rao, M. Lindorfer, A. Legout, and D. Choffnes, "Recon: Revealing and controlling pii leaks in mobile network traffic," in Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys '16, (New York, NY, USA), pp. 361–374, ACM, 2016.

[17] D. Noever "The Enron Corpus: Where the Email Bodies are Buried?", arXiv preprint arXiv:2001.10374, 2020.

[18] Z.R. Nokhbeh, K.S. Barber, "A study of web privacy policies across industries", Journal of Information Privacy and Security,vol.13(4), pp.169-85, 2017.

[19] M.D. Bader, S.J. Mooney, A.G. Rundle, "Protecting personally identifiable information when using online geographic tools for public health research", Am J Public Health, pp. 206-208, 2016.

[20] A. Alnemari, R.K. Raj, C.J. Romanowski, S. Mishra, "Protecting personally identifiable information (pii) in critical infrastructure data using differential privacy", In IEEE International Symposium on Technologies for Homeland Security (HST) , pp. 1-6, 2019.

[21] M.M. Onik, C.S. Kim, N.Y. Lee, J. Yang, "Personal Information Classification on Aggregated Android Application's Permissions", Applied Sciences, (19), pp. 39-97, 2019.

[22] A. Majeed, F. Ullah, S. Lee, "Vulnerability-and diversity-aware anonymization of personally identifiable information for improving user privacy and utility of publishing data", Sensors, vol.17(5), pp.1059, 2017.

[23] J. Venkatanathan, V. Kostakos, E. Karapanos, J. Gonçalves, "Online disclosure of personally identifiable information with strangers: Effects of public and private sharing, Interacting with Computers, vol. 26(6):614-26, 2014.

[24] W.B. Tesfay, J.M. Serna, and S. Pape, "Challenges in Detecting Privacy Revealing Information in Unstructured Text", In PrivOn@ ISWC, 2016.

[25] Y. Liu, T. Song, L. Liao, "TPII: tracking personally identifiable information via user behaviors in HTTP traffic", Frontiers of Computer Science, vol. 14(3):1-4, 2020.

[26] N. Vishwamitra, Y. Li, K. Wang, H. Hu, K. Caine, G.J. Ahn, "Towards pii-based multiparty access control for photo sharing in online social networks', In Proceedings of the 22nd ACM on Symposium on Access Control Models and Technologies, pp. 155-166, 2017.

[27] Y. Hu, J. B-Graber, B. Satinoff, "Interactive topic modeling", Mach Learn 95, pp.423–469, 2014.

[28] B. Krause, L. Lu, I. Murray, S. Renals, "Multiplicative LSTM for sequence modelling", arXiv preprint arXiv:1609.07959,2016.

[29] F. Nielsen, "Hierarchical Clustering. In: Introduction to HPC with MPI for Data Science", Undergraduate Topics in Computer Science. Springer, Cham, 2016.

Evaluation of Data Center Network Security based on Next-Generation Firewall

Andi Jehan Alhasan, Nico Surantha
Computer Science Department, BINUS Graduate Program-Master of Computer Science, Jakarta, Indonesia

***Abstract*—This study aims to create a network security system that can mitigate attacks carried out by internal users and reduce attacks from internal networks. Further, a network security system is expected to overcome the difficulty of mitigating attacks carried out by internal users. The goal of this research is to analyze the effectiveness of the Next-Generation Firewall implemented to improve network security. The method used in this research is the comparison method with a test of TCP SYN attack, UDP flood attack, ICMP smurf attack, and DHCP starvation attack on a company network. From the experiment results, it can be concluded that the Next-Generation Firewall has significantly better performance for protecting mitigating attacks carried out by internal users on a company network. It can increase the security of data communication networks against threats from the internal networks.**

***Keywords*—Network security; next-generation firewall; TCP SYN attack; UDP flood attack; ICMP smurf attack**

I. Introduction

Advances in IT technology today bring security concerns. So, it is crucial to secure network infrastructure [1]. Some widely used mechanisms to secure the network are firewall, Intrusion Detection System (IDS), and Intrusion Prevention System (IPS). A firewall is a software or hardware-based network security system that controls incoming and outgoing network traffic by analyzing packets and determining whether packets are allowed through the firewall or not based on the rules created. The firewall works on Transmission Control Protocol/Internet Protocol (TCP/IP) [2]. Other mechanisms to secure the network are Intrusion Detection System (IDS), which can detect unauthorized attack activity, and Intrusion Prevention System (IPS), which can perform actions to prevent intrusion or attack on the network [3]. The purpose of using Firewall, IPS, or IDS is to protect the internal network from external attacks network and to protect firewall core network from internal attacks.

In the existing network topology, coal companies currently still use traditional firewall devices to provide network security for the company's data center network infrastructure. Traditional firewalls that are currently implemented in the company have not been able to prevent attacks from attackers originating from the local area network, such as DHCP Starvation attacks, TCP Syn Attacks, UDP Flood Attacks and, ICMP Smurf Attacks, to the coal company's core network firewall because traditional firewalls are currently only able to block data/traffic that passes in the form of ip addresses, ports, and protocols and cannot block specific data or detect the contents of data packets that pass through the network [18] [19]. Traditional Firewall rules and policies are not able to block these attacks due to the limitations of features in general in traditional enterprise firewall devices or other traditional firewall devices. The design of the Next Generation Firewall implementation on the pfSense firewall device is expected to be able to prevent attacks originating from the local area network to the core network firewall. The advantage of Next Generation Firewall compared to traditional firewalls is that it is able to provide network security functions by implementing packet blocking and a deep packet inspection feature against a number of malicious packets passing through the network. The Next Generation Firewall will periodically protect against new types of attacks that have not been identified by traditional firewalls. In this study, the Next Generation Firewall was integrated with an existing firewall that functions only as a router that currently exists in the company's network infrastructure.

Based on previous research papers that have been reviewed, previous research only used firewall, Intrusion Prevention System (IPS) and Intrusion Detection System (IDS) devices to prevent SQL Injection attacks, OS Bash Injection Attacks, HTTP Traffic, DDoS attacks, Port Scanners, and Password Brute Force. There has been no research that focuses on preventing attacks from attackers on traditional firewalls, IPS and IDS such as DHCP Starvation attacks, TCP Syn Attacks, UDP Flood Attacks, and ICMP Smurf Attacks. So, our motivation in this research is to contribute by adding new methods of preventing or mitigating types of attacks, namely, DHCP Starvation attack, TCP Syn Attack, UDP Flood Attack, and ICMP Smurf Attack, from local area networks to firewall devices from the internal network.

In this study, a solution for preventing attacks from local area networks to the core network firewall infrastructure is the use of the Next Generation Firewall pfSense which is integrated with the existing firewall that functions as a router. In this study, it will be tested using Kali Linux (Nmap, Hping3) and Yersinia Tool as attackers and Next Generation Firewall pfSense along with its configuration as a prevention of these attacks.

II. Literature Review

Putra and Surantha previously have done research to design and implement vulnerability port Hypertext Transfer Protocol (HTTP), SQL Injection, and OS Bash Injection Attack on servers from internal network or inline mode with Damn Vulnerable Web App and Vega Vulnerability Scanner tools. The study integrated Cisco ISE Network Access Control, Cisco Identity Service Engine (ISE), and Intrusion Prevention System (IPS) on Cisco Firepower 8250 devices. In the study, the attack

test was carried out on the windows server and then SQL Injection attack and OS Bash Injection Attack were performed on the server. The test results obtained that the integration of cisco ISE as Network Access Control (NAC) and Cisco Firepower 8250 devices as IPS can prevent attacks to HTTP ports, SQL Injection, and OS Bash Injection Attack the server from internal attacks. Compared to using only NAC devices because Cisco ISE devices can not prevent such attacks [4].

Erlacher and Dressler researched the implementation of Signature-based Network Intrusion Detection Systems using the Snort tool and in the research, they used Novel Flow-based Network Intrusion Detection Systems-Flow Information Export-based Signature-based Intrusion Detection System (NIDS FIXIDS). In the study, Tool Vermont (Versatile Monitoring Toolkit) served as a network monitoring toolkit and was combined with snort to prevent and detect attacks of massive HTTP traffic or high-speed traffic. The attack was tested using the Cisco Trex Traffic generator tool by sending high-speed HTTP traffic on a network that has been running NIDPS and FIXIDS. The test results improved methods for preventing and detecting a massive traffic attack on the network [5].

Other research by Bul'ajoul, James, and Shaikh researched the implementation of a Network Intrusion detection prevention system (NIDPS) using the Snort tool as NIDPS with a new Novel NIDPS architecture method. In the study, the snort tool that functioned as NIDPS could not prevent and detect attacks optimally when it received a huge traffic attack or high-speed traffic. Testing was done from Wincap, Flooder Packet, and Transmission Control Protocol (TCP) replay tools as attackers transmit TCP and User Datagram Protocol (UDP) traffic attacks. Testing with Novel NIDPS method architecture and integrating with Cisco Layer 3 switch resulted in better attack prevention and detection methods despite massive traffic attacks or high-speed traffic to internal networks [6].

Other research, which was conducted by Kaur and Singh, examined Web-Based attacks or attacks on a website on a network using deep learning-based system methods as hybrid intrusion detection and signature generation to prevent and detect attacks on the web. The prevention and detection of such attacks using D-Sign Architecture consist of Misuse Detection Engine, Anomaly Detection Engine, and Signature Generation Engine. In the study, the attack was tested by experimenting with attacks to a web with various HTTP traffic attacks. The results of D-Sign Architecture Implementation combined with Deep Recurrent Neural Network-based anomaly detection can produce better prevention and detection in the event of a known attack or a new attack on HTTP [7].

Duppa and Surantha did review and comparison testing traditional Intrusion Prevention System and Next-Generation Intrusion Prevention System. In the study, testing Next-Generation IPS to protect the network from attacks that take advantage of traditional IPS weaknesses, namely exploitation to HTTP Port or layer 7. Testing was done using Kali Linux as an attacker from inline or internal network mode with SQL Injection attack and malicious site exploit. The device used as NGIPS is Cisco Fire Power. From the penetration testing results, it can be concluded that Next-Generation IPS using the Cisco Firepower device can prevent SQL Injection attacks and malicious site exploit better than traditional IPS [8].

Ring and Landes researched the implementation using two methods, namely Unsupervised Port Scan Detection (UPDS) and Supervised Port Scan Detection (SPDS) in the research, to prevention of port scanning TCP and UDP protocols in a traffic network, such as scanning ports that open on Switch, Router, and Firewall devices. The scanning port was done in the research with the Nmap tool combined with Open Stack and NetFlow. The research compared the port scanning test algorithm with several TFDS, TRQ-SYN, UPDS, SPDS, and Webster. The test results obtained the results that UPDS and SPDS algorithms on the network can reduce or prevent an attacker from being able to scan open ports on the network or the user client [9].

Three-Tier Novel Architecture for Intrusion Detection and Prevention System in Software Defined Network was for prevention of DDoS attacks on a network. The research compared and tested Novel Three-Tier architecture with Intrusion Detection and Prevention SDN-IoT method to prevent replay attack, Mima attack, forgery attack, DDoS attack, on SDN network. From the test results, from the test results, it can be concluded that using Novel Three-Tier architecture for Intrusion Detection Prevention System (IDPS) can be guided from attacks on the system rather than the old Intrusion Detection and Prevention SDN-IoT [2].

Research conducted by Rengaraju and Ramanan researched the implementation of Intrusion Prevention System on Software Defined Networking network to secure Software-Defined Clouds (SDC) network as a controller for Denial-of-Service (DoS) and ICMP Flooding attack prevention. Attack testing was carried out with the hping3 tool. From the test results obtained, SDC Controller that serves as Access List (ACL) and IPS can prevent Denial-of-Service (DoS) and ICMP Flooding attacks with Signature-based method [10].

Research conducted by Karim and Handa researched Intrusion Detection System, which is increasingly a key element of system security that is used to identify the malicious activities in a computer network or system. Hybrid computing is one of the latest and an emerging area in the Information and Technology (IT) sector, which has given a different dimension to the organizations. Performance and security aspects and the major issues have to be addressed in Hybrid Computing. This research will attempt to give an overall idea about Hybrid computing, Intrusion, types of Intrusion Detection Systems, and earlier works done on Intrusion Detection System [11].

A global intrusion detection system composed by autonomous Internet-distributed detection systems was proposed. In our approach, distributed detection elements cooperate by sending information about a potential threatening flow that traverses its Autonomous System (AS). Distributed Intrusion Detection Systems (DIDS) use Border Gateway Protocol (BGP) updating capabilities in order to spread intrusion warning messages across Internet routing domain so as to notify the SIEM of the attack target. When an anomalous in-transit traffic is detected, the AS integrated IDS gathers all attributes of the anomalous flow in the extended BGP Network Layer Reachability Information (NLRI) field and advertises it

towards the AS target of the intrusion. Then, the SIEM of the target AS can use such information set to manage related protection countermeasures [12] [13].

Other research by Choi and Allison is in the form of a review paper related to the methods used Intrusion Prevention System and Intrusion Detection System. IPDS performs attack prevention and detection using 2 algorithms, namely signature-based detection and anomaly detection. In the study, the authors emphasized the use of IDPS implementation in a small to Medium-size Enterprise in preventing and detecting attacks within computer networks [14].

This research was to analyze the effectiveness of the Next Generation Firewall that was implemented to secure IoT in smart house and company network. The method used in this research was the method of comparison with a test of DDoS attacks, phishing, and SQL Injection on both network, smart house network, and company network. From the results of experiment, it can be concluded that the Next Generation Firewall has significantly better performance for protecting smart house and company network and it can increase security of data communication networks against threats from the Internet [15].

Sri Lanka Institute of Information Technology Computing (Pvt) Ltd representatives presented research in 2016 on how to create a more secure network by integrating firewall capability and firewall technologies. The findings of the experiment prove that the suggested concept is capable of constructing a stable network. This study discussed how firewalls are used to shield infrastructure from outside intruders and how Virtual Private Networks (VPN) allow encrypted access to the corporate network over non secure public networks [16].

Other research by Kishan, Rami, and Lei researched the Traditional firewalls are incapable of coping with emerging threats such as targeted and data-focused attacks. This paper discusses a survey of the different types of current and next-generation firewalls, highlighting their potential functionalities. The different technologies implemented in Next-Generation Firewall (NGFW) for network security were highlighted. Additionally, the advantages of the next-generation firewalls were compared against the traditional firewalls. Also, in this paper, the primary network security goals, their recent emerging security threats, and their potential solutions to protect the network are discussed [17].

Most data centers still use traditional firewalls, Intrusion Prevention System (IPS), and Intrusion Detection System (IDS) to provide network security in data centers. Traditional firewalls, IPS, and IDS applied in this data center cannot detect attacks with different variants, such as TCP SYN attacks, UDP flood attacks, ICMP smurf attacks, and DHCP starvation attacks. By utilizing anomaly-based because traditional firewall, IPS, and IDS weaknesses only understand the identity of data passed based (IP address and port used) and unable to recognize deep packet inspection, unable to know what is in the data package of these attacks.

TABLE I. COMPARISON OF RESEARCH ON IPS, IDS AND FIREWALL

No	Reference	Research Topics	Method	Tools
1	Putra & Surantha, 2019	implementation Cisco ISE and IPS NAC to prevent SQL Injection attacks and OS Bash Injection Attack on servers.	Network Access Control and Intrusion Prevention System (IPS)	Cisco ISE and Cisco Firewall Firepower
2	Erlacher & Dressler, 2019	Implementation of Signature-based Network Intrusion Detection Systems to prevent HTTP Traffic attacks	Signature-based Network Intrusion Detection Systems	Snort, Tool Vermont, and Cisco Trex Traffic generator
3	Bul'ajoul, James, & Shaikh, 2019	Implementation of Network Intrusion detection prevention system (NIDPS) to prevent substantial traffic attacks or high-speed traffic on an internal network	Novel Network Intrusion Detection Systems architecture	Snort, tool Wincap, Flooder Packet, and TCP replay
4	Kaur & Singh, 2019	Research related to Web-Based Attack or attack on a website on a network	Arsitektur D-Sign, Deep Recurrent Neural Network-based anomaly detection	Open Web Application Security Project (OWASP)
5	Duppa & Surantha, 2019	Implementation of Next-Generation IPS Testing to prevent SQL Injection and Exploit Malicious Site attacks	Next Generation Intrusion Prevention System (NGIPS)	Cisco ISE and Cisco Firewall Firepower
6	Ring, Landes, & Hotho, 2018	Testing and prevention of port scanning to TCP and UDP protocols on traffic network for the prevention of port scanning on Switches, Routers, and Firewall	UPDS (Unsupervised Port Scan Detection) and SPDS (Supervised Port Scan Detection).	Nmap, Open Stack, and NetFlow
7	Ali & Yousaf, 2020	Intrusion Detection and Prevention System in Software Defined Network to prevent DDOS attacks on SDN network	Deep Learning Novel Three-Tier	Network simulator environment OMNeT++ 4.6
8	Choi & Allison, 2017	Review paper of attack prevention and detection using signature-based detection and anomaly detection algorithms in small and medium-sized enterprise network	Signature-based detection and anomaly detection of IDS, IPS	
9	Rengaraju & Ramanan, 2017	Implementation of Intrusion Prevention System in Software Defined Networking network for Denial-of-Service attack prevention.	Distributed Firewall with Intrusion Prevention System (IPS) for SDC	Software-Defined Clouds Controller
10	Soewito & Andhika, 2017	Analyze and implemented the effectiveness of the Next Generation Firewall to prevent DDoS attacks, phishing, and SQL Injection in bright house and the company network.	Next Generation Firewall	LOIC (Low Orbit Ion Canon) and NGFW Checkpoint
11	Tharaka, Silva, & Sharmila, 2016	Analyze firewall capacity and other firewall technologies such as packet filtering, network address translation, virtual private network, and proxy services.	Firewall	
12	Neupane, Rami, & Lei, 2018	A survey of the different types of current and next generation firewalls are discussed in details highlighting their potential functionalities.	Next Generation Firewall	Palo Alto Next Generation Firewall

Previous research on Literature Review or refer in the Table I used a firewall, IPS, and IDS devices to prevent SQL injection attacks, OS bash injection attacks, HTTP traffic, DDoS attacks, Port scanner, and passwords brute force only. No research focused on preventing attacks such as TCP SYN, UDP flood, ICMP smurf, and DHCP starvation attacks on firewall or router devices. Thus, the motivation of this research is to later contribute by adding prevention or mitigation with different variants types of attacks that currently often occur in network infrastructure in data centers that come from internal networks.

In this research, a solution was proposed to improve network security to prevent attacks on the company's infrastructure network and prevent attacks from internal networks to the company's core network infrastructure by using Next-Generation Firewall pfSense and Suricata tool.

III. Methodology

A. System Design

As summarized in Table II, the Next-Generation firewall used in this research was open source tool security software, namely pfSense with OS 2.5.1 series. Next-Generation firewall has been connected between the traditional firewall and core switches. The Next-Generation firewall pfSense was integrated with the traditional firewall with all segments in the network infrastructure. Pfsense was able to communicate with a traditional firewall as a router to perform the expected integration according to the objectives in this study. Switch access connected directly to the user's pc using layer 2 switch. Computers used as attackers were Asus type with Kali Linux OS.

Fig. 1 proposes a new topology using Next-Generation firewall pfSense connected to the network in inline mode. In this study, the device was integrated with the existing infrastructure. The integration done in this study was a physical and logical connection in which the traditional firewall as a router internet gateway and the Next-Generation firewall pfSense should be able to connect with existing infrastructure devices and then configure to integrate existing devices. Policy implementation and configuration will be carried out on the Next-Generation firewall pfSense to prevent TCP SYN attack, UDP flood attack, ICMP Smurf attack and DHCP Starvation Attack. To achieve the objectives in this study, a network security system that can reduce internal users who carry out attacks and reduce attacks from internal networks was created by using Next-Generation firewall pfSense.

In the existing network topology, the traditional firewall core network implemented as a router internet gateway, which runs Network Address Translation (NAT) service, Routing Default route, Domain Name System (DNS) Service, and service Dynamic Host Configuration Protocol (DHCP), aims to connect internal network clients to connect the internet safely and reliably. On the proposed added devices topology above, Next-Generation firewall devices implemented the Intrusion Prevention System (IPS) feature in inline mode to prevent attacks to the traditional firewall core network. On the access switch, VLAN and Port Security division are implemented to ensure that only clients' registered Media Access Control (MAC) address and VLAN in the switch can be connected to the network or the internet.

TABLE II. Existing and Proposed System Specifications

Tool	Vendor	OS Version	Function
RB1016-12G	Mikrotik	6.47.8	Traditional Firewall As Router Internet Gateway
PC with 2 LAN Card	PfSense	2.5.1	Next-Generation Firewall

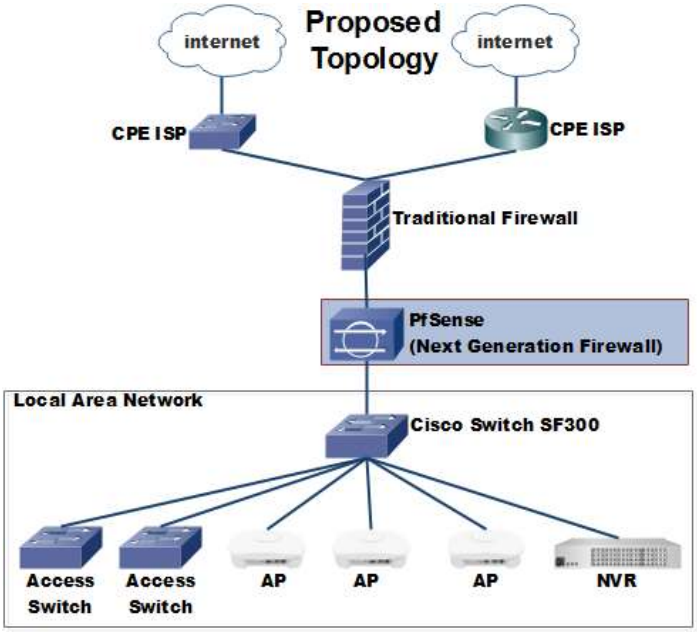


Fig. 1. Proposed Network Topology.

Table III presents existing configuration traditional firewall Mikrotik as router internet gateway, while Table IV and Fig. 2 present the configuration of the proposed Next-Generation Firewall Pfsense as intrusion prevention system for preventing TCP SYN attack, UDP flood attack, ICMP smurf attack, and DHCP starvation attack from an internal network.

TABLE III. Configuration of Traditional Firewall Mikrotik

Service	Configuration	Function
IP Address	ip address add address=192.168.10.1/24 network=192.168.1.0 inte rface=ether1	IP Gateway Local Area Network
DHCP Client	ip dhcp-client add interface=wlan1 use-peer-dns=yes use-peer-ntp=yes add-default-route=yes	Request Internet Connection to ISP
NAT	chain=srcnat action=masquerade out-interface=wlan1 log=no log-prefix=""	Translation from an internal network to external network/internet
Default Route	ip route add dst-address=0.0.0.0/0 gateway=192.168.1.1 distance=1	Routing packet to the internet from LAN
DHCP Server	Add address=192.168.10.0/24 gateway=192.168.10.1	Give out ip address to LAN

TABLE IV. CONFIGURATION OF PROPOSED NEXT-GENERATION FIREWALL PFSENSE FOR PREVENTING ATTACKS

Service	Configuration	Function
IP Address	Interface>LAN>IPv4 address 192.168.1.1, IPv4 subnet /24, description LAN	Default gateway LAN
	Interface>WAN>IPv4 address 192.168.10.2, IPv4 subnet /24, IPv4 gateway 192.168.10.1 description WAN	Connection to Traditional Firewall
Default Route	System>Routing>Gateways>add Interface WAN, address Family IPv4, Gateway 192.168.10.1	Connecting to Internet
NAT	Firewall>NAT>Outbond>Mode Automatic Outbound NAT	Translation from an internal network to external network/internet
DNS	System>General Setup>DNS Server 202.152.254.246	Translate domain names into IP Addresses
Suricata	Interface>Service>Suricata	Enable Mode IPS
Intrusion Prevention System	Interface>Edit>Enable Block Offenders	For the prevention of TCP SYN, UDP flood, ICMP smurf and DHCP starvation attack
	IPS Mode> Legacy Mode	
	Kill State > Enable	
	Which IP to Block > SRC	

Fig. 2. Enable Configuration IPS on pfSense using Suricata.

B. Implementation and Testing

In this research, implementation and testing were carried out to prove the solution given to overcome the existing problems. This implementation and testing were carried out using system design and infrastructure that has been integrated with the traditional firewall like a router internet gateway and Next-Generation firewall pfSense and Suricata as an Intrusion Prevention Systems. By configuring the traditional firewall and Next-Generation firewall pfSense, both systems can communicate and integrate to achieve the objectives of this research. Then, this test took place by trying to simulate an attack by connecting the user's laptop to the internal network. The attack operating system used the Kali Linux tool hping3 and additional tool Yersinia that acted as an attacker. In this study, the traditional firewall Mikrotik targeted the attack connected to the Next-Generation firewall pfSense. Then, the attacker would perform a TCP SYN attack, UDP flood attack, ICMP smurf attack, and DHCP starvation attack on the traditional firewall. Table V presents system specifications used for attack testing simulation and Fig. 3 is the attack testing topology used in this study.

In this test, the Next-Generation firewall pfsense was located in the middle of an inline configured network so that the traditional firewall could immediately decide on the package that has been checked. The package was analyzed by the Next-Generation firewall pfSense based on signatures or anomalies. If the package contains a crime or vulnerability, the Next-Generation firewall pfSense will immediately prevent it by blocking malicious packages. Then, the Next-Generation firewall pfSense can immediately prevent and quarantine the computer that is the source of the attack so that it no longer launched prolonged attacks on the network. In this test, a compliant user was an official network access condition with specific requirements. Attack testing employed several sample attacks, namely TCP SYN attack, UDP flood attack, ICMP smurf attack, and DHCP starvation attack, using tool Hping3 and Yersinia on Kali Linux OS.

TABLE V. SYSTEM SPECIFICATIONS USED FOR ATTACK TESTING SIMULATION

Perangkat	Vendor	OS Version	Function
RB951UI-2HND	Mikrotik	6.43.1	Traditional Firewall
Laptop with 2 LAN Card	Pfsense	2.5.1	Next-Generation Firewall
Notebook	Kali Linux	Kali Linux	Attacker

Fig. 3. Attack Testing Topology.

IV. Evaluation Result and Discussion

The Next-Generation firewall that was used in this study as Intrusion Prevention System was open-source pfSense and Suricata. Next-Generation firewall pfSense will connect traditional firewall Mikrotik device and cisco distribution switch. Integrating traditional firewall Mikrotik and Next-Generation firewall pfSense and Suricata as Intrusion Prevention System is expected to prevent an attack from an internal network to a traditional firewall core network.

Next-Generation firewall pfSense and Suricata will prevent attacks from internal networks to a traditional firewall core network as the purpose of this study expects. The following are the results of attack tests that have been done with the Next-Generation firewall pfSense and Suricata.

A. TCP Syn Attack

Based on the tests that have been done, users tried TCP SYN attacks using Kali Linux with the hping3 tool on the target traditional firewall Mikrotik. The first attacker performed a port scan that was available on target. The commonly used target is the 80/HTTP service, as shown in Fig. 4. Then, the attack used the command hping3 -c 20000 -d 120 -S -w 64 -p 80 --flood --rand-source 192.168.10.1, as shown in Fig. 5, which aims to send the targeted attackers a large number of TCP SYN packets. Using only the traditional firewall Mikrotik, these attack attempts succeeded by exhausting traditional firewall resources and impacting performance. As shown in Fig. 6, these attacks can be performed because the traditional firewall Mikrotik cannot detect the TCP SYN attack pattern. However, using the enhancements of the Next-Generation firewall pfSense in this study, TCP SYN attack attempts were detected and blocked by pfSense. PfSense detects attack patterns and then blocks or prevents them before getting to the traditional firewall Mikrotik so that network performance becomes regular and smooth. As shown in Fig. 7, TCP SYN attacks can be mitigated by pfSense.

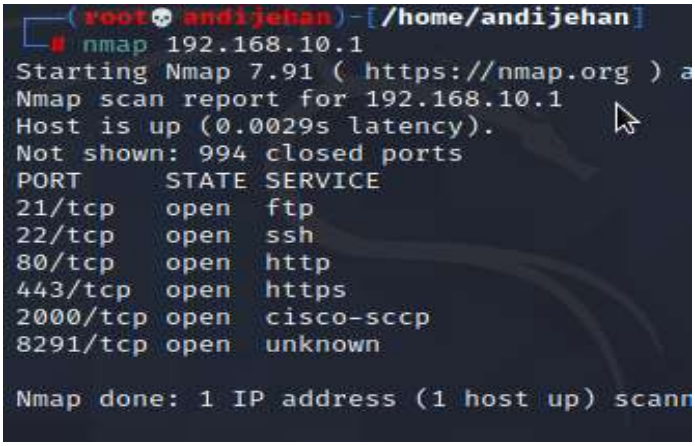


Fig. 4. Scanning Available Ports on a Target.

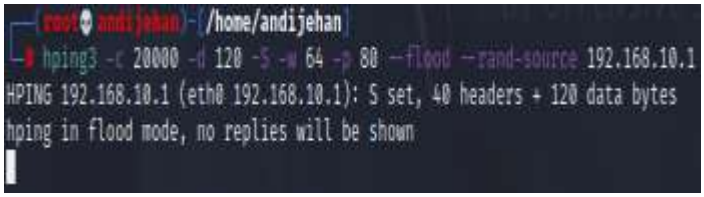


Fig. 5. Attacker Send Large Number TCP/SYN Packets.

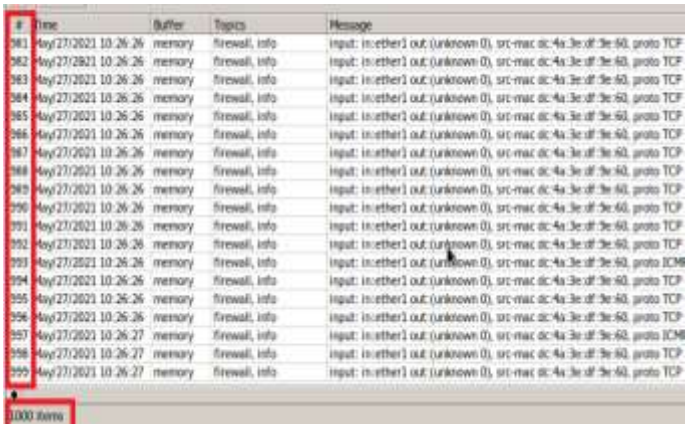


Fig. 6. TCP SYN Sent from Random Source Addresses.

Fig. 7. Result Log Blocked TCP SYN Attack by Pfsense.

B. UDP Flood Attack

Based on the results of the second test, the UDP flood attack was carried out on a traditional firewall Mikrotik as a target by still using the hping3 tool, starting to attack the UDP 53 (DNS) protocol port with hping3, as seen in Fig. 8 using the CLI --flood --rand-source --UDP -p 53 192.168.10.1 command. This attack aimed to create and send many UDP datagrams from the pampered IP to the targeted traditional firewall Mikrotik. When the firewall receives this type of traffic, it cannot process each request and consume its bandwidth by sending ICMP "unreachable destination" packets. Using only a traditional firewall Mikrotik, these attack attempts succeeded by exhausting firewall resources and impacting performance. As shown in Fig. 9, using an additional device that was the Next-Generation firewall Pfsense in this study, UDP flood attack attempts were detected and blocked by pfSense. PfSense can detect attack patterns and then block or prevent them from getting to the traditional firewall Mikrotik so that network performance becomes regular and smooth. As shown in Fig. 10, the TCP SYN attack can be mitigated by pfSense.

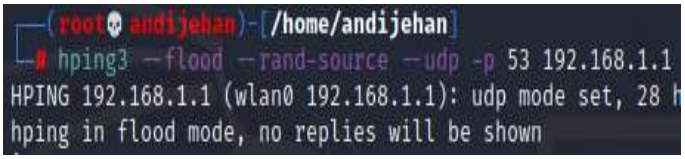


Fig. 8. Attacker UDP Protocol Port 53 (DNS).

Name	Usage
Cpu0	100.0
ethernet	0.0
management	1.5
networking	0.0
unclassified	0.5
winbox	0.5

Fig. 9. Exhausting the Resources CPU of the Traditional Firewall.

Fig. 10. Result Blocked UDP Flood Attack by Pfsense.

C. ICMP Smurf Attack

The third test was an ICMP smurf attack. This type of attack uses a large number of Internet Control Message Protocol (ICMP) ping firewall that targeted internet broadcast addresses, e.g., 192.168.1.255, using the command hping3 --Icmp --flood --rand-source -c 20000 --spoof 192.168.10.1 192.168.10.255. as seen in Fig. 11. This attack is aimed at all replies sent to the victim instead of the IP used for pinging. Using only a traditional firewall Mikrotik, this attack exhausts firewall resources and impacts performance, as shown in Fig. 12, because the traditional firewall cannot detect the pattern of ICMP smurf attack. However, using an enhancement that is the Next-Generation firewall Pfsense in this test, ICMP smurf attack attempt was detected and blocked by pfSense. PfSense can detect attack patterns and then block or prevent them from entering traditional firewalls so that network performance becomes regular and smooth. As shown in Fig. 13, the ICMP smurf attack can be mitigated by pfSense.

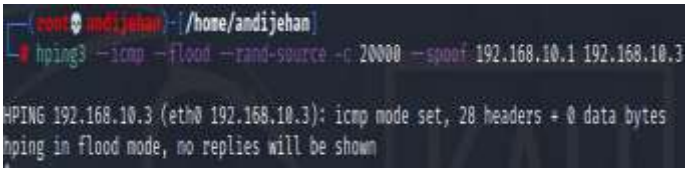


Fig. 11. Attacker Sends Large Number ICMP Packets.

Name	Usage
Cpu0	100.0
ethernet	0.0
management	1.5
networking	0.0
profiling	0.0
unclassified	0.5
winbox	0.0

Fig. 12. Exhausting the Resources CPU of the Traditional Firewall.

Fig. 13. Result Blocked ICMP Smurf Attack by Pfsense.

D. DHCP Starvation Attack

The last attack test was a DHCP hunger attack performed on a traditional firewall Mikrotik as a target using the Yersinia tool. As seen in Fig. 14, using a Mikrotik firewall, this attack can be detected and recognized so that DHCP hunger attacks do not make the traditional firewall Mikrotik down. Similarly, by using the additional Next-Generation firewall in this test, DHCP hunger strike attempts were detected and blocked by pfSense. PfSense can detect attack patterns and then block or prevent them from getting to the traditional firewall Mikrotik. As seen in Fig. 15, starvation attacks can be mitigated by traditional firewall Mikrotik and pfSense.

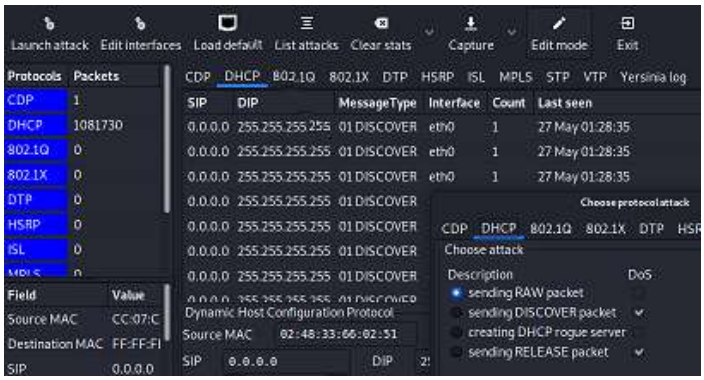


Fig. 14. DHCP Starvation Attack with Tool Yersinia.

Fig. 15. Result of Blocked DHCP Starvation Attack by Pfsense.

Table VI summarizes the results based on the completed tests in this study. It shows significantly different results from using a traditional firewall only or integrating a traditional firewall with a Next-Generation firewall pfSense. The expected test results in this study can be achieved using the proposed solution. The proposed solution demonstrated that the Next-Generation firewall pfSense can prevent attacks from internal users and can reduce attacks from internal networks based on the test scenarios that have been done. With additional devices, Next-Generation firewall pfSense can improve network security compared to traditional firewall Mikrotik only.

TABLE VI. RESULT COMPARISON

No	Type Intrusion	Target	Result	
			Firewall	NGFW Pfsense (Proposed Solution)
1	TCP Syn Attack	Vulnerable Firewall	Allowed	Blocked
2	UDP Flood Attack	Vulnerable Firewall	Allowed	Blocked
3	ICMP Smurf Attack	Vulnerable Firewall	Allowed	Blocked
4	DHCP Starvation Attack	Vulnerable Firewall	Blocked	Blocked

With the results of the tests that have been done in this research, the proposed solution is to integrate traditional firewall Mikrotik as a router with Next-Generation firewall pfSense that can mitigate against internal users who perform attacks from internal networks. Thus, network security with firewall system integration with a Next-Generation firewall can be increased compared to traditional firewall Mikrotik only.

V. CONCLUSION

Based on the test results conducted in this study, the proposed solution demonstrates that the Next-Generation firewall pfSense can prevent attacks from internal users and can reduce attacks from internal networks based on the test scenarios that have been done. With additional devices, the Next-Generation firewall pfSense can improve network security compared to traditional firewalls only. However, this study still has many limitations, especially on the type of attack tested. In this study, attack testing only targeted network devices with traditional firewall from internal networks only. It is recommended for further research to conduct attack testing from the public internet to the internal network in order to improve network security better.

ACKNOWLEDGMENT

The publication of this research is fully supported by Bina Nusantara University.

REFERENCES

[1] Abubakar, R., Aldegheishem, A., Majeed, M. F., Mehmood, A., Maryam, H., Alrajeh, N. A., ... & Jawad, M. (2020). An Effective Mechanism to Mitigate Real-Time DDoS Attack. IEEE Access, 8, 126215-126227.

[2] Ali, A., & Yousaf, M. M. (2020). Novel Three-Tier Intrusion Detection and Prevention System in Software Defined Network. IEEE Access, 8, 109662-109676.

[3] Alzahrani, S., & Hong, L. (2018). Generation of DDoS attack dataset for effective ids development and evaluation. Journal of Information Security, 9(4), 225-241.

[4] Putra, A. S., & Surantha, N. (2019). Internal Threat Defense using Network Access Control and Intrusion Prevention System. International Journal of Advanced Computer Science and Applications, Vol. 10, No. 9, 2019.

[5] Erlacher, F., & Dressler, F. (2020). On High-Speed Flow-based Intrusion Detection using Snort-compatible Signatures. IEEE Transactions on Dependable and Secure Computing.

[6] Bul'ajoul, W., James, A., & Shaikh, S. (2019). A New Architecture For Network Intrusion Detection And Prevention. IEEE Access, 7, 18558-18573.

[7] Kaur, S., & Singh, M. (2019). Hybrid intrusion detection and signature generation using Deep Recurrent Neural Networks. Neural Computing and Applications, 1-19.

[8] Duppa, G., & Surantha, N. (2019). Evaluation of network security based on next-generation intrusion prevention system. Telkomnika, 17(1), 39-48.

[9] Ring, M., Landes, D., & Hotho, A. (2018). Detection of slow port scans in flow-based network traffic. PloS one, 13(9), e0204507.

[10] Rengaraju, P., Ramanan, V. R., & Lung, C. H. (2017, August). Detection and prevention of DoS attacks in Software-Defined Cloud networks. In 2017 IEEE Conference on Dependable and Secure Computing (pp. 217-223). IEEE.

[11] Karim, H. A. R. A., Handa, S. S., & Murthy, M. R. (2017). A Methodical Approach to Implement Intrusion Detection System in Hybrid Network. International Journal of Engineering Science, 4817.

[12] Silva, R. S., & de Moraes, L. F. (2019). A cooperative approach with improved performance for global intrusion detection systems for internet service providers. Annals of Telecommunications, 74(3-4), 167-173.

[13] Silva, R. S., & Macedo, E. L. (2017, October). A cooperative approach for a global intrusion detection system for internet service providers. In 2017 1st cybersecurity in networking conference (CSNet) (pp. 1-8). IEEE.

[14] Choi, Y. B., & Allison, G. D. (2017). Intrusion Prevention And Detection In Small To Medium-Sized Enterprises. In SAIS 2017 Proceedings.

[15] Soewito, B., & Andhika, C. E. (2019, August). Next-generation firewall for improving security in company and IoT network. In 2019 International Seminar on Intelligent Technology and Its Applications (ISITIA) (pp. 205-209). IEEE.

[16] S.C. Tharaka, R.L.C. Silva, S. Sharmila, S.U.I. Silva, K.L.D.N. Liyanage, A.A.T.K.K. Amarasinghe, D. Dhammearatchi. (2016, April). High Security Firewall: Prevent Unauthorized Access Using Firewall Technologies. International Journal of Scientific and Research Publications, Volume 6, Issue 4, pg. 504-508,April 2016, ISSN 2250-3153.

[17] Neupane, K., Rami, H., Lei, C., (2018). Next Generation Firewall for Network Security: A Survey. IEEE Access, 978-15386-6133-18.

[18] Barker, Keith., Scott Morris dan Kevin Wallace, CCNA Security 640-554, 2012.

[19] P. Oppenheimer and T.-D. N. Design, "Cisco Press," ISBN, vol. 1, pp. 57069-57870, 2011.

Analogy of the Application of Clustering and K-Means Techniques for the Approximation of Values of Human Development Indicators

José Luis Morales Rocha, Mario Aurelio Coyla Zela, Nakaday Irazema Vargas Torres, Genciana Serruto Medina
Gestión Pública y Desarrollo Social
Universidad Nacional de Moquegua, Moquegua, Perú

***Abstract*—The objective of this study was to apply Clustering and K-Means' techniques to classify the departments of Peru according to their Human Development Index. In this article, the elbow method was used to determine the optimal number of clusters, applying the classification algorithms to group the departments of Peru according to their similarities, in addition to the Principal Component Analysis (PCA) technique for a better display of clusters. After applying the unsupervised algorithms, the results were more relevant in clusters 2 and 4 according to their HDI, made up of the departments of Arequipa, the Constitutional Province of Callao, Ica, Lima, Moquegua and Tacna, where the most notable is the life expectancy at birth, the population with full secondary education, the number of years of education, the average per capita income, and the state's density index. The results obtained by the K-Means algorithm show more cohesive results than the Clustering algorithm.**

Keywords*—Clustering; K-Means; elbow method; cohesion; separation; human development index

I. Introduction

The main reason for developing this research is to establish indicators of similarities and identify which departments have a lower or higher level of HDI for the characteristics analyzed and manage better levels of life expectancy, access to education, income level and index of density of the State, in the different areas that allow an adequate formulation of public policies and prioritize the social agenda that allows better opportunities and degree of progress and equality of citizens.

The research proposes the application of unsupervised Machine Learning algorithms (K-Means and Clustering) to observe the formation of clusters, with their respective indicators, grouping the departments of Peru into four clusters, according to the similarities between them, to measure human development through life expectancy, access to education and income level.

In this research, unsupervised learning algorithms were proposed to group the departments into clusters, according to optimization criteria; being one of the most used the K-Means; this algorithm ranks the indicators into clusters. In [1] K-Means is a partition grouping technique; the data objects are divided into groups that do not overlap. In [2] The clusters allow interaction in external networks in which information flows and facilitates its transfer. For [3] clustering techniques meta-learning tools are useful to analyze the knowledge produced by modern applications. The elbow method is used to determine the optimal number of clusters and a suitable observation, fixing the distances between each cluster.

The most relevant departments according to their HDI are found in cluster 2 and cluster 1 are perceived with the lowest HDI values, so the State must provide public policies focused on the populations of the departments in cluster 1.

The use of the K-Means and Clustering algorithms require the classification of groups with similar characteristics, according to the Human Development Indicators (HDI) with a high incidence due to altitude and State Density Index; therefore, quality information for decision-making in the design of public policies to improve HDI by departments and regions is provided.

The structure of the research article presents the state of the art according to the study variables, theoretical background emphasizing classification techniques, determined the results of the clusters, the discussion and conclusions of the research.

II. Related Work

To achieve high precision in terms of time and space, in [3] considers K-Means to be the best option for large and categorical data. It concludes that the K-Means genetic algorithm is faster than evolutionary algorithms. In [4] two clustering methods: K-Means and hierarchical clumping in air pollution studies were reviewed, with the aim of providing a review of clustering applications, specifically by using hierarchical clustering and k- mean. It was stated that each grouping technique has its own advantages and disadvantages and there is no a "best" method.

According to [5] the performance of classification algorithms is influenced by certain characteristics of the data sets on which they are modeled, such as imbalance in class distribution, class overlap, and lack of density. At the same time, the circumstances of class overlap and lack of density of the minority class in unbalanced data sets are observed.

As [6] artificial intelligence in medicine shows that ultrasonic imaging technologies have a true diagnosis; Two types of neural network algorithms have been proposed in three categories: USCT images of healthy, fractured and osteoporotic bones. Initially, a Convolutional Neural Network classifier system is presented and then an evolutionary neural network with the AmeobaNet model for the USCT images

classification. In [7] emotion recognition through an artificial neural network that detects spoken expressions, proposing a regularized Bayesian artificial neural network model that recognizes emotions through speech. The Berlin database with 1470 samples of emotions: 500 angry, 300 happy, 350 neutral and 320 sad. The performance of the methodology is compared with other avant-garde ones used for the same purpose, the proposed methodology achieved 95% precision in the recognition of emotions, being one of the highest compared to other methodologies used.

In [8] Hybrid approaches to data classification and optimization algorithm increase the precision of data classification. The study performs Moth Flame (MFO) and Fuzzy Min Max Neural Network (FMMNN) optimization applications to classify medical data. In terms of classification, the experiment achieved 97.74% accuracy for liver disorders and 86.95% accuracy for the diabetes data set that is related to the achievement of good human health.

As [9] states, data analysis is used as a tool in different fields, clustering plays an important role in the composition of the data analysis, thus dealing with the segmentation of the data structure in an unknown segment; using the K-Means algorithm. This article explains the applications of clustering methods and the objectives of clustering with big data. It also introduces the clustering technique for identifying data patterns by performing sample data analysis.

In [10] the research examines the CatBoost ranking algorithm on loan approval and staff promotion. This algorithm outperforms other implemented classifiers. Two types of analysis were carried out, in the first one the amount, the type, the income of the applicant and the purpose of the loan that help to predict the approvals of the loan were considered, in the second case the division, the schooling abroad, the geopolitical zones, qualification and working years, which had a high impact on the promotion of personnel. Based on the performance of CatBoost, the algorithm is interesting for a better prediction of loan approvals and staff promotion.

In [11] the K-Nearest Neighbor algorithm is used in multidimensional and outlier data due to its precision. A hybrid K-Nearest Neighbor approach with optimized particle scoring to improve K-Nearest neighbor performance, which is implemented in two stages: it first resolves multidimensional data by selecting the features with the swarm optimization algorithm and the second resolves the presence of outlier's values with the results of stage 1 and applying a new K-Nearest Neighbor technique scored.

In [12] computer diagnosis of tumors is important, as their segmentation is difficult to diagnose. The Fuzzy K Means fast clustering algorithm based on super pixels was used. These images bring a multi-scale morphological gradient reconstruction operation that allows getting segmentation precision. The results reveal that this approach is fast and accurate compared to segmentation algorithms; which provide a high precision of 99.58% and an improved RFN value of 8.34% compared to other methods analyzed. In [13] the logistic regression, K-NN applied to the data set in breast cancer, was found to determine the well based prediction of the data set. Also, with logistic regression an accuracy of 91% was achieved, and the detection was early and accurate. Likewise, it is seen that to reduce and classify heart disease, the support vector machine (SVM) has been adopted, the closest K-NN neighbors and the linear discriminant analysis. It has been shown that the vector machine turned out to be a better classifier with an accuracy of 80.4%.

In [14] the paper uses fine-tuning transfer learning on RNA-Seq gene expression data, classifying 5 types of cancer that affect women. The data comes from the genomic data commons (GDC) portal, with 2166 samples, along with 19,947 common genes. Spearman's correlation was used to narrow down the number of genes, eliminating those that are highly correlated. Gene expression is filtered by selecting values greater than 0.25 in the samples. In the gotten profile, the samples are transformed into 2D images as data, adapting to the convolutional layer of the CNN architecture. We fit four previously trained models on the RNA-Seq gene expression data, namely ResNet50, DenseNet, Xception and VGG16. The Xception architecture shows the highest and most accurate performance 98.6%, recovery 97.8%, and F1 score of 98% in a five-time cross-validation test and training approach.

According to [15] the study proposes a model based on machine learning to predict new infections expected by COVID 19. The model is tested in Egypt and in the 10 highly rated countries in September 2020. The proposed model is implemented based on algorithms supervised machine learning regression. Then compared with one of the more accurate prediction models The Bayesian crest, and the results show the power of the model compared to its counterpart in all the countries studied.

III. Theoretical Background

The rapid development of data collection techniques and new storage technologies [11] have allowed organizations to retain a large amount of data. With the help of machine learning algorithms, the quality of decision-making can be supported thus human error can be avoided. Classification [16] are supervised techniques that categorize unknown data into a specific class or group. In classification, the classes are known in advance.

A. K-Means

In [17] the K-Means technique is a clustering algorithm, machine learning technique. In [18] K-Means is a partition clustering technique; data objects are divided into groups that do not overlap.

The K-Means algorithm [3] groups clusters iteratively. Calculate the distance means, using an initial centroid, with each class that is represented by the centroid, using the distance as a metric and giving the k classes in the data set. In the K-Means algorithm, the mean value of the elements within the group is represented in the center of each group. The K-Means algorithm [19] groups the data into groups, defining a fixed number of groups, assigning data iteratively to the groups formed by adjusting the centers in each group.

The K-Means technique [18] learns the characteristics of a data set and forms partitions with them, these partitions are called clusters, which represent data with similar features. For

numerical data, each group is represented by a centroid, which is the mean of the elements in the group. For categorical variable data, it corresponds to the object that occurs most frequently, which is used as the group prototype.

K-Means uses the squared Euclidean distance as a measure of similarity for cluster membership:

$$d_{sq} = \sum_{i=1}^{D}(x_i - y_i)^2 \quad (1)$$

In (1) x and y are points in a D-dimensional space. The number of clusters k is determined by minimizing the sum of squared errors (SSE), which is given by the sum of the squared error in each data pair and its closest centroid. It is given by (2).

$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\left\|x_j - c_j\right\|^2 \quad (2)$$

where c_j is the centroid of the j-th group, and $w_{ij} = 1$ if the data pair x_i is in group *j* and w_i, j = 0, if x_i is not in group *j*.

The K-Means algorithm [20] randomly selects k data points from an original data set to later add them as the center of the initial clustering. First, each piece of data is considered a data point. Then, the Euclidean distance algorithm is used to determine the distance between the data points and the cluster center, the data set is preliminarily clustered according to the distance. Finally, the average distance of the data in each group is calculated, and the center of the group is adjusted, and the final result of the grouping is obtained through multiple iterations.

B. Clustering

Clustering techniques [21] are used in different areas of research, such as data classification, taxonomy, document retrieval, image segmentation and pattern classification. The Clustering algorithm [18] is the technique of grouping elements using a similarity measure. The grouping can be hierarchical or partitioned, exclusive, overlapping or fuzzy, and complete or partial. The Clustering algorithm [3] presents as a result the reduction of the dimensionality of a data set. The goal of a clustering algorithm is to identify the various groups within a data set.

The clustering technique [22] is a method to group data into classes with identical characteristics in which the similarity between classes is maximized or minimized. Grouping is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes.

C. Types of Clustering

Clustering is divided into two types:

1) Hard clustering: each data point is or is not part of a cluster. It means that each element is grouped into one of the k groups.

2) Soft clustering: A probability is assigned to the data point to be in certain clusters instead of placing each data point in a separate cluster, a probability of being in k groups is assigned to each element.

D. Clustering Methodologies

Because the Clustering technique is subjective. Cluster analysis is not an automated activity, but an iterative information discovery process or a multi-objective collaborative optimization that involves trial and error.

E. Validation of the Classification Algorithms

As the goal of clustering is to group similar objects in the same cluster and different objects to be placed in different clusters, [23] internal validation metrics are usually based on two criteria:

1) Cohesion: The element of each cluster must be as close as possible to the other elements of the same cluster.

The Sum of Squared Within (SSW), internal measure to evaluate the Cohesion of the clusters the grouping algorithm generated is:

$$SSW = \sum_{i=1}^{k}\sum_{x\in C_i} dist^2(m_i, x) \quad (3)$$

where *k* is the number of clusters, *x* a point of cluster c_i and m_i the centroid of cluster c_i.

2) Separation: Clusters must be widely separated from each other. There are different approaches to measure this distance among cluster: distance between the closest member, distance between the most distant members, or the distance among centroids.

The Sum of Squared Between (SSB), a measure of separation used to evaluate the inter-cluster distance is given by:

$$SSB = \sum_{j=1}^{k} n_j\, dist^2\left(c_j, \bar{x}\right) \quad (4)$$

where k is the number of clusters, n_j is the number of elements in cluster *j*, c_j is the centroid of cluster *j*, and x is the mean of the data set.

IV. Results

This section shows the results obtained from the application of Unsupervised Machine Learning algorithms (K-Means and Clustering).

A. Indicators used

The following indicators were used in the application of Machine Learning techniques for the classification of the Human Development Index.

- Human development Index.
- Life expectancy at birth.
- Population with full secondary education (18 years).
- Years of education (Population aged 25 and over).
- Family income per capita.
- Altitude.
- State Density Index.

B. The Elbow Method

The criterion used to establish the number of clusters to be used was determined by the elbow method.

The elbow method uses the mean distance of the observations to their centroid. The larger the number of clusters k, the intra-cluster variance decreases more. The smaller the intra-cluster distance the better it is, since it means that the clusters are more compact. The elbow method looks for the value k that satisfies that an increase in k does not substantially improve the mean intra-cluster distance.

According to Fig. 1, 4 clusters were established for the classification of the Human Development Index.

C. K-Means

To graphically illustrate the formation of the clusters and because there are seven indicators and it is not possible to make a graph that represents all these characteristics, a technique called Principal Component Analysis (PCA) was used, which reduces the quantity of variables to be analyzed, in this case to be visualized, creating a smaller quantity of new variables that best represents the original variables.

In Fig. 2, the graph of the HDI classification is shown, by means of the two main components, coloring it according to the cluster to which each department belongs according to its HDI.

In this graph, it is observed that each of the departments are well defined according to their HDI, in components 1 and 2, each of the departments is represented with the points and with the colors the cluster to which they belong.

The clusters are organized according to the following colors:

Cluster 1: Blue

Cluster 2. Green

Cluster 3: Red

Cluster 4: Yellow

According to Fig. 3, it is observed that there is a significant difference between the four clusters, the number 2 presents a higher HDI, followed by cluster 4, meanwhile cluster 3 and cluster 1 present a low HDI. In addition, it can be seen that cluster 3 presents greater dispersion and cluster 4 less dispersion.

Table I shows that the most relevant departments according to their HDI are found in cluster 2, made up of the departments of Arequipa, the constitutional province of Callao, Ica, Lima, Moquegua and Tacna. The positions in favor of these departments are found in almost all their dimensions, making life expectancy at birth more noticeable, in the population with full secondary education, years of education, average per capita income and the state's density index.

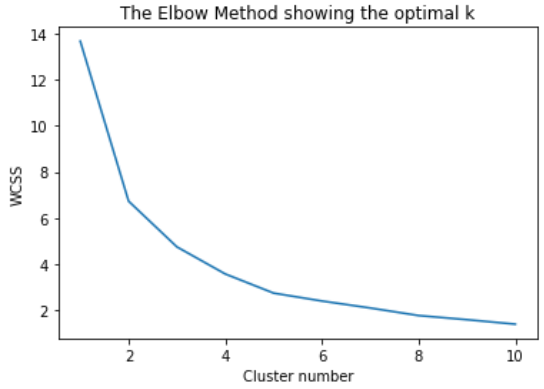


Fig. 1. The Elbow Method.

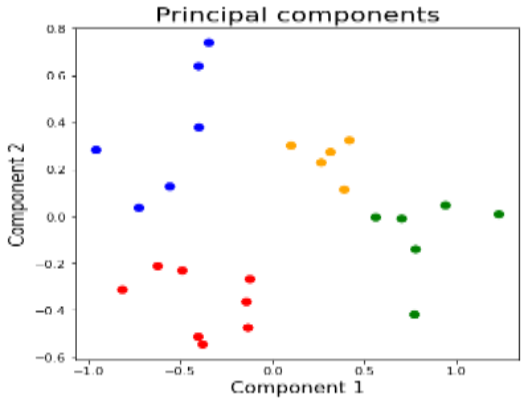


Fig. 2. Principal Components – Kmeans.

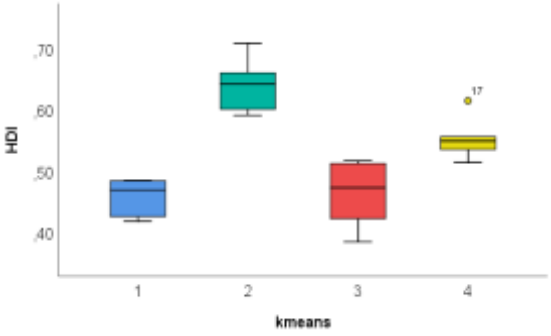


Fig. 3. Box Plot – K-Means.

TABLE I. CLASSIFICATION USING K-MEANS

Cluster	HDI	Life expectancy at birth	Population with full secondary education	Years of education	Family income per capita	Altitude	State Density Index
1	0.457	71.77	50.12	7	728.96	1359.00	0.66
2	0.639	76.85	74.11	10	1184.19	843.50	0.79
3	0.463	73.08	66.37	8	633.27	3383.75	0.72
4	0.552	75.74	64.92	8	937.99	74.40	0.74

In the departments of cluster 1, the lowest HDI values are noted, made up of the departments of Amazonas, Cajamarca, Huánuco, Loreto, San Martin and Ucayali. Although life expectancy at birth is relatively high (approximately 72 years), the figures for the population with full secondary education are approximately 50%, 7 years of education on average, it is appreciated that there is an average per capita family income of S /. 728.96 and a state density index of 0.66.

Cluster 4 is made up of the departments of La Libertad, Lambayeque, Madre de Dios, Piura and Tumbes; cluster 3 made up of the departments of Ancash, Apurímac, Ayacucho, Cusco, Huancavelica, Junín, Pasco and Puno.

D. Clustering

In Fig. 4, the dendrogram for the classification of the departments of Peru according to the HDI is shown.

In Fig. 5, the graph of the classification of the departments of Peru according to their HDI is shown, by means of the Clustering algorithm, through the two main components, coloring it according to the cluster each department belongs based on its HDI.

In this graph it is observed that each of the departments are also well defined according to its HDI.

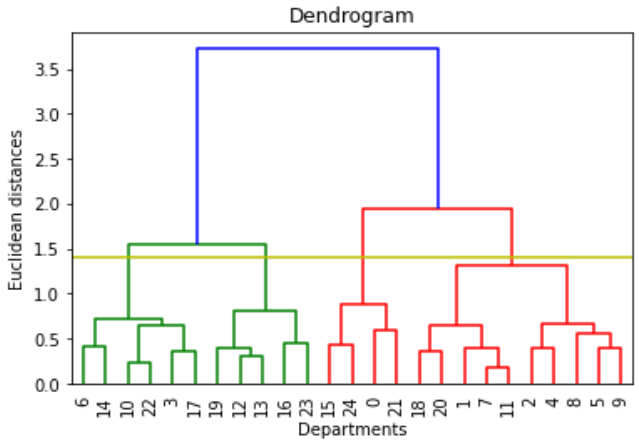


Fig. 4. Dendrogram – Clustering.

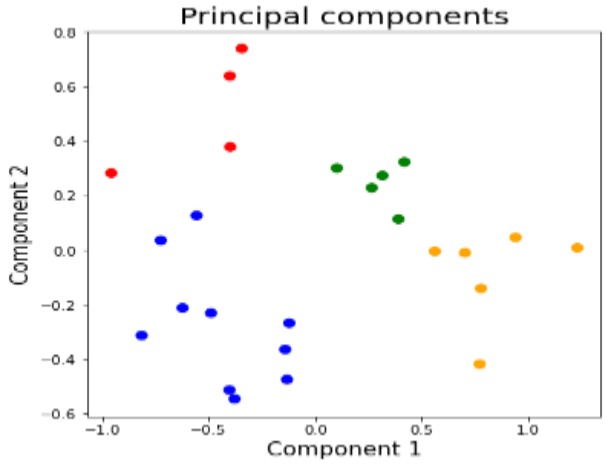


Fig. 5. Principal Components – Clustering.

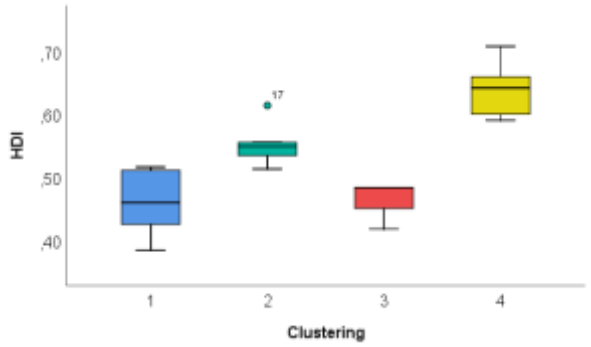


Fig. 6. Box Plot – Clustering.

According to Fig. 6, in the box-and-whisker plot of the Clustering algorithm, it is also observed that there is a significant difference among the four clusters, cluster 4 has a higher HDI, followed by cluster 2, cluster 3 and cluster 1 have a low HDI. In addition, it can be seen that cluster 1 presents greater dispersion and cluster 2 lower dispersion.

Table II displays that the most relevant departments according to their HDI are found in cluster 4, made up of the departments of Arequipa, the constitutional province of Callao, Ica, Lima, Moquegua and Tacna. The positions in favor of these departments are also given in almost all their dimensions, highlighting the life expectancy at birth, the population with full secondary education, the years of education, the average per capita income and the density index of the state.

In cluster 1 departments, made up of Ancash, Apurímac, Ayacucho, Cajamarca, Cusco, Huancavelica, Huánuco, Junín, Pasco and Puno, the lowest HDI values are perceived. Although life expectancy at birth is relatively high (approximately 73 years) and the population with full secondary education (64%), above 7 years of education on average are shown, an average per capita family income of S /. 635.10 and a state density index of 0.71.

The departments of La Libertad, Lambayeque, Madre de Dios, Piura and Tumbes compound cluster 2, and cluster 3 the departments of Amazonas, Loreto, San Martin and Ucayali.

E. Internal Validation Metrics

1) Cohesion: internal measure to evaluate the Cohesion of the clusters of the clustering algorithms:

$SSW_{Clustering} = 1.90$

$SSW_{K\text{-}Means} = 1.74$

The Sum of Squared Within (SSW) of the K-Means algorithm (1.74) shows more cohesive clusters than the Clustering algorithm.

2) Separation: measure of separation used to evaluate the inter-cluster distance.

$SSB_{Clustering} = 4.9856$

$SSB_{K\text{-}Means} = 4.9859$

The Sum of Squared Between (SSB) of the K-Means and Clustering algorithms show similar inter-cluster distances.

TABLE II. CLASSIFICATION USING CLUSTERING

Cluster	HDI	Life expectancy at birth	Population with full secondary education	Years of education	Family income per capita	Altitude	State Density Index
1	0.459	73.05	64.00	7	635.10	3172.20	0.71
2	0.553	75.74	64.92	8	937.99	74.40	0.74
3	0.467	71.19	47.91	8	772.25	875.50	0.66
4	0.640	76.85	74.11	10	1184.19	843.50	0.79

V. DISCUSSION

According to the objective, to apply Clustering and K-Means techniques to classify the departments of Peru according to their Human Development Index, the results exhibit in Table I show that the departments with greater relevance according to their HDI are in cluster 2, positions in favor of these departments arise in almost all their dimensions, making life expectancy at birth more noticeable, in the population with full secondary education, the years of education, the average per capita income and the density index of the state, these results were achieved using the K-Means technique. According to the Clustering technique, Table II shows the results obtained with the most relevant departments according to their HDI in cluster 4, positions in favor of these departments are also given in almost all their dimensions, making the hope of life at birth, population with completed high school, years of education, average per capita income, and state density index more relevant. The results obtained by the K-Means algorithm show more cohesive results than the Clustering algorithm.

Results that when compared with what was found by [3], who determined that the K-Means algorithm shows better results for the classification of big data. The author in [4] claims that K-Means and hierarchical clumping techniques have their own advantages and drawbacks and there is no "best" method. These results can affirm that the K-Means algorithm shows significant results regarding to the Clustering classification algorithms, especially in the cohesion measures.

On the other hand [24] K-Means is a classic prototype-based clustering technique that attempts to group data into K groups specified by the user. In [22], [25] Clustering is a method to group data into classes with identical characteristics, in which intraclass similarity is maximized or minimized.

VI. CONCLUSION

According to the K-Means algorithm, it identifies cluster 2 with the highest HDI because it groups the departments in the Coastal Region with higher population density on average, higher per capita income, strategic geographic location in metropolitan areas and zones of industrial, commercial, agricultural and mining activity that contribute to development, greater employment, health and education, contribution by mining canon and increase in government investment.

The K-Means algorithm accurately determines the grouping by departments in cluster 3, which shows a lower level of HDI doing its classification by similar characteristics of geographical location, belonging to the Sierra Region, which have a relationship and incidence due to higher altitude and lower relative population density, with inequalities and inadequate application of equitable public policies and less development of economic activities in these departments, in which, also, the level of human development of the population decreases.

Through the Clustering classification, the highest level of HDI is confirmed by the same characterization in the grouping of cluster 4, made up of the departments of the Coastal Region and considered as metropolitan cities, with greater mining development and lower HDI than cluster 1, which integrates the departments of the Sierra Region.

In both cases of application of the K-Means (cluster 1) and Clustering (cluster 3) algorithms, demonstrate better effectiveness in the separation of the clusters in a broad way, through the grouping of the departments of Loreto, San Martín and Ucayali that have the lowest HDI level and are located in the Jungle Region, characterized by lower population density and less development of economic activities.

To sum up, the study concludes that the K-Means and Clustering techniques require the classification of groups, cohere and optimize the information for decision-making in the departments under study, in order to be used to manage and achieve better levels of DHI.

VII. FUTURE WORK

Future work will be related to the measurement of Quality-of-Life Indices (ICV) of the adult population and human development indicators at a comparative level among Ibero-American countries.

Plan to analyze the Regional Competitiveness Indices and their relationship with economic and social development and compare the indicators by regions and departments in order to know their evolution and determining factors for changes in position.

In addition, project studies on problems, trends and progress in development policies by measuring management indicators according to results and products structural gaps, which guarantee the application, follow-up and monitoring of public policies with equality and equity criteria in all regions and departments of Peru.

REFERENCES

[1] D. L. Pineda-Ospina, E. G. Rodríguez-Guevara, and D. A. García-Bonilla, "Regional clusters as a strategy to overcome competitive disadvantages," Res. Dev. Innov. Mag., vol. 11, no. 1, pp. 49–62, 2020, doi: 10.19053/20278306.v11.n1.2020.11682.

[2] V. Bhagat, Y. Izad, J. Jayaraj, R. Husain, K. Che Mat, and M. Moe Thwe Aung, "Emotional Maturity Among Medical Students and Its Impact on Their Academic Performance," Tost, vol. 4, no. 1, pp. 48–54, 2017, [Online]. Available: http://transectscience.org/.

[3] M. Faizan, M. F. Zuhairi, S. Ismail, and S. Sultan, "Applications of Clustering Techniques in Data Mining : A Comparative Study," vol. 11, no. 12, pp. 146–153, 2020.

[4] P. Govender and V. Sivakumar, "Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)," Atmos. Pollut. Res., vol. 11, no. 1, pp. 40–56, 2020, doi: 10.1016/j.apr.2019.09.009.

[5] M. R. Ayyagari, "Classification of Imbalanced Datasets using One-Class SVM, k-Nearest Neighbors and CART Algorithm," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 1–5, 2020, doi: 10.14569/ijacsa.2020.0111101.

[6] M. Fradi, M. Afif, and M. Machhout, "Deep Learning based Approach for Bone Diagnosis Classification in Ultrasonic Computed Tomographic Images," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 12, pp. 80–87, 2020, doi: 10.14569/IJACSA.2020.0111210.

[7] M. Iqbal, S. Ali, M. Abid, F. Majeed, and A. Ali, "Artificial Neural Network based Emotion Classification and Recognition from Speech," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 12, pp. 434–444, 2020, doi: 10.14569/ijacsa.2020.0111253.

[8] A. K. Dehariya and P. Shukla, "Medical Data Classification using Fuzzy Main Max Neural Network Preceded by Feature Selection through Moth Flame Optimization," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 12, pp. 655–662, 2020, doi: 10.14569/IJACSA.2020.0111276.

[9] M. Faizan, M. F., S. Ismail, and S. Sultan, "Applications of Clustering Techniques in Data Mining: A Comparative Study," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 12, pp. 146–153, 2020, doi: 10.14569/ijacsa.2020.0111218.

[10] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, "Comparison of the CatBoost Classifier with other Machine Learning Methods," vol. 11, no. 11, 2020.

[11] R. Kadry and O. Ismael, "A New Hybrid KNN Classification Approach based on Particle Swarm Optimization," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 291–296, 2020, doi: 10.14569/IJACSA.2020.0111137.

[12] M. Rela, S. N. Rao, and P. R. Reddy, "Liver Tumor Segmentation using Superpixel based Fast Fuzzy C Means Clustering," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 380–387, 2020, doi: 10.14569/IJACSA.2020.0111149.

[13] T. A. Khan, K. A. Kadir, S. Nasim, M. Alam, Z. Shahid, and M. S. Mazliham, "Proficiency Assessment of Machine Learning Classifiers: An Implementation for the Prognosis of Breast Tumor and Heart Disease classification," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 560–569, 2020, doi: 10.14569/IJACSA.2020.0111170.

[14] F. Alharbi, M. K. Elbashir, M. Mohammed, and M. E. Mustafa, "Fine-Tuning Pre-Trained Convolutional Neural Networks for Women Common Cancer Classification using RNA-Seq Gene Expression," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 676–683, 2020, doi: 10.14569/IJACSA.2020.0111182.

[15] T. Sh. Mazen, "A Novel Machine Learning based Model for COVID-19 Prediction," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 523–531, 2020, doi: 10.14569/IJACSA.2020.0111166.

[16] A. D. Dondekar and B. A. Sonkamble, "Harmonic Mean based Classification of Images using Weighted Nearest Neighbor for Tagging," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 240–244, 2020, doi: 10.14569/IJACSA.2020.0111131.

[17] T. A. Sipkens and S. N. Rogak, "Technical note : Using k -means to identify soot aggregates in transmission electron microscopy images," J. Aerosol Sci., no. September, p. 105699, 2020, doi: 10.1016/j.jaerosci.2020.105699.

[18] E. Patel and D. S. Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model," Procedia Comput. Sci., vol. 171, no. 2019, pp. 158–167, 2020, doi: 10.1016/j.procs.2020.04.017.

[19] J. Morales, N. Vargas, M. Coyla, and J. Huanca, "Classification model of municipal management in local governments of Peru based on K-means clustering algorithms," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 7, pp. 568–576, 2020, doi: 10.14569/IJACSA.2020.0110770.

[20] W. Yang, H. Long, L. Ma, and H. Sun, "Research on clustering method based on weighted distance density and k-means," Procedia Comput. Sci., vol. 166, pp. 507–511, 2020, doi: 10.1016/j.procs.2020.02.056.

[21] M. Mateen, J. Wen, M. Hassan, and S. Song, "Text clustering using ensemble clustering technique," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 9, pp. 185–190, 2018, doi: 10.14569/ijacsa.2018.090925.

[22] B. M, J. Tomy, U. A, and P. Jacob, "Clustering Student Data to Characterize Performance Patterns," Int. J. Adv. Comput. Sci. Appl., vol. 1, no. 3, pp. 138–140, 2011, doi: 10.14569/specialissue.2011.010322.

[23] E. León Guzmán, "Metrics for Clustering Validation," 2019, [Online]. Available: http://www.disi.unal.edu.co/profesores/eleonguz/cursos/md/presentaciones/Sesion13_validacion_Clustering.pdf.

[24] T. Tamer, A. Haydar, and I. Ersan, "Data Distribution Aware Classification Algorithm based on K-Means," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 9, 2017, doi: 10.14569/ijacsa.2017.080946.

[25] M. Khalid, N. Pal, and K. Arora, "Clustering of Image Data Using K-Means and Fuzzy K-Means," Int. J. Adv. Comput. Sci. Appl., vol. 5, no. 7, pp. 160–163, 2014, doi: 10.14569/ijacsa.2014.050724.

Analysis of Momentous Fragmentary Formants in Talaqi-like Neoteric Assessment of Quran Recitation using MFCC Miniature Features of Quranic Syllables

Mohamad Zulkefli Adam, Noraimi Shafie, Hafiza Abas, Azizul Azizan
Razak Faculty of Technology and Informatics
UniversitiTeknologi Malaysia
Kuala Lumpur, Malaysia

***Abstract*—The use of technological speech recognition systems with a variety of approaches and techniques has grown rapidly in a variety of human-machine interaction applications. Further to this, a computerized assessment system to identify errors in reading the Qur'an can be developed to practice the advantages of technology that exist today. Based on Quranic syllable utterances, which contain Tajweed rules that generally consist of Makhraj (articulation process), Sifaat (letter features or pronunciation) and Harakat (pronunciation extension), this paper attempts to present the technological capabilities in realizing Quranic recitation assessment. The transformation of the digital signal of the Quranic voice with the identification of reading errors (based on the Law of Tajweed) is the main focus of this paper. This involves many stages in the process related to the representation of Quranic syllable-based Recitation Speech Signal (QRSS), feature extraction, non-phonetic transcription Quranic Recitation Acoustic Model (QRAM), and threshold classification processes. MFCC-Formants are used in a miniature state that are hybridized with three bands in representing QRSS combined vowels and consonants. A human-guided threshold classification approach is used to assess recitation based on Quranic syllables and threshold classification performance for the low, medium, and high band groups with performances of 87.27%, 86.86%and 86.33%, respectively.**

Keywords*—Speech processing; MFCC-Formant; Quranic recitation assessment; human-guided threshold classification

I. INTRODUCTION

The recitation of the Qur'an, which although uses Arabic words, however, is quite different from the recitation of ordinary Arabic texts. This is due to the presence of certain pronunciation rules (Tajweed) which must be followed during the recitation [1]. As a result, it can be agreed that those who are Arabs and practice Arabic are also required to learn pronunciation that conforms to Tajweed rules while reading the Quran. Tajweed rule basically emphasizes the correct and accurate pronunciation, which is called Makhraj (or plural Makhaarij). It involves the articulation point of each letter and together with determining the specific quality or characteristics (Sifaat) of each letter that distinguishes it from other sounds.Most of the applications provide Al Quran contents in text, audio and video formats without interactive tools to perform assessment of recitation.

Speech processing is widely used in human-computer interaction. Speech signals rich in speech information can be utilized by frequency modulation, amplitude modulation and time modulation that carry various components such as resonance movement, harmonic, tone intonation, force and even time. The research may lead to the approach of spectral energy and its temporal structure that can be used in speech processing in the recitation of the Quran. The measurement of the parameters and features extracted will be used to capture the nature of the speech and the parametric features to reveal the errors of Quranic recitation based on Tajweed measured from the likelihood of the parametric features.

The speech signal properties are used to be the main reference in the Quran recitation assessment computing machine, where the same method was developed and demonstrated in the Intelligent Quran Recitation Assistance (IQRA) computational engine proposed in the study presented in this paper. Firstly, the unique and salient features are identified, investigated and used to represent the digitized Tajweed rules that embedded in the recited syllable of particular Quranic word. This is then creatively and experimentally led to the creation of extractor and classifier design to underpin the task of dissimilarity grouping of Tajweed rules, where the assessment will take place. The main concern of this paper is to reveal the analysis process of the significance (momentous) level of the miniature features (fragmented formants) in producing the digital representation of the Tajweed rules (based on the syllable). By strategically using the threshold approach in the experiments, the conventional Talaqi-like approach seemingly realized digitally and formed the new modern (or neoteric) assessment.

In the remaining of the sections, the content of the paper is divided into various sections for the purpose of conveying the understanding of the proposed problem and solution. Section II discusses the signification of fragmented formants (each of several frequency bands) or momentous fragmentary formants that derived from the Mel Frequency Ceptral Coefficients (MFCC). This is then followed by Sections III and IV for the experiment and human-guided results, respectively. The outcome of the paper is concluded in Section V with comments and recommendation.

II. Momentous Fragmentary MFCC-FormantS

Although speech recognition techniques have evolved drastically and have begun to improve in application construction, they are still the most challenging method to analyse spoken language based on pattern recognition or machine ability in learning pattern development more interactively. Speech recognition commonly used in Arabic and Quran recitation are such as Arabic coding and synthesis research, dialect detection, speaker recognition, memorization and sentence retrieval. A large number of analyses use word or sentence utterance approach techniques [2] to identify and evaluate from signal speech representation. Spoken or readings are present as a form of language because speech also contains basic acoustic sounds or also known as phonemes. Each phoneme sound released is usually influenced by a neighbourhood phoneme delivered with a syllable or word. The recitation of the Quran conveys words spoken with a certain rhythm that can be formulated as an acoustic-phonetic symbol and prosody. Challenges of developing such a system are centralized to the modelling of features extraction and matching process that significantly are able to describe the recitation errors and intelligently propose the Tajweed error detection. Acoustic phonetics symbols of Arabic language can be formed as consonants and vowels. Each of combination Arabic phonetic symbols can be represented as a syllable and word. There are six pattern combinations of vowels which are CV, CV: CVC, CVCC, CV:C, CV:CC, where C represents the consonant and V as a vowel[3] while V: as a long vowel. The approach of analysis concerns on sequences of voiced or unvoiced sound because of recitation are related to phonetic and prosody. The sequences are segmented in a series of frames and represented by formant frequencies[3]and [4]. The production of voice or speech involves the movement of air from the lung to vocal tract towards the lips. The combination of voice production mechanism produces a variety of vibration and spectral-temporal composition that produce different speech sound. Apparently, the Arabic phonetic sound was produced from the specific articulation places or regions in the vocal tract. Speech or voice response is produced through the vocal tract filter that is characterised by series of formant or resonant frequencies [5].

The sound spectrum can be represented by formant frequencies which show the greater intensity of sound quality. The quality of sound is greatly shown using formant frequencies, especially the characteristic of sound of the consonant [6]. In this case, the formant frequencies of f_1, f_2, f_3 and $f_4$as illustrated in Fig. 1 are used as features to model the Quranic alphabet pronunciation [6]. Theoretically, the combination of formant frequencies of f_1, f_2, f_3 and $f_4$from speech production should be able to describe the characteristics of the letters during pronunciation for each of the 28 hijaiyah letters (Arabic letters).Furthermore, the changes of formant frequencies of f_1, f_2, f_3 and f_4 can be used to represent the characteristic of vowel, consonants and its combination [7][8]. There are large number of techniques used in speech processing and feature extraction can be used such as Mel Frequency Cepstral Coefficient (MFCC), PLP and LPC techniques [9]. MFCC is the most popular feature extraction technique compared to other techniques because it is related to the human auditory system [10]. In addition, MFCC can produce better accuracy with less computational complexity.

The momentous fragmentary MFCC-formant frequency is experimentally invented and introduced for representing the syllable feature with detail analytical approach. This is done by dividing the MFCC and its derivative feature into Band-1, Band-2 and Band-3 for representing the formant frequency ranges of low frequency, medium frequency and high frequency respectively. Each band has been broken into four (4) co-efficient of MFCC which represent the frequencies from the filter triangular bank. Each of the co-efficient derives a sequence of frames of power energy. The value energy in every frame is basically the total energy from multiple filters in the MFCC. The combination of frames from the first frame to the 'n = 1, 2, 3, …' frame is the concatenated $MFCC_n$ and its derivatives ($\Delta MFCC_n$ and $\Delta\Delta MFCC_n$). The co-efficient and their appropriated power energy frames can be considered as the miniature feature that will characterize the respected syllable feature. This approach takes into account the evaluation of vowel and consonant features in syllable pronunciation based on low, medium and high formant frequency ranges. The selection of band is also closely related to vowel and consonant of phoneme speech spectrum. Table I shows how the three bands have been fragmented by dividing the selected range of frequencies based on experiments. Vowel and consonant that are categorized as voiced phoneme have high power energy characteristics. In Band1, the MFCC coefficient (C1, C2,…, C12) as described in Table I are categorized based on the formant frequency range. The formant of f_1 which has coefficients C1, C2, C3 and C4 is indicative of High-Voiced Vowel and Low-Voiced Consonant Characteristic. While band2 represents the formant frequency of f_2 and $f_3$are represented by the coefficient C5, C6, C7 and C8 indicate the characteristics of High-Voiced Consonant and Low-Voiced Vowel Characteristic. The final category is band3 for the formant range, $f_4$and above. This category is represented by C9, C10, C11, and C12 which exhibit Voiceless and Low-Voiced Consonant Characteristic features. For the experiments conducted, band1 and band2 are very practical to reveal information about vowels, while information about sound consonants is more appropriately revealed on a combination of band1, band2 and band3. Band3 is therefore used to reveal voiceless consonant information. Table I lists the frequency ranges for the 3 bands and its coefficients.

Each band shows the characteristic of formant frequency of syllable pronunciation that is produced from the vocal tract filter response[11].

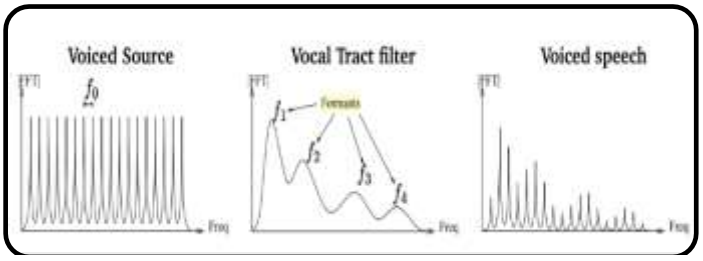


Fig. 1. The Source and Filter Response for a Typical Vowel Sound [5].

TABLE I. THE BAND OF FREQUENCY RANGE

Fragmentary Frequency	Band1 (300Hz-1077Hz)	Band2 (861.3Hz-2089Hz)	Band3 (1787Hz-3704Hz)
Syllable Pronunciation	High-Voiced Vowel and Low-Voiced Consonant	High-Voiced Consonant and Low-Voiced Vowel	Voiceless and Low-Voiced Consonant
MFCC-Formant co-efficient	C1, C2, C3 and C4	C5, C6, C7 and C8	C9, C10, C11 and C12

III. EXPERIMENT

The structure of IQRA implementation for the Al-Quran recitation assessment has been divided into three (3) stages as shown in Fig. 2. The first stage is data acquisition andpre-processing. The second stage is feature's extraction and last stage is Human-Guided Threshold classification.

The assessment algorithm of Quranic recitation uses tactical hybrid methodical DSP approaches by combination of various machine learning [12] conventional approaches. The flow of proposed computational engine for Quranic recitation assessment is described in the following sub-sections.

A. Data Acquisition and Pre-Processing

Data acquisition is recorded from numerous Malay reciters of various backgrounds. This includes male and female of Malay ethnic, ranging from the age of between 20 to 65 years old. There are two categories of selected reciters which are experts and learners. This Quranic Recitation Speech Signals (QRSS) originally also contained unwanted audio such as noise or any surrounding audio that is difficult to predict. However, the signal compensation method is used to eliminate these unwanted signals, which include such as the 60Hz Hum AC-DC signal [13], the silent signal [14], breaths sound signal, clicks and pops sound [9] that can interfere the performance of the computing engine.The wav format is a commonly recorded audio data format, for example with 16 bits, 44,100 samples [9] and uses mono channels.

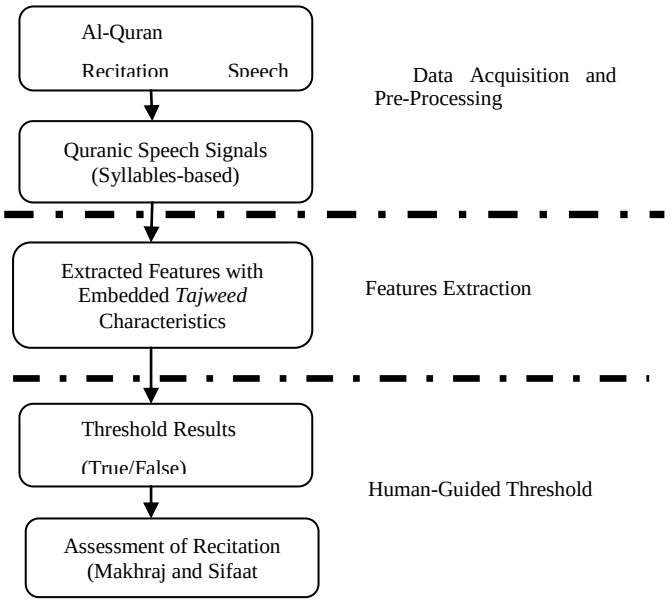


Fig. 2. The Structure of IQRA Implementation.

The main aim of signals initialization is to prepare the signals with several selected techniques that should be able to enhance signals representation. The steps of initialization are start-end point detection [14], pre-emphasis [15] and amplitude normalization. The end point detection is used to define the start point and end point of Quranic speech signals. Each of learners or experts have different start point and end point while do recitation. Combined zero crossing and short term energy function are used to determine start point and end point [16]. Therefore, the amplitude normalization is used to compensate the speaker health condition, age and gender and change the amplitude range between 0 and 1. Meanwhile, pre-emphasis converts the QRSS to the higher frequency with the co-efficient of 0.95. There will be more information can be extracted by converting the signal into high frequency spectrum as compared to the one in low frequency.

In confronting the variability and complexity of the continuous QRSS, the recitation of the experts and learners should be parameterized by a single warp factor. Based on vocal tract speech production, the air flow of speech production among reciters is differently delivered and it's involved of Vocal Tract Length (VTL). VTL is varied across different reciters around 18cm and 13cm for males and females, respectively. The positions of formant frequency are inversely proportional to VTL, and the formant frequency can vary around 25% [17]. The main purpose of speaker adaptation is to get the same rhythm, tone and length between expert and learner QRSS that can be compared in same word/utterance articulation from the Vocal Tract Length Normalization (VLTN)[18]. The DTW is used to warp QRSS energy of speech in the same length of recitation in the time series frame.

B. Feature Extraction and Prediction Model

The speech signal is basically a non-linear signal and needs to be handled with systematic processing. Thus, in this paper, the approach of Mel-Frequency Cepstral Coefficients (MFCC) and formant frequencies features (MFCC-Formant) are selected to reveal the characteristic of syllables by manipulating the power energy. The speech signals are segmented by time frame and also by frequency domain to derive the cepstral coefficients. The MFCC-Formant-like features are used as an acoustic model to indicate the pattern similarity and dissimilarity of Al-Quran recitation. The characteristic of the shape of the energy spectrum can be aligned as an acoustic model of Al-Quran recitation which represents the energy, rhythm and tone. The significant miniature feature for cepstrum energy and its derivative feature are extracted with the aid of cepstral analysis [16].

MFCCs have been widely used in the field of speech recognition and have successfully demonstrated dynamic features as they extract linear and non-linear properties of signals. MFCC and its derivatives (Δ and ΔΔ MFCC) are formed and grouped together to represent the transformed syllable. This QRSS is produced from the articulated speech production that consists of information of rhythm and intonation energy. ΔMFCC is known as delta coefficient (differential coefficient), while ΔΔMFCC is known as shift delta coefficient (acceleration coefficient) where both show the properties of trajectory of power energy between

segmented frame of syllable representation. MFCCs are more general acoustic features which are largely used in the systems that applied Good of Pronunciation (GoP) [19]. Firstly, use the Hamming window to find the magnitude of the signal by using the Fourier Transform. Secondly, map the power spectrum in frequency domain that obtained from the mel scale by using the triangular overlapping windows filters. Then, take logarithm power at each mel frequency and apply Discrete Cosine Transform (DCT) of mel log powers as if it were a signal. Lastly, MFCC represents the amplitude of the resulting spectrum. The block diagram of MFCC is show in Fig. 3.

Furthermore, Fig. 3 depicts the step and design approaches of feature extraction method for MFCC, ΔMFCC and ΔΔ MFCC then represent as MFCC-Formant miniatures features. In step 1, spectral analysis is used to determine the frequency formant content of the arbitrary signals of QRSS. The overlap frame that uses the hamming window is used to reduce the spectral leakage effect. On the side of hamming window, lobe is overlapped, and the main lobe captures the characteristic of spectral energy by using the Discrete Fourier Transform (DFT). The selection of hamming window is performed because of its least amount of distortion. The frame size must be controlled and not too large in order to prevent the QRSS syllable properties from being too much across the window, thus affecting the resolution of time, whereas if the frame size is too short, the resolution of the narrow-band component will be sacrificed, and this will adversely affect the frequency resolution. A large number of previous experiments using MFCC have stated that frame measurements for spectrograms preferably between 20ms and 40ms to optimize a sample sufficient to obtain reliable spectrum estimates and depend on the length of utterance. The frame size of 20ms and the frame shift of 10ms also have shown reliable spectrum estimation [21]. In this experiment, the chosen shape of spectrogram is framed between 25ms and frame shift is 10ms based on phoneme formant representation.

In step 2, the Mel scale is used to obtain the power spectrum for each frame. This can be done by using a triangular window filter where each of them is not the same size in terms of amplitude. The amplitude decreases with increasing frequency range, and this is to get the characteristics of low frequency and high frequency that can be heard by the human ear. The human ear is basically more sensitive to low frequencies.

In step 3, the logarithm power at each of filters is measured for every segmented frame. Thus, each of bin per frame per filter holds the log-energy for each filter channel. In the experiment done in this thesis, 20 numeric values are obtained for each frame at the output. The outputs are stored in a matrix form with the number of row represent the frame (size frame of QRSS syllable) and the number of columns equal to 20 (which is the number of filters in the filter bank).

In step 4, DCT converts the power spectrum log generated by the mel scale in the frequency domain to the time domain. The DCT will rearrange the co-efficient cepstral from small order to a large sequence based on the evaluation of cosine signal characteristics.

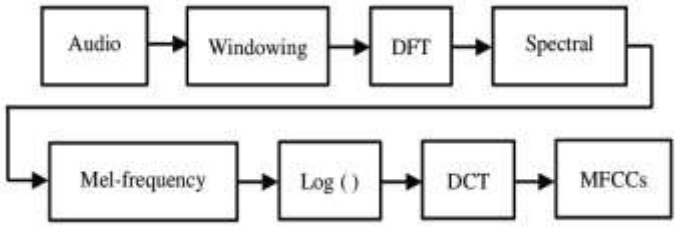


Fig. 3. The Block Diagram of MFCC [20].

In step 5, there are 13 MFCC-Formant coefficients generated from the QRSS syllable but only 12 coefficients are selected. The first coefficient (C0) representing the natural frequency (Pitch property) of the syllable indicates the amount of power energy but, not included in the analysis in this thesis. There are only from C1 to C12 MFCC-formant co-efficient are used in the analysis and taking the frequency of the band between 300Hz and 3700Hz. The 12 delta (ΔMFCC) and 12 ΔΔMFCC were concatenated together to represent the MFCC-Formant features of each QRSS syllable.

The speech signal is required as a stationary signal to estimate the parameters. The stationery signals were parameterised as features coefficient in such a manner before measuring the similarity through the matching or recognition process. Recitation of Al-Quran commonly can be assessed or evaluated using non-phonetic or phonetic transcription. In this paper, non-phonetic transcription is an approach by designing the prediction model without reference set of transcription. The parameter estimation algorithm of model prediction is estimate by using MLLR (Maximum Likelihood Linear Regression). These algorithms are integrated with GMM to classify the feature pattern as statistical model approach. These GMM statistical models have their characteristic which represent the signal characteristic as a static pattern [16]. The MLLR computes a set of transformations which reduces the mismatch between an initial model set and the adaptation data [22].

Parameter estimation is used to represent the acoustic model based on MFCC-formant-liked features and is designed to measure the similarity and dissimilarity (likelihood) of syllable pronunciation. The machine learning approach has great attention on parameter estimation in speech processing as data modelling. The QRAM is obtained by establishing a few tasks and methods to be applied. The fragmentary MFCC-formant features are proposed and modelled by using the Gaussian Mix Model (GMM). The GMM is a probabilistic model to represent the subpopulation and works well with the parameter estimation strategy. Generally, GMM is one of statistical-based clustering methods and an effective model that capable of achieving the high identification accuracy for short utterance length. Although MFCC is not robust to noise, the model-based approach used in this thesis able to eliminate the noise by the cancellation performed by Maximum Likelihood Estimation (MLE). MLE is a standard approach to estimate the model parameters from the sampling data. MLE is measured based on Expectation Maximization (EM) for parameters estimation approach. The EM algorithm is an iteration method to find the MLE of latent of hidden variables. The estimated parameters based on mean, covariance and

weight indicated the similarity and dissimilarity of each syllable pronunciation for every learner or reciter. The expected mean, covariant and weight of GMM for 4-Dimensional data are figured out by EM as mentioned before, where the EM algorithm is not possible to optimize the log likelihood of $\log[p(x|\lambda)$ directly with respect to λ. This means that the observation data, X= x1, x2,…,xD can be introduced by the discrete random variable, Z= z1, z2, …, zD and model parameters $\lambda=\{\overline{w}_i, \bar{\mu}_i, \bar{\Sigma}_i\}$. The log likelihood of model λ is given by

$$\sum_{d=1}^{D} \log p(x_d|\lambda) =$$

$$\sum_{d=1}^{D} \log p\left(x_d, z_{d,}|\lambda\right) - \sum_{d=1}^{D} \log p\left(z_d, x_{d,}|\lambda\right) \quad (1)$$

The expression also impresses that the statistical distribution of 4-dimensional observation of MFCC data can be clustered into 4-cluster of GMM. When the model of λ from different observation for different reciters is considered, each model will calculate the MLE parameters to represent the likelihood among the reciters for different band of MFCC. The parameters are trained as unsupervised classification. This model is designed by combining 4 GMM clusters using 4 dimensional fragmentary MFCCs to find the MLE that represents the data distribution for each of these frames. Furthermore, this model represents the sequence of MFCC-Formant sample frames that are considered parametric distribution models. The resulting MLE parameters show the maximum data calculated from the GMM model generated from the data that have been observed. This parameter is defined as a blueprint for the model. In avoiding the GMM overfitting, Bayesian Information Criterion (BIC) is used to estimate the reasonable amount of data prediction done by GMM. For instance, if the BIC value is much lower, the model is considered better in predicting data. BIC is an asymptotically optimal method for estimating the best model using only sample estimates [23]. BIC is defined as

$$BIC = -2 \ln I(x, M) + k \ln(n) \quad (2)$$

where x are the sample data, l(x,M) is the maximized likelihood function under a model M. While k is the number of estimated parameters, and n is the sample size.

The statistical clustering GMM Model approach is used to measure the similarity and dissimilarity of QRAM by estimating the maximum likelihood of fragmentary band of MFCC miniature features. It is a prototype-based algorithm which is consists of the feature vectors and representing as a mixture of Gaussian distribution. A mixture model is a probabilistic model for representing the presence of subpopulations within an overall population. The mixture shows probability distribution of parameters and can represent as the number of mixture components approaches to infinity. However, the appropriate number of mixtures must be determined for each model so that the mixtures are able to show the best distribution for the parameters or data where the distribution shows the characteristics of the parameters. Thus, the data will be segmented based on similarities or differences between observations in the dataset by 4-mixtures GMM as shown in Fig. 4. The similarities or differences are represented by the maximum likelihood estimation (MLE) as illustrated in Fig. 5. Each model of recitations is a set of model parameters with estimated mean vector, covariance matrix and mixture weight. Each of parameter models is trained as unsupervised classification by using the expectation maximization (EM).

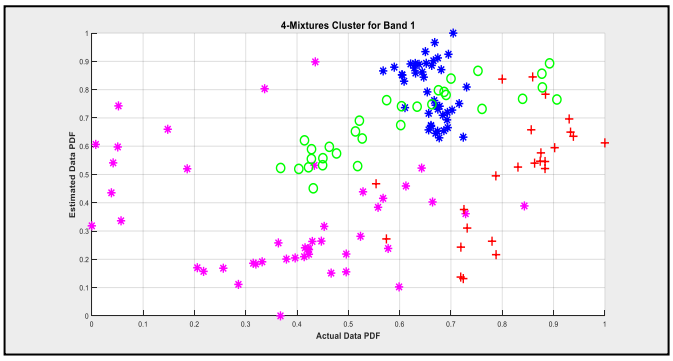


Fig. 4. The 4-Mixtures of GMM.

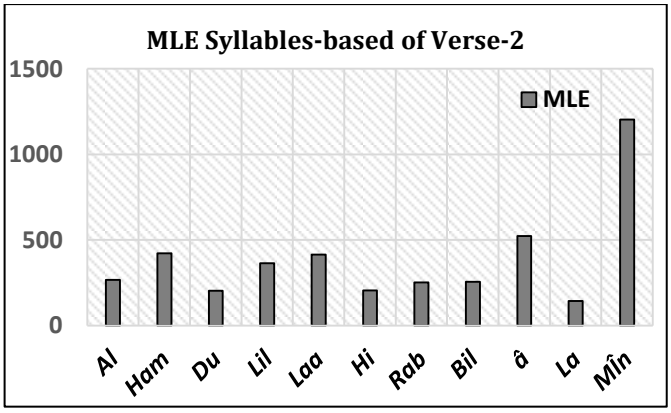


Fig. 5. The MLEs Value of Quranic Syllable-based Verse-2.

C. Human-Guided Threshold Classification

The computational engine score threshold process is used to evaluate the similarity and the dissimilarity based on human-guided threshold classification. This successful threshold process shows conventional Talaqi processes (experts evaluate the recitation by learners based on how to pronounce syllables in the verses of the Quran) are transformed to a machine evaluation approach. Computational engines must have salient features that can distinguish between correct and incorrect readings. Therefore, in determining the score, analysis of salient features, and matching process is used to obtain reading assessment based on the actual assessment by experts called a human-guided or Talaqi-Like assessment. Talaqi-like process has been used in the training phase and also the testing phase process in the computational engine. This is to ensure that the assessment by the expert is always included in the assessment made by the machine.

In this process, the MLE parameters are used as representations to each syllable recited by the learners. Initially, the MLE values from the expert readings were used as the initial reference in determining the initial threshold by assuming that all MLEs produced by the expert readings were within acceptable thresholds. After that the learners' reading is

assessed by the initial experts' threshold. Thus, the initial threshold will change to a new threshold after undergoing the training process by Human guided assessment (conducted by prominent expert). This process will be repeated until all MLE parameters have been evaluated by a prominent expert. Finally, the value of the threshold range has been completely obtained and can be used as a benchmark the reading made by the learners is correct or otherwise. The value of this Human-Guided Threshold classification will be tested in the testing phase and the performance is calculated.

IV. Result of Human-Guided Classification

A. Talaqi-Like Training Phase

In this training phase of the classification stage, the initial parameters of MLEs are taken from the calculation of 12 expert recitations. The starting point of training phase is when the input given to this designed system begins to create a change of pattern or minimum and maximum MLEs value that limits the correctness of a Tajweed in the reading of the Al-Fatehah chapter. This is seemingly caused by the changes of the acceptable lowest and highest values of that correspondingly due to the variability demonstrated by various reciters, but remains accepted (Acceptance Threshold) by the expert. The process of correcting (or training) the minimum and maximum values (threshold range) is firstly performed on the group of experts' MLEs data. This is the initial threshold range and used as reference values to be compared with the learner recitations. Secondly, the MLEs values obtained from the recited syllables of 40 learners are matched with the expert threshold range. Besides the setting of minimum and maximum values, the indication of True Acceptance (TA), False Rejection (FR) and False Acceptance (FA) of the calculated MLEs are counted and accumulated. Tabulates the MLEs values of syllables verse-2 of Al-Fatehah recited by 40 learners have been matched with the threshold range of expert's recitations. In the classification process performed in this experiment, the threshold range selected based on this expert indicates that most syllables are categorized as FR (False Rejection). This is logically agreeable and reasoned by the experts that most of the learners' performance has not been perfectly pronounced, but the Tajweed rules are acceptable.

The learning process for the machine evaluation to accurately perform is by allowing the human expert to guide the evaluation manually (Talaqi-Like approach). This is where the core of operations in the transfer of knowledge from human to a machine has taken place in the process of training the machine. Experts have individually altered the assessment performed by the machine in the case of (True) Rejection by the machine. This situation occurs by manually record or mark in the form that has been prepared for each learner. At the same time, the corresponding MLE values will be re-accepted as correct Tajweed recitation and assigned as True Acceptance (TA), which in turn will change the threshold range to new values (Minimum or Maximum). Fig. 6 shows the comparison of performance of classification for MLE band-1between initial expert threshold and Talaqi-Like threshold.

The acceptance of true recitation is based on three band threshold range categories of MLE, which indicate the similarity and dissimilarity for each syllable in Al-Fatehah verses, as compared with the adjusted reference threshold range. Similarity range indicates the acceptance and dissimilarity indicate unacceptance of recitation. The overall performance of acceptance recitation threshold is higher 80% for all MLEs. It shows that the MLEs parameter estimation can used indicator to assess the Quranic recitation assessment. However, the testing phase is used as validation stage to proof the experiment of Quranic assessment reliability. Fig. 7 shows the performance of Human-Guided Threshold classification based on true acceptance (TA) in training phase.

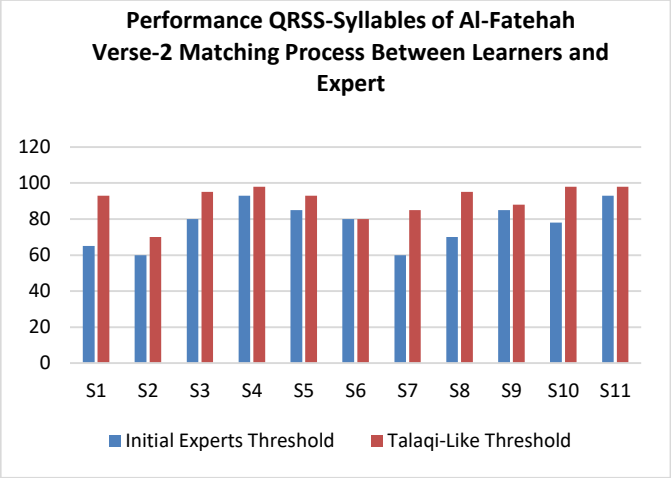


Fig. 6. Comparison Performance between Initial Experts and Talaqi-Like Threshold.

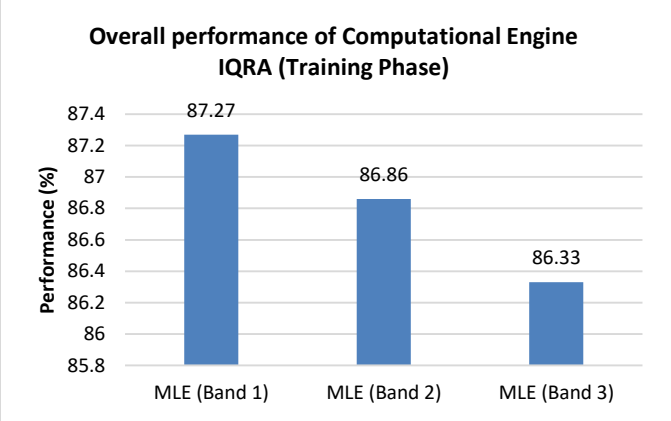


Fig. 7. Performance of IQRA in Training Phase.

B. Talaqi-Like Testing Phase

In the testing phase, the main objective is as linked from the training phase, which is to test the computational engine that has been designed in the context of reliability of the miniature salient feature, extractor and classifier. The trained range of MLEs is used to assess the performance of test data. Each syllable is tested according to the threshold determined based on MLE Band1, Band2 and Band3 (Human-Guided Threshold range). A total of 40 different learners from the training phase took their readings and the readings of each syllable in Al-Fatehah were extracted and matched with the reference MLEs from the training phase. Each test data is also evaluated manually by an expert and the performance of the reading truth that refers to Tajweed rules is calculated in a

technical context, namely, true / false positive acceptance (TP and FP), false rejection (FR) and false acceptance (FA). The comparison of errors will be made and analyzed between the machine evaluation and human evaluation. From here, the performance of the machine in terms of performing as an evaluator is then measured with respect to human expert performance.

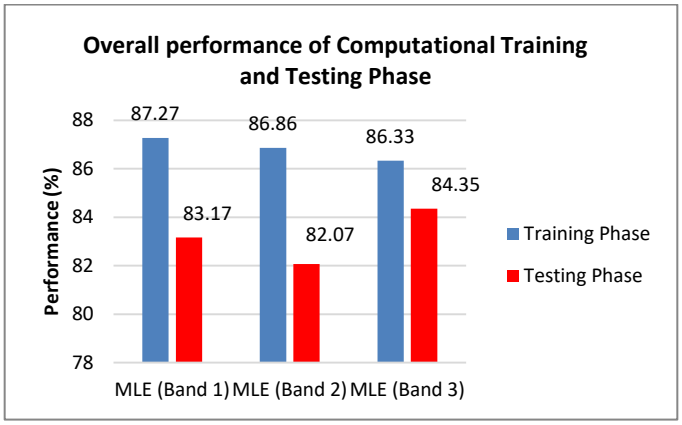


Fig. 8. IQRA Computational Training and Testing Phase Performance.

Based on Fig. 8, the classification performance using the threshold method for these MLEs parameters can be used to evaluate syllable-based Al-Quran recitation that Tajweed rules are embedded in the syllable. With this performance, the conclusion that can be expressed is that each band-1, band-2 and band-3 MLEs are able to show the characteristics of vowel and consonant combinations in each syllable based on fragmentary frequencies. These features have shown impressive performance of over 80% for representing Tajweed rules based on Makhraj, attributes and also derivatives rules.

V. Result and Discussion

The True Acceptance (TA) indicates both true positive and negative of syllables recitation based on location of MLEs. True positive shows that the learners have MLEs parameters are in the range of independent assessment threshold. It also shows that learners pronounced the syllables correctly. While true negative show that the learners pronounced the syllables incorrectly but MLEs parameters are out from threshold range. In addition, FR shows that the learners pronounced the syllable correctly but the location of MLE is located out from the threshold range. While FA indicates the pronunciation of syllables is incorrect, but the parameters MLE are in the range of independent assessment threshold.

Referring to Fig. 9, the total of FRR is revealed as 81.98%, while the total of FAR is 18.02%. This data represents all 40 learners involved in this testing phase. The plotted graph depicts the value of accuracy in evaluating readings depends on the ability to isolate either FA or FR. This ultimately leads to the sensitivity of the selected threshold change, where a reduction in FA will result in an increase to FR. In a simple interpretation, it is important for a learner to prioritize accuracy in reading in full compliance with Tajweed law. Therefore, the reduction in FA is considered better although it will lead to an increase in FR.

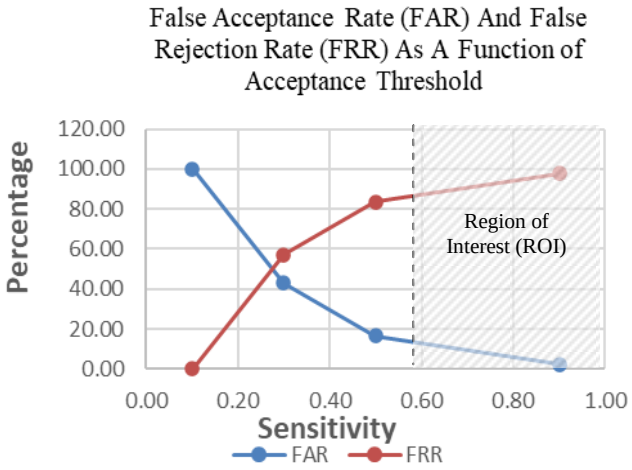


Fig. 9. FAR and FRR-The Function of Acceptance Threshold.

The goodness of pronunciation evaluation is used to evaluate the range and performance of recitation acceptance. The performance is measured by finding the threshold pattern for each syllable based on MLE parameters using GMM. In the training phase, there are two threshold processes involved, which are initial (or reference) threshold evaluation that based on expert recitation and expert-guided machine assessment. The overall performance of MLE Band-1, -2, -3 for Al-Fatehah verses are 86.33%, 86.86%, and 87.27%, respectively. The designed computational engine for IQRA (recitation assessment) system demonstrated through the use of fragmented frequency parameters (3 Bands) with the creation of salient miniature features of MLE along with machine learning has been implemented perfectly.

VI. Conclusion

Human-guided threshold classification process is studied and updated repeatedly based on observations given by prominent experts by looking for MLE parameters. This paves the way for the machine learning process through human-driven threshold values where the machine is able to assess learner recitation using MLE. The threshold is based on the probability or similarity of the MLE parameters of MFCC-Formant features for the syllables spoken by the reciters. The matching process practiced by this machine that uses human-guided threshold limit values can be interpreted as equivalent to the Talaqi approach as in the conventional evaluation process. In other words, the technological assessment in computer machines highlighted in this paper has successfully matched the way the assessment process of Quran recitation is done in conventional practice. Indeed, processes with highly systematic tactical and methodical approaches and techniques in combination with the role of human expertise and the advantages of the application of technology have been successfully demonstrated to produce a practical evaluation model.

Acknowledgment

Thank you to Universiti Teknologi Malaysia (UTM) for the financial support in sponsoring this research project, under

the program of UTM Enhancement Research Grant (UTMER).

REFERENCES

[1] Y. Mohamed, M. Hoque, T. H. S. Bin Ismail, M. H. Ibrahim, N. M. Saad, and N. N. M. Zaidi, "Relationship between Phonology, Phonetics, and Tajweed: A Literature Review," vol. 518, no. ICoSIHESS 2020, pp. 407–411, 2021, doi: 10.2991/assehr.k.210120.153.

[2] L. Marlina et al., "Makhraj recognition of Hijaiyah letter for children based on Mel-Frequency Cepstrum Coefficients (MFCC) and Support Vector Machines (SVM) method," 2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018, vol. 2018-Janua, pp. 935–940, 2018, doi: 10.1109/ICOIACT.2018.8350684.

[3] R. D. Kent and H. K. Vorperian, "Static measurements of vowel formant frequencies and bandwidths: A review," J. Commun. Disord., vol. 74, no. November 2017, pp. 74–97, 2018, doi: 10.1016/j.jcomdis.2018.05.004.

[4] S. Khairuddin et al., "Classification of the Correct Quranic Letters Pronunciation of Male and Female Reciters," in IOP Conference Series: Materials Science and Engineering, 2017, vol. 260, no. 1, doi: 10.1088/1757-899X/260/1/012004.

[5] M. N. Stuttle, "A Gaussian Mixture Model Spectral Representation for Speech Recognition," no. July, p. 163, 2003.

[6] S. Ahmad, S. N. S. Badruddin, N. N. W. N. Hashim, A. H. Embong, T. M. K. Altalmas, and S. S. Hasan, "The Modeling of the Quranic Alphabets' Correct Pronunciation for Adults and Children Experts," 2nd Int. Conf. Comput. Appl. Inf. Secur. ICCAIS 2019, pp. 1–6, 2019, doi: 10.1109/CAIS.2019.8769590.

[7] M. Farchi, K. Tahiry, S. Mounir, B. Mounir, and A. Mouhsen, "Energy distribution in formant bands for arabic vowels," Int. J. Electr. Comput. Eng., vol. 9, no. 2, p. 1163, 2019, doi: 10.11591/ijece.v9i2.pp1163-1167.

[8] K. Tahiry, B. Mounir, I. Mounir, L. Elmazouzi, and A. Farchi, "Arabic stop consonants characterisation and classification using the normalized energy in frequency bands," Int. J. Speech Technol., vol. 20, no. 4, pp. 869–880, 2017, doi: 10.1007/s10772-017-9454-9.

[9] T. Roy, T. Marwala, and S. Chakraverty, "Precise detection of speech endpoints dynamically: A wavelet convolution based approach," Commun. Nonlinear Sci. Numer. Simul., vol. 67, pp. 162–175, 2019, doi: 10.1016/j.cnsns.2018.07.008.

[10] M. Bezoui, A. Elmoutaouakkil, and A. Beni-Hssane, "Feature extraction of some Quranic recitation using Mel-Frequency Cepstral Coeficients (MFCC)," Int. Conf. Multimed. Comput. Syst. -Proceedings, vol. 0, pp. 127–131, 2017, doi: 10.1109/ICMCS.2016.7905619.

[11] N. Shafie, M. Z. Adam, H. Abas, A. Azizan, and K. Lumpur, "Sequential Classification for Articulation and Co-Articulation Classes of al-Quran Syllables Pronunciations Based on GMM- MLLR," AIP Conf. Proc., 2020.

[12] G. Aggarwal and L. Singh, "Comparisons of Speech Parameterisation Techniques for Classification of Intellectual Disability Using Machine Learning," Int. J. Cogn. Informatics Nat. Intell., vol. 14, no. 2, pp. 16–34, Feb. 2020, doi: 10.4018/ijcini.2020040102.

[13] L. R. Rabiner and R. W. Schafer, "MATLAB Exercises in Support of Teaching Digital Speech Processing," IEEE Int. Conf. Acoust. Speech Signal Process, pp. 2480–2483, 2014.

[14] M. Asadullah and S. Nisar, "A Silence Removal and Endpoint Detection Approach for Speech Processing 3 rd International Multidisciplinary Research and Information Technology," 3rd Int. Multidiscip. Res. Conf. 2016, no. September 2016, pp. 119–125, 2016.

[15] K. Livescu, P. Jyothi, and E. Fosler-Lussier, "Articulatory feature-based pronunciation modeling," Comput. Speech Lang., vol. 36, pp. 212–232, 2016, doi: 10.1016/j.csl.2015.07.003.

[16] N. Shafie, M. Z. Adam, S. Mohd Daud, and H. Abas, "A Model of Correction Mapping for Al-Quran Recitation Performance Evaluation Engine," Int. J. Adv. Trends Comput. Sci. Eng., vol. 8, pp. 208–213, 2019.

[17] L. Saheer, J. Dines, and P. N. Garner, "Vocal tract length normalization for statistical parametric speech synthesis," IEEE Trans. Audio, Speech Lang. Process., vol. 20, no. 7, pp. 2134–2148, 2012, doi: 10.1109/TASL.2012.2198058.

[18] N. Shafie, M. Z. Adam, and H. Abas, "Al-Quran Recitation Speech Signals Time Series Segmentation for Speaker Adaptation using Dynamic Time Warping," J. Fundam. Appl. Sci., vol. 10, pp. 126–137, 2018, doi: 10.4314/jfas.v10i2s.11.

[19] M. Maqsood, A. Habib, and T. Nawaz, "An efficient mispronunciation detection system using discriminative acoustic phonetic features for Arabic consonants," Int. Arab J. Inf. Technol., vol. 16, no. 2, pp. 242–250, 2019.

[20] H. C. Junho Son, Chon-Min Kyung, "Practical Inter-Floor Noise Sensing System with Localization and Classification," MDPI, no. August, 2019.

[21] M. Al-Ayyoub, N. A. Damer, and I. Hmeidi, "Using deep learning for automatically determining correct application of basic quranic recitation rules," Int. Arab J. Inf. Technol., vol. 15, no. 3A Special Issue, pp. 620–625, 2018.

[22] D. P. Lestari and A. Irfani, "Acoustic and language models adaptation for Indonesian spontaneous speech recognition," ICAICTA 2015 - 2015 Int. Conf. Adv. Informatics Concepts, Theory Appl., pp. 1–5, 2015, doi: 10.1109/ICAICTA.2015.7335375.

[23] J. Ding, V. Tarokh, and Y. Yang, "Model Selection Techniques: An Overview," IEEE Signal Process. Mag., vol. 35, no. 6, pp. 16–34, 2018, doi: 10.1109/MSP.2018.2867638.

IoT-based e-Health Framework for COVID-19 Patients Monitoring

Fahad Albogamy
Computer Sciences Program, Turabah University College
Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia

***Abstract*—The COVID-19 pandemic, produced by the SARS-CoV-2 virus, has caused global public health emergency, with the rapid evolution and tragic consequences. The fight against this disease, whose epidemiological, clinical, and prognostic characteristics are still being studied in recent works which is forcing a change in the form of care, to include transforming some face-to-face consultations into non-face-to-face. Recently, various initiatives have emerged to incorporate the Internet of Things (IoT) in different sectors specially the health sector generally and in e-Health systems specifically. Millions of devices are connected and generating massive amounts of data. In this sense, based on the experience in the health sector in the management of the pandemic caused by COVID-19, it has been determined that monitoring potential patients of COVID-19 is still a great challenge for the latest technologies. In this paper, an IoT-based monitoring framework is proposed to help the health caregivers to obtain useful information during the current pandemic of COVID-19, thus bringing direct benefits of monitoring patient's health and speed of hospital care and cost reduction. An analysis of the proposed framework was carried out and a prototype system was developed and evaluated. Moreover, we evaluated the efficacy of the proposed framework to detect potentially serious cases of COVID-19 among patients treated in home isolation.**

***Keywords*—COVID-19; IoT; healthcare; e-Health**

I. Introduction

Currently, the Internet of Things (IoT) is one of the technologies that have been widely disseminated in different application contexts [1]. IoT can be defined as a technology that consists of millions of devices that are connected to the Internet [2]. This network of devices exchange, add and process information about their physical environment to provide value-added services to the end users. Thanks to the advantages provided by IoT, its use has been made possible in different sectors such as industry, tourism, health, and the environment, which has made possible the construction and formation of smart cities [3].

There has been a remarkable growth in the number of connected devices in recent years. Since 2008, these have already surpassed the number of inhabitants on earth, and could reach 75 billion connected devices by 2025 [4]. An example of this scenario is the smart lamps, which even when they are not emitting light must remain connected, waiting for a switch on, whether from a human user or from another computer system. Another example is wearable devices [5], which constantly monitor a person's physical activity, and can issue alerts if any of the indices fall outside the desirable range [6]. With this view, the trend is that it is increasingly necessary for isolated items to connect to share information. Therefore, just as it is now uncommon to have a computer disconnected, it will soon be unusual to have an air conditioner or a coffee maker in this same situation.

Worldwide, the commercial automation and home automation areas are, among the IoT technologies, those that currently attract the most investment and with the largest installed IoT parks in the world [7]. Despite this, many other areas can also take advantage of the growth, evolution, and cost-effectiveness of this technology to expand its possibilities. Health care is one of these areas, where the use of IoT can help in various fields, whether streamlining medical care, emergency response time, predicting the occurrence of serious events, among other possibilities [8]. In all cases, the expected benefit is the improvement in the quality and life expectancy of the public using IoT technology [9].

Coronavirus is a new disease that comes from the group of coronaviruses which has seven distinct types in this family. four types of this family are mild, like a catarrh. COVID-19 was primary identified in Wuhan City, China, and has been officially named "SARS-CoV2." It causes cough, fever and difficulty of breath. The incubation period is officially two to fourteen days after exposure. Infection can range from very mild to advanced pneumonia, but it appears that 80% of cases are mild [10].

With the COVID-19 pandemic, the researches on personal health sensing equipment became even more intense [11]. Thermometers and pulse oximeters, for example, have become even more popular. One of the impacts of this was that many wearable devices, such as smart watches, incorporated these features into their latest models. using these devices, it is possible to monitor interesting data about the user's health, such as heart rate, blood oxygenation and body temperature - depending on the device model [12].

The demands of the health area that can be supported by IoT are extensive, ranging from the needs in hospitals and health care establishments, through ambulances and emergency environments to home care environments. Even with all these possibilities, the healthcare area still does not make full use of IoT technologies [13].

In intensive care unit environments, for example, patients are routinely monitored by hospital equipment that usually displays data on parametric monitors, located within sight of healthcare professionals. Its main function, arguably, is to

display the values obtained in real time, as well as to provide audible and visual alerts when values leave the normal range. In general, the generated historical data is stored in the equipment for a few hours, just so that it is possible to observe the average, maximum and minimum values for the period. However, the definitive record is made by the nursing team: periodically a professional must go near the monitor to write down the information presented in the patient's medical record [14].

As for ambulances, operating protocols may vary according to the management and purpose of the displacement. A common case of use of ambulances is when a patient is transferred in a state of urgency or emergency to a specialized hospital. When an emergency call is received requesting, for example, an ambulance to go to the home of a patient in home care, the ambulance environment (equipment and professionals) is quickly prepared based on the health status information that was provided when the call was made. If the patient (before the arrival of the ambulance) experiences any sudden change of state (for example, cardiac arrest), it will be necessary for the ambulance team to be notified by telephone, while still on the outward journey. In a perfect setting [15], during the COVID-19 pandemic, it became necessary to simultaneously monitor millions of infected people, to monitor their symptoms and carry out, if necessary and at the right time, their hospitalization[16]. In this scenario, there is a great benefit in using IoT devices to collect and analyze patient health data in real time, which is the possibility of predicting the capacity of sectors and scheduling internal patient transfers. Not only that, but there could also be an optimization in the process of external transfers, as the team could, through real-time patient data, choose the ideal moment to perform each transfer.

These factors indicate that the current way of using health sensing data can be improved, with the objective of generating benefits for both the patient and for the professionals and health services involved. These benefits are related to the agility in identifying health problems, measurable by sensors, the quick notification of health caregivers and the possibility of obtaining prognoses about the patient's health.

The contribution of this paper is to propose an IoT-based monitoring framework that can help the health caregivers to obtain useful information during the current pandemic of COVID-19. Moreover, a prototype system is developed and evaluated.

The rest of this paper is organized as follows: Section II discusses related works. Section III describes the proposed IoT-based e-health framework. Section IV elaborates the prototype simulated implementation. Section V discusses the experiment. In Section VI, we discuss our findings. A conclusion if this work in conducted in Section VII. Finally, Future work is proposed in Section VIII.

II. Related Work

This section presents a set of related works that were considered for the development of this work. In [17], a review is made of the main potentialities of IoT as a mechanism to mitigate the impact of the pandemic caused by COVID-19. Likewise, a set of applications that have been developed in different countries for the identification and control of patients with COVID-19 are presented, several of which have been articulated to the health system. In the same way, the authors highlight the possibilities of IoT to make remote and autonomous monitoring of variables such as heart rate, blood pressure and blood glucose by patients. In the same way, as challenges of this type of technology, the security of the data and the interoperability of the different devices stand out.

In [18], the authors propose an IoT system for heart rate monitoring in usability tests. The system proposed by the authors obtains the heart rate and heart rate variation data from a Bluetooth belt and sends them to a desktop application, which receives the data through a Bluetooth adapter and stores them in a database at the time that displays them graphically as a function of time. From the data captured, the system determines the level of mental stress of the user at different moments of the test.

In [19], a system for the detection of heart conditions and the identification of mental stress is proposed using the free Arduino hardware platform and a heart rate sensor compatible with it. From the data captured by the sensor, the system allows the graph of the heart rate to be displayed in real time as a function of time, as well as the possible level at which the obtained heart rate is classified (bradycardia, normal, tachycardia). The system presented by the authors does not allow the storage of the data and therefore neither the analysis of the history of the captured data.

In [20], the authors present an IoT system for the self-diagnosis of heart diseases using a probabilistic method for the study of cardiac dynamics. The proposed system is framed in the service-oriented architecture and consists of a bluetooth LE heart rate sensor, which captures the data and sends it in real time to a mobile application, which is responsible for visualizing the data and calculating the probability. Heart rate takes a previously defined critical value. Although the proposed system includes the capture, storage and analysis layers, it does not present the behavior of the heart rate in real time at a graphical level, nor does it take into consideration the level of oxygen saturation as a study variable.

Among other applications of IoT for health care, the proposal of [21] stands out who propose a monitoring system for chronic obstructive pulmonary disease-COPD through constant measurement of body temperature, oxygen saturation and heart rate with inexpensive sensors to issue early prevention alerts. For their part, [22] present a literature review oriented to internet of things applications for home health care using IoT technology, known as IoT Health.

All the works presented in this section show IoT Health as a promising alternative for the health industry by allowing the personalization of the health service with lower labor and operational costs and facilitating the early warning of health alterations in people.

III. Proposed Framework

In this section, a detailed description for the proposed framework is given. A generic diagram of the framework is shown in Fig. 1. It aims to provide support for healthcare

specialist decision-maker. Additionally, it works as a component capable of generating useful information about patients through IoT technology. The proposed framework can receive data collected from sensors, process and generate relevant information and notifications in a timely manner. The proposed framework was designed for COVID-19 patients.

A. IoT Data Collection Module

An important part of data analysis is collecting patient data. IoT data collection module considered as the bridge between wearable sensors and the proposed framework. This module is built based on the IoT infrastructure for health applications. In this experiment, the Mysignals platform is used. MySignals is a hardware development platform built for e-health system [23]. Fig. 2 shows MySignals device and its different types of connected IoT sensors. The information collected by MySignals is useful for the analysis of COVID-19 patients.

The following information can be collected through MySignals:

- Body position - important to know which side the patient is on or has been in the same position for many hours (warning to prevent pressure ulcers from forming). In the case of patients with COVID-19, this data may indicate whether the patient is walking or lying down, in addition to indicating the patient's level of discomfort and stress.
- Body temperature: the body temperature data is collected more often for the caregivers. This data can also be used by crossing data from the external environment, since temperature and other external factors can also affect the patient.
- Electromyography: This sensor detects abnormal muscle electrical activity. Although not directly related to the symptoms of COVID-19, the caregivers can use it as a crossover to indicate other comorbidities that may weaken the patient in general.
- Airflow: to check the number of breaths and identify the patient's discomforts for breathing.
- Galvanic skin response: can be used to check patient relaxation and stress levels.
- Blood pressure: check hypertension and hypotension. It is also not directly related to COVID-19 but can be used to assess other comorbidities and patient weakness.
- Glucometer: check changes in normal blood glucose levels. It is also not directly related to COVID-19 but can be used to assess other comorbidities and patient weakness.
- Pulse Oximeter: in intubated patients, it is important to verify the efficiency of mechanical ventilation. For patients being monitored before the intensive care unit environments, it is essential to indicate lung involvement by the virus.

B. Communication Module

The purpose of the communication layer is to connect different types of sensors devices with the rest modules. It is mainly the communication network that receives the data and transmits it through 5G networks. This module is also composed of microservices to receive, transform to a standard format, and then persist and transmit the data to other modules.

C. Processing Module

The processing Module is directly linked to the Data collection module and communication module. The objective is that this module is to manage the received information; pre-process it links it with the notification module. as shown in Fig. 1, this module can identify an immediate data analysis flow, right after the pre-processing, reducing the waiting time for certain notifications, as well as the complete flow with data processing to generate new information. The decision module would use a knowledge base to interpret what values need to generate notifications. In this case, the decision module required specific information from the patient, to carry out the correct analysis as a knowledge base.

D. AI Module

This module is considered as a general AI component. It may deploy different AI techniques based on the purpose of the given medical application; thus, this module can be considered also as an expert system.

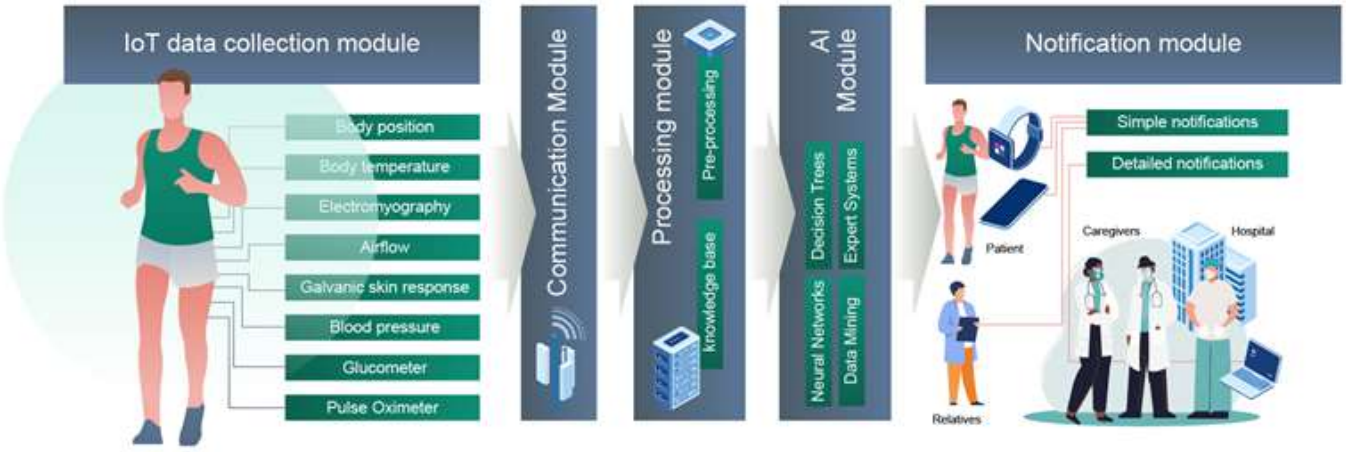


Fig. 1. Conceptional Diagram of the Proposed Framework.

Fig. 2. MySignals Device and its different IoT Connected Sensors [23].

There are many algorithms that use AI techniques, which help in the automatic grouping and classification of data. As an example of these AI techniques, it is possible to mention:

- Neural Networks: are specific Data Mining techniques that aim to learn through examples and try to apply the same rules learned to new unclassified data.
- Data Mining: are techniques to explore data in search of patterns. The goal is to discover hidden patterns, such as time sequences or association rules between data, to classify new data.
- Decision Trees: is a technique for creating a sequence of rules that determine results through a series of logical tests.
- Expert Systems: are software that simulates a professional's decisions expert in the area, through a combination of rules or heuristics that analyze information, just like a human being would.

E. Notification Module

The notification module represents an important part of the proposed framework. In this module, each patient will receive notifications in two distinct categories: simple notifications and detailed notifications. These notifications are used by the caregivers, and can be displayed on a mobile application, monitoring screen or SMS messages sent to the patient relatives for immediate action. As mentioned before, two categories of notifications can be issued by the notification module. These categories are as follows:

- Simple notifications: This category is related to events that occur by direct analysis of data received from a sensor, by searching the knowledge base. For example, a notification could be triggered by a heart rate outside the pre-established normality or a low percentage of oxygen when analyzing the glucometer. These events trigger alerts quickly.
- Detailed notifications: This category is that of notifications triggered after analyzing the data received from various sensors, using data crossing and analysis of historical evolution. The information obtained is checked against the knowledge base to generate definitive results. For example, a patient could be categorized with their degree of evolution (stable, improving, getting worse), or an alert could be issued for the possibility of a heart attack in the next few hours.

IV. Prototype Implementation

In order to validate the proposed framework, a system implementation was performed. Java is used as the programming language for the main modules. Fig. 3 shows the main dashboard for individual patient. Here, caregivers can monitor the patient health status data.

The first aspect that can be observed is the information about which patient is currently being viewed, and the option to switch to monitoring another patient. This makes it easy to monitor multiple patients on one device. Since this is a prototype implementation, only three types of sensors have been predefined, BPM (heart beats per minute), Body Temperature and Pressure. The last valid value received by the sensors will be displayed in the corresponding field. The simulated dashboard will receive data which would normally be sent directly by the IoT Data collection module; for example connected to a MySignals sensors. Although not using real sensors, the system will make the simulated data follow the same data flow within the processing module. The reported data aims to send generated linear sequence of data, so that specific situations can be tested, such as by example, a patient with a fever for many hours. Once received, this data may generate a notification to the health caregivers and the patient relatives.

Fig. 3. Prototype Dashboard for the Collected Data.

V. Experimental Results

In order to test the proposed Validation Module, specific rules were created to categorize patients infected with COVID-19 into four profiles, based on disease severity.

To this end, the risk values were averaged, considering the lowest and highest values recorded in sensors of patients who recovered or not from COVID-19. Accordingly, the following profiles were created:

- Normal state (patients who have recovered).
- COVID aggravation stage 1.
- COVID aggravation stage 2.
- COVID aggravation stage 3.

After configuring the system and adding specific rules, the tests performed, using the module's simulated data linear sequence generation algorithm, it was possible to see that the simulated system correctly categorized the patients tested, according to the severity of the disease, indicated by the average values of the sensors.

Within a hospital environment there are different profiles of patients, with different diagnoses and being followed up in different ways. To reduce the scope of this work and simplify the evaluation of the results, the group of patients infected with COVID-19 and waiting (at home or in a hospital) for natural recovery or for the evolution of symptoms to proceed with hospitalization in a specific sector was chosen of the hospital environment.

The rules for generating predictions (which will be stored in the knowledge base) are not the direct target of this work, since this information would depend on the analysis of a specialist in the health area. Some rules are simple and even easily discovered, including in the manuals of health monitoring devices [24]. For example, the minimum glucose rate setting (which normally triggers an alert) could be set to 70 mg/dl and the maximum normal temperature could be 36.7°C. Other measures, however, need constant adjustments by health specialists.

VI. Discussion

In this paper, we proposed an IoT-based e-Health framework for COVID-19 patients monitoring, which is proved to be well-established as a useful and safe approach. Its use in for COVID-19 cases of risk allows adequate medical control, detects in advance the deteriorating of the disease, supports the care in times of high demand, helps to maintaining social isolation by avoiding calls to the emergency room and helps the patient and his relatives.

Although many previous works have been proposed in the control of acute infectious diseases, most of the existing evidence comes from the management of patients with chronic diseases. The evidence available on the role those new technologies such as IoT and advanced e-Health systems can play in controlling an epidemic is scarce.

For mitigating COVID-19, there are different studies (please see Section II) that provide guidance on the working method and the way to implement it. However, we believe that

proposing a lightweight and effective framework for monitoring COVID-19 patients is beneficial. It is definitely a solid basis for future research, more extensive and with control groups, to define the role that IoT technologies have to play a major role in current and future pandemics.

The approach that combines IoT and proactive healthcare follow-up systems has been well accepted by the patients. Probably a main reason is that this monitoring strategy gives the user a security in days of uncertainty. The proactive attitude, the response to alerts or calls from patients in acceptable times gives a feeling of vigilance and control. We understand that, at least in part, this justifies that the connecting COVID-19 patients with health care givers is essential.

When assessing this work, a number of limitations must be considered. Firstly, the data used is simulated from small size of data and for a single hospital center as a use case. Since the findings may not be identical in areas with a different incidence of COVID-19, it is important to reproduce this study in other health areas.

In summary, this work suggests that IoT monitoring framework, used effectively, enabling useful and safe follow-up for high-risk COVID-19 patients, although steady at the time of diagnosis. More researches are required to validate these results and assess their possible application in potential pandemics.

VII. Conclusion

New technologies for patient monitoring are very welcome and with the COVID-19 pandemic, they have become even more necessary. This work proposed a COVID-19 IoT-based framework that receives and analyzes patient data through a IoT sensors. An analysis of the proposed framework was also carried out, as well as implementing a prototype system.

As a result of this work, the proposed framework that works on an IoT-based sensor infrastructure is developed to provide real-time health information. The proposed framework is interconnectable with other systems, which process data and maintain a patient history. With the use of this framework, a general improvement in the quality of patient follow-up by the healthcare team is expected. This work was developed considering the monitoring exclusively of patients infected with COVID-19, but it is necessary to investigate whether it can be applicable to other diseases or health conditions.

VIII. Future Work

As for future work, it would be interesting to implement an application layer that can provide display data for medical caregivers, nurses and especially first responders. A possible hospital module could present information about each patient, including their current and past status, notification history and possible prognosis. A possible ambulance module could indicate this information and also all monitoring data in real time during the process of transferring the patient from home to the hospital, its alerts and prognoses. This could improve the efficiency of the entire process, as rescuers could anticipate events. Moreover, it would be essential to compare the efficiency of the proposed framework with relevant literature through the application of AI algorithms for detecting and monitoring COVID-10 cases.

References

[1] M. Bansal, A. Goyal, and A. Choudhary, "Industrial Internet of Things (IIoT): A Vivid Perspective," Inventive Systems and Control, pp. 939-949: Springer, 2021.

[2] M. Lombardi, F. Pascale, and D. Santaniello, “Internet of Things: A General Overview between Architectures, Protocols and Applications,” Information, vol. 12, no. 2, pp. 87, 2021.

[3] G. Uganya, Radhika, and N. Vijayaraj, “A survey on internet of things: Applications, recent issues, attacks, and security mechanisms,” Journal of Circuits, Systems Computers, vol. 30, no. 05, pp. 2130006, 2021.

[4] R. Ande, B. Adebisi, M. Hammoudeh, and J. Saleem, “Internet of Things: Evolution and technologies from a security perspective,” Sustainable Cities Society, vol. 54, pp. 101728, 2020.

[5] A. Alamri, A. Gumaei, M. Al-Rakhami, M. M. Hassan, M. Alhussein, and G. Fortino, “An effective bio-signal-based driver behavior monitoring system using a generalized deep learning approach,” IEEE Access, vol. 8, pp. 135037-135049, 2020.

[6] M. S. Al-Rakhami, A. Gumaei, M. Altaf, M. M. Hassan, B. F. Alkhamees, K. Muhammad, and G. Fortino, “FallDeF5: A Fall Detection Framework Using 5G-based Deep Gated Recurrent Unit Networks,” IEEE Access, vol. 9, pp. 94299-94308, 2021.

[7] J. Delsing, J. Eliasson, J. van Deventer, H. Derhamy, and P. Varga, "Enabling IoT automation using local clouds." pp. 502-507, 2016.

[8] H. Magsi, A. H. Sodhro, N. Zahid, S. Pirbhulal, L. Wang, and M. S. Al-Rakhami, “A Novel Adaptive Battery-Aware Algorithm for Data Transmission in IoT-Based Healthcare Applications,” Electronics, vol. 10, no. 4, pp. 367, 2021.

[9] P. Valsalan, T. A. B. Baomar, and A. H. O. Baabood, “IoT based health monitoring system,” Journal of critical reviews, vol. 7, no. 4, pp. 739-743, 2020.

[10] Y. Shi, G. Wang, X.-p. Cai, J.-w. Deng, L. Zheng, H.-h. Zhu, M. Zheng, B. Yang, and Z. Chen, “An overview of COVID-19,” Journal of Zhejiang University-SCIENCE B, vol. 21, no. 5, pp. 343-360, 2020.

[11] A. Gumaei, M. Al-Rakhami, M. M. Al Rahhal, F. R. H. Albogamy, E. Al Maghayreh, and H. AlSalman, “Prediction of COVID-19 confirmed cases using gradient boosting regression method,” Computers, Materials Continua, vol. 66, no. 1, 2021.

[12] M. Otoom, N. Otoum, M. A. Alzubaidi, Y. Etoom, and R. Banihani, “An IoT-based framework for early identification and monitoring of COVID-19 cases,” Biomedical signal processing control, vol. 62, pp. 102149, 2020.

[13] X. Yang, X. Wang, X. Li, D. Gu, C. Liang, K. Li, G. Zhang, and J. Zhong, “Exploring emerging IoT technologies in smart health research: a knowledge graph analysis,” BMC Medical Informatics Decision Making, vol. 20, no. 1, pp. 1-12, 2020.

[14] G. B. Rehm, S. H. Woo, X. L. Chen, B. T. Kuhn, I. Cortes-Puch, N. R. Anderson, J. Y. Adams, and C.-N. Chuah, “Leveraging IoTs and machine learning for patient diagnosis and ventilation management in the intensive care unit,” IEEE Pervasive Computing, vol. 19, no. 3, pp. 68-78, 2020.

[15] O. Udawant, N. Thombare, D. Chauhan, A. Hadke, and D. Waghole, "Smart ambulance system using IoT." pp. 171-176, 2017.

[16] J. Lu, R. Jin, E. Song, M. Alrashoud, K. N. Al-Mutib, and M. S. Al-Rakhami, “An Explainable System for Diagnosis and Prognosis of COVID-19,” IEEE Internet of Things Journal, 2020.

[17] R. P. Singh, M. Javaid, A. Haleem, and R. Suman, “Internet of things (IoT) applications to fight against COVID-19 pandemic,” Diabetes Metabolic Syndrome: Clinical Research Review, vol. 14, no. 4, pp. 521-524, 2020.

[18] P. L. Penmatsa, and D. R. K. Reddy, "Smart detection and transmission of abnormalities in ecg via bluetooth." pp. 41-44, 2016.

[19] J. Van Zaen, O. Chételat, M. Lemay, E. M. Calvo, and R. Delgado-Gonzalo, “Classification of cardiac arrhythmias from single lead ECG with a convolutional recurrent neural network,” arXiv preprint arXiv:.01513, 2019.

[20] C. Gutiérrez-Ardila, C. Montenegro-Marin, and P. Gaona-García, "IOT System for Self-diagnosis of Heart Diseases Using Mathematical Evaluation of Cardiac Dynamics Based on Probability Theory," Information Systems Technologies to Support Learning: Proceedings of EMENA-ISTL, vol. 111, pp. 433, 2018.

[21] M. A. Ponce-Gallegos, G. Pérez-Rubio, E. Ambrocio-Ortiz, N. Partida-Zavala, R. Hernández-Zenteno, F. Flores-Trujillo, L. García-Gómez, A. Hernández-Pérez, A. Ramírez-Venegas, and R. Falfán-Valencia, "Genetic variants in IL17A and serum levels of IL-17A are associated with COPD related to tobacco smoking and biomass burning," Scientific reports, vol. 10, no. 1, pp. 1-11, 2020.

[22] H. Ahmadi, G. Arji, L. Shahmoradi, R. Safdari, M. Nilashi, and M. Alizadeh, "The application of internet of things in healthcare: a systematic literature review and classification," Universal Access in the Information Society, vol. 18, no. 4, pp. 837-869, 2019.

[23] X. Hu, A. M. Abdulghani, M. Imran, and Q. H. Abbasi, "Internet of Things (IoT) for Healthcare Application: Wearable Sleep Body Position Monitoring System Using IoT Platform." pp. 76-81, 2020.

[24] D. Boiroux, A. K. Duun-Henriksen, S. Schmidt, K. Nørgaard, S. Madsbad, N. K. Poulsen, H. Madsen, and J. B. Jørgensen, "Overnight glucose control in people with type 1 diabetes," Biomedical Signal Processing Control, vol. 39, pp. 503-512, 2018.

Brown Spot Disease Severity Level Detection using Binary-RGB Image Masking

N.S.A.M Taujuddin[1], N. H. N. A Halim[2]
Z.H Husin[5], A.R.A Ghani[6], Tara Othman Qadir[7]
Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia (UTHM), 86400 Parit Raja, Johor, Malaysia

M.Siti Norsuha[3], R. Koogeethavani[4]
Paddy and Rice Research Centre, Malaysian Agriculture Research and Development Institute (MARDI) Seberang Perai, 13200 Kepala Batas
Pulau Pinang, Malaysia.

***Abstract*—Agriculture is known as one of the main factor for a growth of a country. Paddy plantation is the most widely planted crop in Malaysia. The rice produced is the main food source to Malaysian's people and source of income to this country as well. However, a disease known as Brown Spot (BS) attacks the paddy plants and threats their quality. This disease caused by bipolar fungus, which represent by the development of an oval, dark brown to purplish-brown spot on leaf. This disease observed as among the hazardous disease that may result in degradation of paddy production. Brown Spot disease could spread through airborne spores from plant to plant on the field. In this research, a system that could help people, especially farmers, to detect the disease at early stage is developed. The real image capture at paddy field is processed in the MATLAB software with image enhancement, background removal as well as binary and RGB image masking process. To determine the Brown Spot area, pixel intensity between the infected and non-infected areas is calculated. The severity level table developed by Horsfall and Heuberger is then used as reference to classify the severity level of Brown Spot disease. A GUI is created to detect the Brown Spot disease automatically. From the study conducted, the accuracy of Brown Spot detection is approximately 89% accurate compared to manual evaluation by plant pathology.**

Keywords*—Brown spot; image enhancement; binary image; RGB image; masking process

I. Introduction

Over many years, Malaysia has grown-up its economy through agriculture sector. Agriculture has a significant economic contribution [1], with paddy being the third most widely cultivated crop. Rice produced by paddy, is the main source Malaysian and gain profit to the country [2]. Thus, a lot of efforts have been done to ensure a high quality and quantity production of this crop [3].

Farmers need to take a good care of their paddy growth in order to ensure the production of a good rice quality. However, paddy is very vulnerable as it often threatened by various pest and disease such as leaf folders, stem borer, plant hopper, Brown Spot (BS) disease [4], Bacterial Leaf Blight (BLB) diseases and Leaf Blast (LB) diseases [5].

Brown Spot (BS) disease (see Fig. 1) is caused by the bipolar fungus, making development of an oval, dark brown to purplish-brown spots on leaf. Brown Spot is considered as one of the most severe conditions [6] of paddy plants and can affect paddy leaves by as much as 50 to 90 per cent of the yield product. Brown Spot can quickly spread by airborne spores from plant to plant in the field. It occurs when the contaminated seeds are sown at a prevailing low temperature of 18-22˚C [7] [8].

Brown Spot disease is a fungal disease that affects the development of rice plants, killing young seeds and reduces their quality. This disease may manifest itself at any stage of crop development [9] but is most severe at full crop maturation. Brown Spot disease is easily identifiable by its characteristic oval to circular form and size of a sesame seed on the surface of the paddy leaf. Typically, the colour is yellow-brown with dark brown patches [10].

Fig. 1. Brown Spot Disease on Paddy Plant.

II. Literature Review

Plant diseases provide a broad scientific field of study in agriculture and emphasize the biological features of diseases. Today, detecting plant diseases is challenging and requires particular care. Fungal infections, bacteria, viruses, and nematodes produce spots on leaves or stems, brown or black lesions, death of lower leaves, yellowing of lower leaves, and black specks [11]. Each disease has its technique of prevention. Standard techniques used include cultural tradition [12], disease resistance cultivars, and chemical usage.

Previously, researcher in [13-14] has developed a system that can detect disease on leaf automatically by using image processing techniques. This, in turn, it allows farmers to recognize the diseases at an early stage and provides valuable knowledge to monitor the crop condition.

Image acquisition, image pre-processing, image segmentation, feature extraction, and classification are the image processing stages required for Brown Spot disease diagnosis used in [15-17]. These processes are done on the captured image of infected plants. The diagnosis of plant diseases is usually based on the presence of different colors [18], shapes [19] and abnormalities on plants leaves [20].

The present technique on detecting plant diseases is via professional observation using their bare eyes [21]. To do so, a big team of specialists is required, which is very expensive especially when the number of farms is enormous. At the same time, the farmers do not have enough facilities [22] to do so. In addition, because of this, the expense of consulting experts is often high and time consuming.

In [23], the researcher discusses the detection of two diseases known as Leaf Spot and Leaf Blotch. The procedure used is by divide the process into several stages using image processing techniques. The first step was to convert RGB image to HSI colour space, which retains just the hue component for further processing. The disease part on the leaf is then extracted using K-Means segmentation. After that, the resulting image is analysed for feature extraction using GLCM texture analysis. The output is then used to train and classify a multi-class SVM classifier.

Colour is always the most important element in image processing [24] and serves as an important indication for class identification. Digital image processing produces objective colour measures that are very useful for early lesion detection. The pixel in a colour image is often represented in RGB space, where the colour of each pixel is defined as a triplet (R, G, B), where R, G, and B correspond to Red, Green, and Blue, respectively [25]. The colours in the defect area of the images were analysed to determine the stage of the disease. At this stage, diseases can be categorised into their level according to their condition, from early to the worst stages [2] [26].

Based on these previous researches, it can be seen that the image processing has widely being used in detecting the paddy diseases. Hence, this study is conducted to propose another enhanced technique named Binary-RGB image masking technique to detect Brown Spot disease.

III. Methodology

The Brown Spot (BS) disease detection start with image acquisition, followed by image enhancement using histogram equalization, background removal, image masking, obtaining the pixel value of segmented RGB image and masked image and determination of Brown Spot Disease based on Disease Severity Scale developed by Horsfall and Heuberger [27] (refer Fig. 2).

The process of the analysis begins with taking pictures of paddy plants that are infected by Brown Spot disease at paddy plot in Malaysian Agricultural Research and Development Institute (MARDI) Seberang Perai using a 24 Megapixel camera with an Optical Image Stabilizer (OIS). The process begins with taking pictures of paddy leaves with a distance of approximately 30 cm.

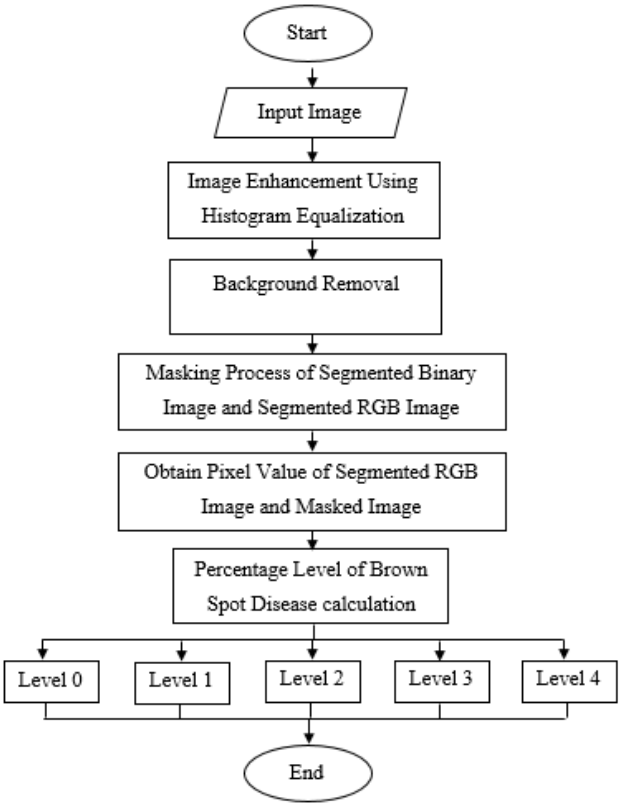


Fig. 2. Flowchart of the Brown Spot Detection Technique.

The quality of the images is then enhanced by using histogram equalization technique. It followed with background removal process to eliminate the similar colour characteristics of the image background with the disease that infected the leaves (see Fig. 3).

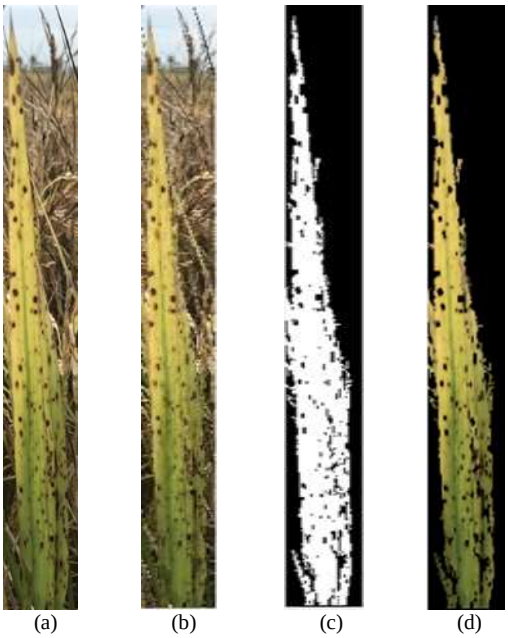


Fig. 3. (a) Original Image (b) Enhanced Image by using Histogram Equalization (c) Binary Image for Background Removal (d) RGB Image.

Then, the Brown Spot disease on a leaf is detected by using a masking concept. The procedure starts with masking the segmented RGB image with a segmented binary image. The resulted masked image will show the presence of Brown Spot on the image sample.

As can be seen in Fig. 3(d), the background of an image is successfully removed, leaving the Region of Interest (ROI) with Brown Spot spotted on the leaf. The Brown Spot disease spot are detected after a masking process is applied. To get the pixel of Brown Spot disease, all pixel except black pixel value from a segmented RGB image are then subtracted with the non-black pixel value form by masked image. The resulting value shows the affected area of the Brown Spot disease (see Fig. 4).

This step is continued by obtaining the pixel value for the non-black pixel of the segmented RGB image and masked image. After that, the Brown Spot severity level on the paddy leaf is determined by calculating the lesion and leaf area ratio. The equations used to express this process are as below:

$$S = \frac{Ad}{A1} \quad (1)$$

or

$$S = Pd/P1 \quad (2)$$

Where;

S is severity extent,

Ad is diseases leaf area,

Al is total leaf area,

Pd is total pixel in diseased area,

Pl is total pixel of leaf.

(a) (b)

Fig. 4. (a) Original Image (b) Masked Image with Brown Spot Disease Spotted on the Leaves.

TABLE I. DISEASE SEVERITY SCALE DEVELOPED BY HORSFALL AND HEUBERGER [15]

Level	Severity
0	Apparently infected
1	0-25% leaf area infected
2	26%-50% leaf area infected
3	51%-75% leaf area infected
4	>75% leaf area infected

The severity level of the infected area is classified according to the Disease Severity Scale developed by Horsfall and Heuberger as shown in Table I.

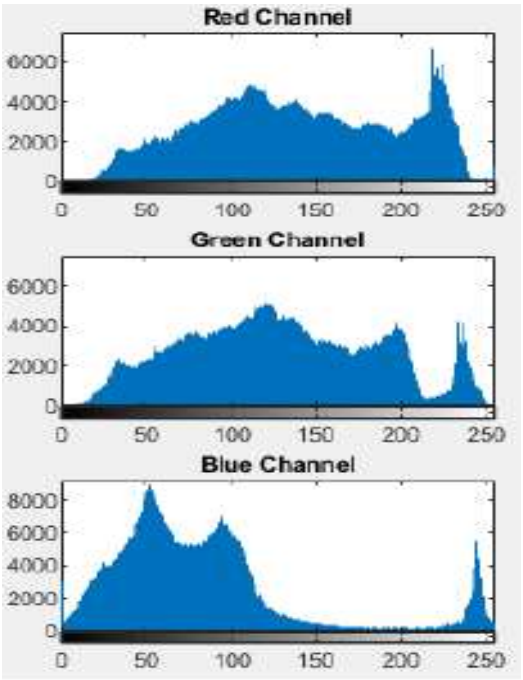
IV. ANALYSIS

To get a high quality images, it is essential to pre-process the collected images. In this study, the Histogram Equalization technique is used to improve the image quality. In this technique, there are two parameters used which are 'Radius' and 'Amount' value. The 'Radius' value is used to control the region's edge pixel dimension. The greater the value, the broader is the region around the edge. While a lower value narrows the region around the edge. Besides, the 'Amount' value function is used to increase the contrast of the sharpened pixels. Therefore, with a larger value of 'Amount', the brightness of the image will increase.

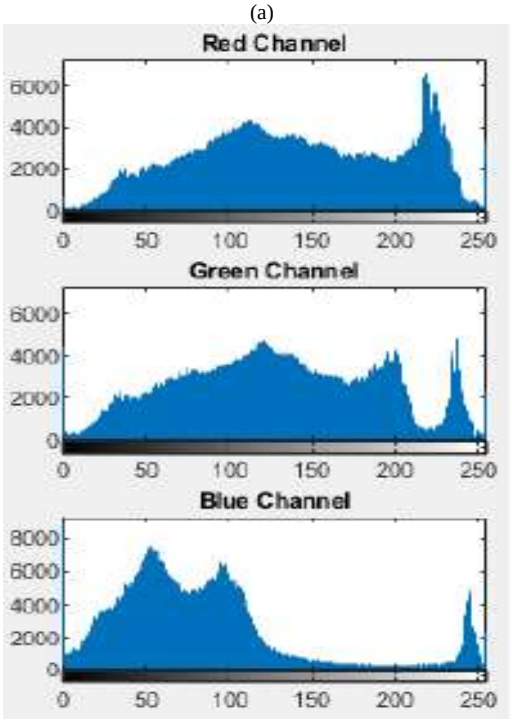
As shown in Fig. 5, there are some changes happen in the histograms of the Red, Green, and Blue channels between the original and enhanced images. The x-axis indicates the tonal scales, while the y-axis indicates the number of pixels in an image. If the pixel's value on the X-axis is close to 0, it indicates that the pixel has a darker black colour. While if its value near to 255, that certain pixel is merely get a lighter colour or white.

By using Histogram Equalization technique the most common intensity pixel values are efficiently spreads out the while stretching the image's intensity range. This technique, in particular, is resulting in an improved image intensity as well as image quality of the Region of Interest (ROI), which in this instance is the paddy leaf.

As for example, the peak value of blue channel is at 50, while the green channel is on range of 0-50, and red channel is on the range of 0 to 250. After the Histogram Equalization process, the pixel intensity value is being stretched out to it closes contrast value. As for the pixel value of blue channel, the colour pixel of 50 changes from 8000 to 7000, where the colour pixel of this channel is stretched out to its near intensity range. For green channel, the colour pixel is spreading towards 0 which indicate the image at the certain pixel will become darker. At the red channel, the colour pixel is spreading more towards the 250 where it's resulting to a brighter pixel. This will increases the image contrast when its data is represented by close contrast values.



(a)



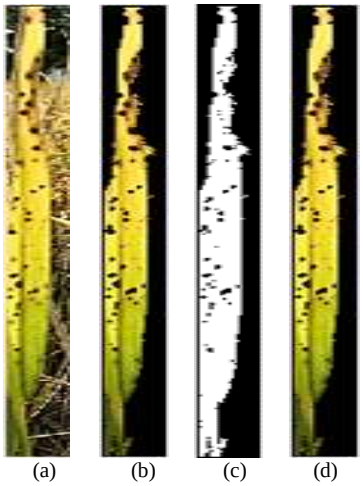
(b)

Fig. 5. (a) Histogram for Original Image (b) Histogram for Enhanced Image.

The image background on enhanced image is then removed and keep only the object of interest, which is the paddy leaves. Then, the affected area on the paddy leaf is detected by subtracting the segmented RGB image with the masked image. The output of two different sample can be seen in Fig. 6 and Fig. 7.

(a) (b) (c) (d)

Fig. 6. (a) Original Image (b) Segmented RGB Image (c) Segmented Binary Image (d) Masked Image.



(a) (b) (c) (d)

Fig. 7. (a) Original Image (b) Segmented RGB Image (c) Segmented Binary Image (d) Masked Image.

In this technique, it is compulsory to get the total pixels of leaf and the total pixel of Brown Spot affected area first. To get the pixel value of the Brown Spot affected area, the non-black pixel is obtained from masked image while for total pixel of leaf is obtained from subtracting the non-black pixel of masked image and the not black pixel of binary image.

In order to obtain the Region of Interest (ROI), the value of the non-black pixel for the segmented RGB image and segmented binary image is calculated. Here, the Horsfall and Heuberger method [15], is applied to calculate the severity level of the Brown Spot disease on the paddy leaf. The equation for Horsfall and Heuberger method is as follows:

$$S = 100\ x\ [\frac{B - K}{B}]$$

Where;

B is the non-black pixel of the segmented RGB image,

K is the non-black pixel of the masked image,

For example, for sample image in Fig. 6, the non-black pixel value of segmented RGB image is 64966, while non-black pixel value of masked image is 16728. So, the calculated affected area is;

Affected area

=100×((B-K)/B)

=100*((64966-43922)/64966)

=32.39%

Based on the Disease Severity Scale, if the percentage obtained is in the range of 26%-50%, the sample image is classified as affected with Brown Spot disease at Level 2. Since the sample image as in Figure 6 is calculated to have an affected percentage of 32.39% so, it fall under Brown Spot severity Level 2.

As for image sample in Fig. 7, the non-black pixel value of segmented RGB image is 47501, while non-black pixel value of masked image is 36981. So, the calculated affected area is;

Affected area

=100×((B-K)/B)

=100*((47501-36981)/47501)

=22.15%

So, based on Disease Severity Scale, this sample fall under Brown Spot severity Level 1.

These image samples are also being forwarded to Plant Pathology for a manual evaluation. The manual observation shows that sample image in Fig. 6 was 30% effected by Brown Spot. While for sample image in Fig. 7, it was manually evaluated as 20% effected with Brown Spot.

Table II shows the result of Brown Spot area percentage by using system calculation and the manual evaluation by plant patology on five collected samples. While Table III shows the Brown Spot severity level by using system calculation and the manual evaluation.

As can be seen in Table II, the result of Brown Spot area percentage by using system calculation and manual evalution are sigltly different but still tolarable. The precision in detecting the Brown Spot disease by using proposed system is about 89%. Although the percentage area of Brown Spot detected on image sample are different, but the severity level by using system calculation and manual evaluation are the same.

To ease the process so that the layman can use the proposed system, a GUI as shown in Fig. 8 is developed. Here, the user need to load the original image to the system. Then, the system will automatically shows the enhanced image, segmented RGB image, segmented binary image and masked image. At the same time, the non-black pixel value of the RGB and masked image, the percentage of disease severity and its severity level is calculated and the result will be appear in the table of severity level box. The reset button is also available to clear all the input data as preparation to receive a new image sample.

TABLE II. BROWN SPOT AREA PERCENTAGE BY USING SYSTEM CALCULATION AND THE MANUAL EVALUATION

Sample	Calculation on System	Manual Evaluation
1	39.48%	35%
2	32.65%	30%
3	57.58%	55%
4	21.30%	20%
5	70.96%	60%

TABLE III. BROWN SPOT SEVERITY LEVEL BY USING SYSTEM CALCULATION AND THE MANUAL EVALUATION

Sample	Calculation on System	Manual Evaluation
1	Level 2	Level 2
2	Level 2	Level 2
3	Level 3	Level 3
4	Level 1	Level 1
5	Level 3	Level 3

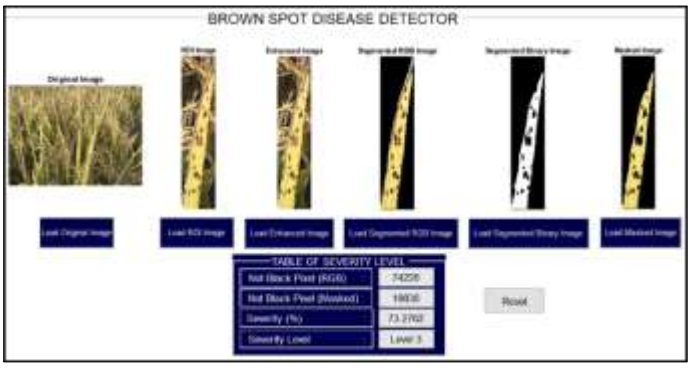


Fig. 8. GUI of the Proposed System.

V. CONCLUSION

This project is intended to help people, especially farmers who work in the agriculture sector. Instead of using too many pesticides that could affect people's health, early detection of the disease should help get rid of the disease before it gets worse and endanger the paddy's quality. Therefore, by using the proposed technique, it will help the farmer to classify the severity level of Brown Spot disease and it may assist the farmers in the early detection of the Brown Spot disease before it spread widely in the paddy field.

ACKNOWLEDGMENT

The authors would like to thank the Research Management Centre (RMC), Universiti Tun Hussein Onn Malaysia (UTHM) for facilitating this research activity under Multi Disciplinary Research (MDR) Grant Vote H485 and Malaysia Agricultural Research and Development Institute (MARDI) Seberang Perai for technical and fieldwork consultation.

REFERENCES

[1] S. C. Omar, A. Shaharudin, and S. A. Tumin, The Status of the Paddy and Rice Industry in Malaysia. Khazanah Research Institute, 2019.

[2] S. Mutalib, M. H. Abdullah, S. Abdul-Rahman, and Z. A.Aziz, "A Brief Study on Paddy Applications with Image Processing and Proposed

Architecture," in 2016 IEEE Conference on Systems, Process and Control (ICSPC 2016), 2016, no. December, pp. 124–129, doi: 10.1109/SPC.2016.7920716.

[3] M. M. M. Najim, T. S. Lee, M. A. Haque, and M. Esham, "Sustainability of rice production: a Malaysian perspective," J. Agric. Sci., vol. 3, pp. 1–12, 2007, doi: 10.4038/jas.v3i1.8138.

[4] B. S. Anami, N. N. Malvade, and S. Palaiah, "Classification of yield affecting biotic and abiotic paddy crop stresses using field images," Inf. Process. Agric., vol. 7, no. 2, pp. 272–285, 2020.

[5] P. A. C. Ooi, "Common insect pests of rice and their natural biological control," Agric. Sci. J., vol. 1, no. 1, pp. 49–59, 2015.

[6] R. P. Narmadha, "Detection and Measurement of Paddy Leaf Disease Symptoms using Image Processing," pp. 26–29, 2017.

[7] L. Aryal, G. Bhattarai, A. Subedi, M. Subedi, B. Subedi, and G. K. Shah, "Response of Rice Varieties to Brown Spot Disease of Rice at Paklihawa , Rupandehi," Glob. Insititute Res. Educ., vol. 5, no. 2, pp. 50–54, 2016.

[8] Z. Mohamed, R. Terano, M. N. Shamsudin, and I. A. Latif, "Paddy Farmers' Sustainability Practices in Granary Areas in Malaysia," Resources, vol. 5, pp. 1–11, 2016, doi: 10.3390/resources5020017.

[9] A. Sharma, D. Satish, and S. Sharma, "Indian major basmati paddy seed varieties images dataset," Data Br., vol. 33, p. 106460, 2020.

[10] I. Ihsan, E. W. Hidayat, and A. Rahmatulloh, "Identification of Bacterial Leaf Blight and Brown Spot Disease In Rice Plants With Image Processing Approach," J. Ilm. Tek. Elektro Komput. dan Inform., vol. 5, no. 2, pp. 59–67, 2019, doi: 10.26555/jiteki.v5i2.14136.

[11] N. S. A. M. Taujuddin et al., "Detection of Plant Disease on Leaves using Blobs Detection and Statistical Analysis," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 8, pp. 407–411, 2020.

[12] R. G. De Luna, E. P. Dadios, and A. A. Bandala, "Automated Image Capturing System for Deep Learning-based Tomato Plant Leaf Disease Detection and Recognition," in IEEE Region 10 Conference, Proceedings of TENCON, 2018, vol. 2018-Octob, no. October, pp. 1414–1419, doi: 10.1109/TENCON.2018.8650088.

[13] R. S. Indumathi, N.Saagari, V.Thejuswini, "Leaf Disease Detection And Fertilizer Suggestion," in Proceeding of International Conference on Systems Computation Automation and Networking, 2019, pp. 1–7.

[14] S. Kn, M. Suresha, and H. N. T, "A Novel Segmentation and Identification of Diseases in Paddy Leaves Using Color Image Fusion Technique," pp. 17–22, 2021.

[15] S. Ramesh and D. Vydeki, "Recognition and classification of paddy leaf diseases using Optimized Deep Neural network with Jaya algorithm," Inf. Process. Agric., vol. 7, no. 2, pp. 249–260, 2020, doi: 10.1016/j.inpa.2019.09.002.

[16] M. V. Overbeek, "Identification of Maize Leaf Diseases Cause by Fungus with Digital Image Processing (Case Study : Bismarak Village Kupang District – East Nusa Tenggara)," in 5th International Conference on New Media Studies, 2019, pp. 125–128.

[17] S. Ramesh and D. Vydeki, "Application of machine learning in detection of blast disease in South Indian rice crops," J. Phytol., vol. 11, pp. 31–37, 2019.

[18] G. Dhingra, V. Kumar, and H. D. Joshi, "Study of digital image processing techniques for leaf disease detection and classification," Multimed. Tools Appl., pp. 19951–20000, 2018.

[19] M. M. Tin, M. M. Khin, S. S. Hlaing, P. P. Wai, and K. L. Mon, "Leaves Disease and Damage Rate Classification based on Features," pp. 419–420, 2021.

[20] T. Islam, M. Sah, S. Baral, and R. Roychoudhury, "A Faster Technique on Rice Disease Detectionusing Image Processing of Affected Area in Agro-Field," Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2018, no. Icicct, pp. 62–66, 2018, doi: 10.1109/ICICCT.2018.8473322.

[21] N. N. Kurniawati, S. N. H. S. Abdullah, S. Abdullah, and S. Abdullah, "Texture analysis for diagnosing paddy disease," Proc. 2009 Int. Conf. Electr. Eng. Informatics, ICEEI 2009, vol. 1, no. August, pp. 23–27, 2009, doi: 10.1109/ICEEI.2009.5254824.

[22] V. Singh, Varsha, and A. K. Misra, "Detection of unhealthy region of plant leaves using image processing and genetic algorithm," in 2015 International Conference on Advances in Computer Engineering and Applications(ICACEA 2015), 2015, pp. 1028–1032, doi: 10.1109/ICACEA.2015.7164858.

[23] G. Kuricheti and P. Supriya, "Computer Vision Based Turmeric Leaf Disease Detection and Classification: A Step to Smart Agriculture," in Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019), 2019, no. Icoei, pp. 545–549, doi: 10.1109/ICOEI.2019.8862706.

[24] N. S. A. M. Taujuddin, R. Ibrahim, and S. Sari, "Progressive Pixel-to-Pixel Evaluation to Obtain the Hard and Smooth Region for Image Compression," in Proceedings - International Conference on Intelligent Systems, Modelling and Simulation, ISMS, 2015, Vol. 2015.

[25] M. V Latte and S. Shidnal, "Multiple Nutrient Deficiency Detection in Paddy Leaf Images using Color and Pattern Analysis," 2016 Int. Conf. Commun. Signal Process, pp. 1247–1250, 2016.

[26] N. N. Kurniawati, S. N. H. Sheikh Abdullah, S. Abdullah, and S. Abdullah, "Investigation on Image Processing Techniques for Diagnosing Paddy Diseases," in 2009 International Conference of Soft Computing and Pattern Recognition, 2009, pp. 272–277, doi: 10.1109/SoCPaR.2009.62.

[27] S. B. Patil and S. K. Bodhe, "Leaf Disease Severity Measurement using Image Processing," Int. J. Eng. Technol., vol. 3, no. 5, pp. 297–301, 2011.

A Novel Feature Extraction for Complementing Authentication in Hand-based Biometric

Mahalakshmi B S[1]
Assistant Professor
Department of Information Science and Engineering
BMS College of Engineering, Bangalore, India

Sheela S V[2]
Professor
Department of Information Science and Engineering
BMS College of Engineering, Bangalore, India

***Abstract*—With an increasing usage of hand-based biometrics in authentication system, there is a need to evolve up with more potential security owing to increasing evolution of threats. The security of the hand authentication system completely depends upon uniqueness and distinct selection of features from hand image which has the properties of robustness, fault tolerance, and simpler implication. Review of existing feature extraction literatures shows more inclination towards sophisticated process as well as it also suffers from various other limitation. Therefore, this manuscript resolves this limitation by presenting a novel model of feature extraction which is carried out in more progressive form and less iterative form, unlike existing approaches. The proposed system achieves its research goal by introducing simplified feature extraction operation via storage, blurring, color space conversion, binary image conversion, the modelling aspect of the study emphasizes on image enhancement along with fuzzification for yielding more efficient result. An experimental study has been carried out using Python considering hand-biometric dataset in order to carry analysis where the outcome shows significant supportability over any palmprint recognition system. The study outcome is compared with most standard implementation of feature extraction to find that proposed system offer better accuracy performance in contrast with existing system.**

***Keywords*—Biometric; security; feature extraction; hand geometric; authentication; palmprint recognition**

I. INTRODUCTION

Authentication is defined as the process of verifying something to be genuine. In our case however, we are trying to verify if the identity of the person is genuine or not. Biometric authentication is a method of authentication where the system focuses on 'who you are' rather than 'what you know' (password) [1]. Biometrics is defined as the biological measurements which will identify an individual [2]. With these studies we can conclude that more unique the data is, more robust the system will be in identifying the individuality. Thus it can be concluded that the system which efficiently identify a person with the data of 'who he is' (bio-metric) rather than what he knows (password) is more secure as passwords are considered to be more insecure compared to biometrics [3]. Even if there are many types of modalities that can be collected for biometric authentication of a person, obtaining the hand geometry require less complex systems of input data acquisition in comparison with other modalities [4]. In the complete mechanism of authentication mechanism, feature extraction plays a vital role that contributes towards dimensional reduction as well as it also offers significant information for precisely carry out the next step of authentication operation. The generalized definition of feature could be stated as a functional information of one or more set of information that when subjected to computation yields at quantifying potential characteristic of an object. Theoretically, a particular form of an image feature can be defined with respect to particular structure within an image that is always feasible to represents in multiple mechanisms. It is to be noted that selection of feature representation can be potential concern while developing a computer vision system. There could be dependencies towards adopting higher level of information in the structure of a feature in order to find a solution to the problem; however, it will require more processing attempts. A complete image or a segment of an image can be represented by the feature vector. This can be accomplished by exploring measurement of the group of features. Usually, a feature vector can be n-dimensional vector which consist of such measurements. Some of the widely used conventional features are Zernike Moments [5], Local orientation histograms [6], local brightness [7], binary object feature [8], etc. However, this will eventually differ from one to another image. From the palmprint image viewpoint, there are various forms of features e.g. wrinkles, textures, ridges, valleys, minute points, pores, ridge width, etc. Conventional approaches [9], [10] mainly make use of line-based methods as elementary representation of palmprint. Studies claims to offer stabilized authentication of palm using principle lines. It should be also noted that a high resolution image of palm print is required in order to perform a detection of features from palm print. The challenges increases more as palm print images are basically of two types viz. two and three dimensional palm images. Two dimensional palm images are further classified into low and high resolution image where low-resolution two dimensional palm print images are further classified into contact-based and contactless forms. With an increasing usage of hand-based biometric system in commercial products and services, it is essential to ensure that a robust modelling be carried out in this regards. Apart from this, with the presence of archives of literatures towards biometric authentication, not much emphasize is offered towards feature extraction. Hence, this loophole in existing system motivates to carry out a novel research work to address the above mentioned issues of existing system.

This paper introduces a novel and simplified mechanism of feature extraction of hand where the emphasis is offered

towards achieving cost effective implementation process unlike existing sophisticated extraction process. The paper presents a simplified model which emphasize on the feature extraction process in order to assist a robust form of biometric-based authentication system. Existing approaches towards biometric authentication system mainly emphasize on using complex form of feature extraction process which could not effectively balance between computational efficiency and accuracy at same time. At the same time, it is also noticed that not much emphasis is offered towards the contextual informative contents with respect to foreground and background. Non-inclusion of this fact will eventually lead to outliers in the authentication system. However, it should be noted that proposed system doesn't develop any form of authentication system; it only introduces a feature extraction mechanism which is meant to be used for any form of hand-based biometric system in future. Hence, this leads to evolution of certain research questions:

- What are the dominant level of features essential for hand-based biometric system to ascertain both accuracy and computational efficiency?
- How to find the artifacts and mitigate them for the fluctuation of illumination state in foreground and background to ensure better removal of outliers?

Without addressing the above two critical research question, developing a feature extraction method will not encapsulate the practical issues pertaining to the authentication system in biometric. The proposed system also emphasize on image enhancement process which offers a superior form of accuracy in contrast to existing scheme of recognition system of hand. The organization of this manuscript is as follows: Section II discusses about the existing approaches of feature extraction followed by discussion of research problem that are confirmed after reviewing existing approaches in Section III. Further, proposed system design is discussed in Section IV followed by discussion of result analysis in Section V and conclusion in Section VI.

II. Related Studies

This section discusses about all the significant research work carried out in recent time towards feature extraction process. It should be known that feature extraction is an intermediate process within recognition of the palm. Hence, this part of the studies discusses about majority of studies that has considered hand recognition system based on palm geometry. While equal emphasis is also given for other image object-based recognition system in order to understand the effectiveness of unique feature extraction approaches.

Existing literatures has consideration of multispectral palm images for recognition process as seen in work of Attallah et al. [11]. The study has emphasized on using spiral feature as well as Linear Binary Pattern for the purpose of feature extraction. Finally, K-nearest neighborhood is used for matching purpose. The outcome is analyzed with multiple dataset to prove its effectiveness. However, the owing to statistical computational adoption, the study doesn't emphasize on possibilities of outliers when color images are taken. Most recent, a unique feature extraction process is discussed by Bakheet and Al-Hamadi [12] where both region and boundary based process of feature extraction is discussed in order to construct a multi-modal descriptor from hand silhouette. Further supervised learning approach is used for training this feature in order to carry of classification. The works carried out by Deshpande et al. [13] have used Harris Corner Detection and Discrete Wavelet Transform in order to carry out feature extract from palm print images. The formulated feature is transformed in binary matric that is subjected for comparison while performing training.

The work of Zhang et al. [14] have used an iterative learning scheme for extracting hidden information in the form of feature.

The work of Yang et al. [15] discusses about the importance of offset feature while performing analysis of feature. The study has used infinite Dirichlet process in order to carry out analysis.

Existing system of recognition also addressed the problems associated with partial occlusion. The study of Liu et al. [16] has used a graph matching mechanism for addressing this problem of recognition system of dorsal hand. The authors have used conventional shape-based feature extraction mechanism for recognition of vein along with edge attributes. Existing literatures has offered a solution towards adaptability problems in palmprint recognition. The work of Zhao et al. [17] has used Least Square Regression as a mechanism for feature extraction. The work carried out by Gupta and Gupta [18] has presented a study of authentication model using multi-biometrics of hand images in the form of hand geometry, dorsal vein of palm, and slap fingerprints. The essential feature in this case are location of fingers which is carried out using segmentation-based process. Nearly similar form of study is also discussed by Izakian et al. [19] where segmentation using trajectories are adopted for carrying out feature extraction. The feature in this study is the alterations of the specific movement profile captured in sliding window. Existing studies has been also reported to make use of Doppler map for carrying out feature extraction. The work carried out by Ryu et al. [20] have used evolutionary approach in order to carry out feature analysis. The study has considered feature specific to radar characteristic along with statistical outcomes to be used in machine learning approach. The feature extraction in this study is carried out using low level descriptor, tracking scattering center, and obtaining connectivity between channels.

Most recently, there are studies which are carried out using machine learning approach towards processing features from hand images. The study carried out by Xin and Wang [21] has discussed about utilization of CNN for assisting in classification of image considering different requirements of feature extraction. Such study is carried out by Du et al. [22] where Convolution Neural Network (CNN) is used for recognition of hand gesture. This study make use of micro-Doppler feature from target areas in order to generate more logical and adaptive feature. CNN was used further to improve upon the recognition system. Similar category of usage of CNN is also seen in work of Abd-Ellah et al. [23] where CNN is used for high-level feature extraction in order to obtain

feature map. Exactly similar approach is also reported in work of Li et al. [24] have carried out the study towards feature extraction using Micro-Doppler associated with recognition of hand gesture. The study has also used orthogonal-based matching map for performing this extraction process where further CNN is used for training.

Adoption of CNN is also reported in work of Ghrabat et al. [25] integrated with search optimization scheme of evolutionary approach. The feature in this study are the color and texture. This is accomplished using co-occurrence matrix of gray level and image intensity for extracting color and texture feature respectively. Further K-means clustering is used for grouping the feature while classification is carried out by Support Vector Machine (SVM) and CNN. Existing mechanism has also witnessed usage of CNN for feature extraction associated with semantic aspect of recognizing vein. This work is reported by Pan et al. [26] where all the feature map are concatenated where the selection of optimal feature is carried out on the basis of semantic weights. A unique adoption of machine learning approach was discussed by Wu et al. [27] towards optimizing the extracted feature. The scheme make use of inferential statistics and neural network in order to carry out recognition. The features used in this study are time-frequency domain, frequency domain, and time domain feature. Yoo et al. [28] have carried out used recurrent neural network in order to study articulation of hand by extracting sequential feature of joints in fingers. Further Long Short Term Memory (LSTM) was reported to be used in study of Yuan et al. [29] for deep feature extraction. The work carried out by Wibisono and Mursanto [30] have also presented a sophisticated feature extraction scheme using deep learning.

Hence, there exists various studies towards feature extraction in existing system. The next section discusses about research problems associated with it.

III. Research Problem

After reviewing the existing literatures towards feature extraction considering hand geometry, various conclusive remarks were obtained associated with its downsides. Following were the research problems that has been identified and addressed in current work:

- Complex Mechanism of Feature Extraction: Existing system are mainly reported to use spiral feature as well as Linear Binary Pattern, Harris Corner Detection and Discrete Wavelet Transform, infinite Dirichlet process, segmentation using trajectories, micro-Doppler feature, etc. All these approaches are quite conventional and sophisticated in its operation with respect to various traits of hand geometry. Owing to inclusion of maximum number of operational steps to carry out feature extraction, these methods may offer poor scalability in its performance.
- More usage of Learning Operation: Machine learning and its different variants were discussed for improving the performance of feature extraction. However, machine learning methods implemented in existing schemes are more focused on achieving accuracy and less on offering granularity in feature extraction from dynamic environment. Moreover dependencies of trained data will further act as an impediment towards feature extraction if images with multiple orientation and modalities are used.
- Biased Adoption of Illumination Factor: Majority of the existing approaches considers the illumination factors over the foreground object and not the background object. The silhouette generated by the hand structure also induce variation in background illumination factor and significant results in outliers (conjoining fingers). Therefore, there is a need of using a simplified scheme which offers equal importance to the illumination factor for extracting precise feature.
- Less Preference to Image Enhancement: Majority of the existing implementation follows a dual steps of operation viz. feature extraction after preprocessing and subjecting the feature for further process of authentication or recognition. However, there are few literatures to highlight further importance for enhancing the image, which is also a challenging step to be carried out. Without this operation, it is a challenging process in order to ensure better visual quality of the outcome image. It also affects the accuracy of the outcome image.
- Common preprocessing steps: Not much literatures has witnessed the use of unique and simplified preprocessing operation. There is a need to preprocessing to eliminate noise in simplified manner, there is a need to consider different form of color space apart from conventional RGB, and also, there is a need to evolve up with a different binarization scheme. At present, there are very studies that has emphasized on such operations while performing feature extraction.

It should be noted that research problem in hand is associated with the feature extraction. All the above mentioned points represents the limiting factors of existing approaches. Out of all this, the first critical problem which requires to be prioritized primarily is associated with preprocessing operation. A better and simplified preprocessing operation will eventually lead to better artifact removal from the biometric input given, which unfortunately lacks in existing approach. This will also contribute towards addressing enhancement issues in an image. The second essential problem is towards dealing with the illumination factor without which proposed model will not be able to deal with different variants of images of hand-based images. The next section discusses about the system design implemented in proposed system as solution.

IV. System Design

The proposed study is focused on feature extraction methodology rather than the recognition model itself. The hypothesis is that if the feature extraction is carried out in efficient manner than it will positively affect the recognition modelling too. The system design of the proposed system is motivated from the work carried out by Afifi et al [31]. A Convolution Neural Network (CNN) system is used to extract

the features and Support Vector Machine (SVM) is used to authenticate the person in existing system. When a CNN is being trained, its performance can be enhanced if right preprocessing steps are added to it [32]. It should also be noted that feature enhancement will improve the performance of the recognition system without making any fundamental changes to the system itself [33]. Though the explicit method for handling the contrast and enhancement can be used if there exists a varying lighting condition, which is the sole purpose of the proposed system. In the proposed model, a novel type of preprocessing system is being proposed which will enhance the performance of the hand geometry recognition and authentication algorithm. As shown in Fig. 2, the proposed system contains five distinct modules viz. i) image blurring, ii) HSV conversion iii) Fuzzy logic binary converter, iv) contour detector v) image enhancer. The purpose of all above modules is to detect the important features in the hand geometry and enhance them so that they are better exposed to next algorithm which will be used for classification/recognition. The block diagram of the system is as shown in the Fig. 1.

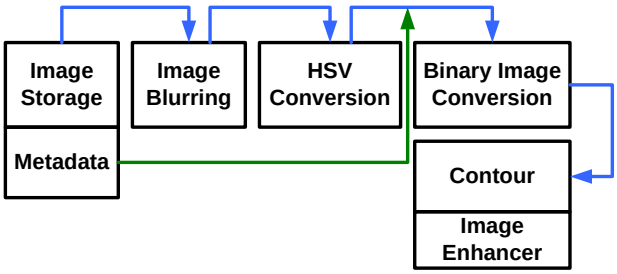


Fig. 1. Proposed Preprocessing System.

The brief outlines of above exhibited blocks of operation are as follows:

A. Image Storage

This block of operation retains images associated with palm geometries that is considered for study implementation. The proposed system uses 11K hand dataset that consists of 11,000 images of hands belonging to 190 people. This also provides additional details about the images in the form of metadata [31]. The metadata are used further prior to the process of binary image conversion. Since in 11K hands dataset, even the metadata is given, It helps in present study to recognize if a person is having any accessory while processing the image. If person is wearing a ring or nail polish, the results may vary hence these taken into account during fuzzy image binarization.

B. Image Blurring

The proposed system implements image blurring in order to eliminate certain features as well as to eliminate noise [34]. Amongst various techniques used for blurring, the proposed system implements Gaussian blurring approach. It is the result of blurring the image with the help of Gaussian function. The one-dimensional Gaussian function is mathematically expressed as follow:

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \quad (1)$$

However, while dealing with images, the formula must be expanded to two dimensions. The two-dimensional mathematical expression of proposed blurring is now as follows:

$$G(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (2)$$

In the above expression (2), the variable σ represents the window size. In proposed study, window size of 3 is selected and the above mathematical expression is applied to every pixel of the image. Due to Gaussian blur, noises in the image are removed and also reflection caused due to shininess of aged skin and blood veins are also removed [35].

C. HSV Conversion

This is the third step of operation in proposed system where color space of Hue, Saturation and Value (HSV) is deployed. The prime justification of using HSV is because it is a standard to represent a color in a more human understandable way. In proposed study, the image is converted into HSV format so that various colors of skin can easily be represented in HSV format rather than regular RGB format. To detect human skin, HSV is more advantageous compared to RGB since it is easy to calculate and ignore ambient light [36]. In order to convert RGB to HSV, the channels must be flipped if the implementation is being done in OpenCV as it used BGR format instead of RGB. In order to do this, all three colors must be normalized using minmax scaler. Formula for same is shown below.

$$Y' = \frac{Y - Y_{min}}{Y_{max} - Y_{min}} \quad (3)$$

However it must also be noted that pixel values always range between 0 to 255 hence Y_{max} is always retained at the value of 255 and Y_{min} value is retained at 0. This will yield to in terms of following mathematical expression.

$$R' = \frac{R}{255}, G' = \frac{G}{255}, B' = \frac{B}{255} \quad (4)$$

Once all three color values are converted, the following algorithm is used to make the conversion. In the algorithm which is being used, the HSV conversion is done with the help of modulus function hence the conversion happens in a much faster way as contrary to using trigonometric functions. First, the algorithm calculates Hue by calculating the angle of color from nearest elementary color (RGB) Saturation is calculated as a measure of percentage difference between most dominant and least dominant color and finally value is nothing but the value of dominant color.

Algorithm : RGB to HSV Conversion

Input : Normalised RGB image (Img)

For pixels in img:

C_{max} = max(R' G' B')

C_{min}= min(R' G' B')

Diff = C_{max} – C_{min}

If C_{max} ==C_{min}:

H = 0

Else if C_{max} == R':

H = ((60 * (G'-B')/Diff) + 360 % 360)

Else if C_{max} == G':

H = ((60 * (B'-R')/Diff) + 120 % 360)

Else if C_{max} == B':

H = ((60 * (R'-G')/Diff) + 240 % 360)

If C_{max} == C_{min}:

S = 0

Else:

S = (Diff/C_{max}) * 100

V = C_{max} * 100

D. Image Binarization

The image must be converted to binary image in order to make it easy for processing. In case of this study, the HSV values are converted to binary values. The complete process of image binarization is shown in Fig. 3.

According to Fig. 2, the proposed system takes an input image and considers all the pixels present in an image for further processing. The next step of operation is to check if the HSV value is more than upper value of the threshold T_{up}. If this condition is satisfied, the proposed system checks of the value of HSV is more than lower threshold T_{low}. In either cases of the condition is not satisfied than proposed system carry out fuzzification followed by a conditional check if the adjacent pixel P_{ad} is white. If the adjacent pixel is found to be white than the proposed model converts them in black or else it makes it white. In the exhibited Fig. 3, the process of fuzzification is the process where probability of the pixel being skin or not skin. In order to do this an algorithm will run in the background and decide the values of Hue so that the algorithm works for various skin tones. The proposed system performs the following steps of implementation towards hand image for eigenvectors.

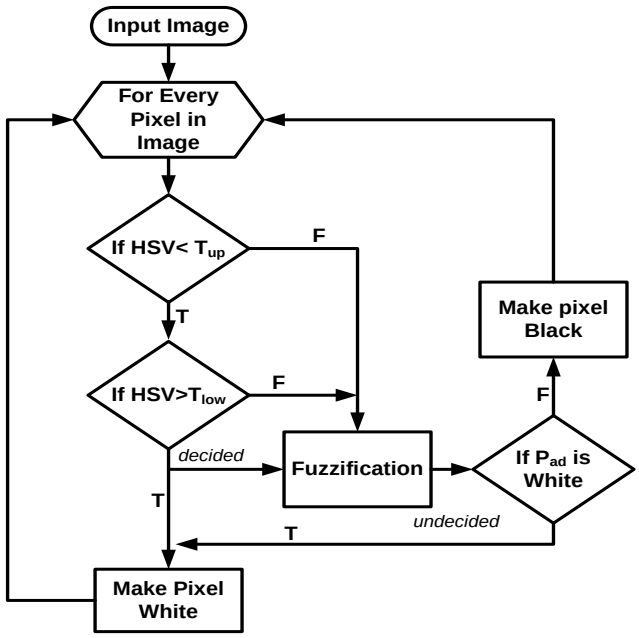


Fig. 2. Process flow of Proposed Image Binarization.

The first steps towards this implementation are to consider a sample of 100 images from the dataset (Fig. 3).

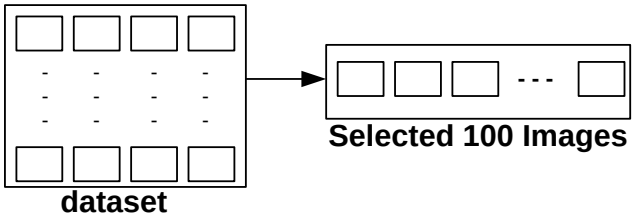


Fig. 3. Selection of Images from Dataset.

The next step of operation is to flatten every images and re-arrange it into matrix T.

Following are the mathematical operations being carried out:

M_1=Image-1 ⊗I

M_2=Image-2 ⊗I

-

-

M_n1=Image-n ⊗I (5)

In the above expression (5), the variable *I* represents identity matrix. Therefore, the resultant matrix T is mathematically expressed as follow,

$T=M_1 \bowtie M_2 \bowtie \ldots\ldots M_n$ (6)

The next step of the operation will be to calculate and subtract mean as follows:

$$T = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ v_{n1} & v_{n2} & \cdots & v_{nm} \end{bmatrix} = A = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_n \end{bmatrix} \quad (7)$$

$$\mu_i = \frac{1}{m}\sum_{j=1}^{m} T_{ij} \quad (8)$$

$$m = T - A \quad (9)$$

Subtract every column of S with A. The next step of operation will be to compute the covariance matrix S as follows:

$$S = Cov(X) = \begin{bmatrix} X_1(t_1) & X_2(t_1) & \cdots & X_n(t_1) \\ & & & \\ X_1(t_n) & X_2(t_n) & \cdots & X_n(t_n) \end{bmatrix} \quad (10)$$

The next process is to compute the eigenvalues for S matrix.

$$\det(S - \lambda I) = 0 \quad (11)$$

Finally, an eigenvector is created

$$E = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdots \\ \lambda_n \end{bmatrix}$$

Fig. 4 highlights the process involved in flattening the image.

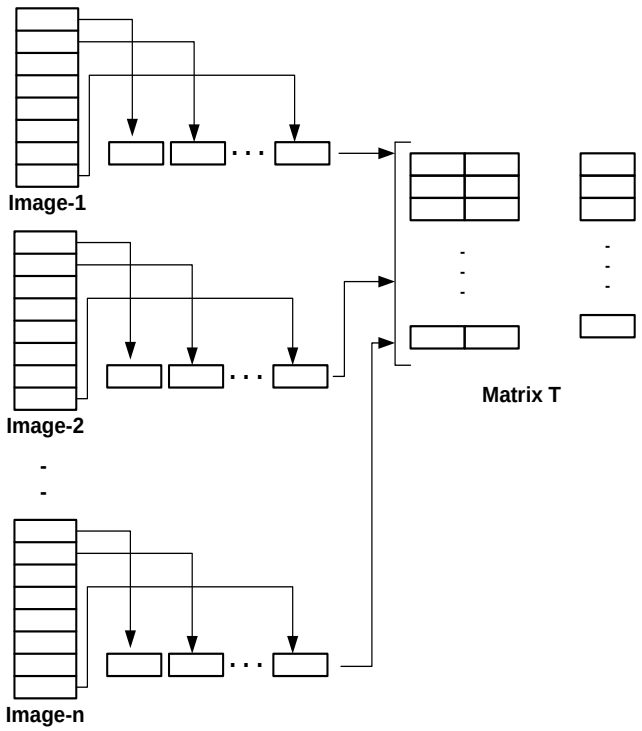


Fig. 4. Process of Flattening Image.

The step of fuzzification block is shown in Fig. 5 as follows.

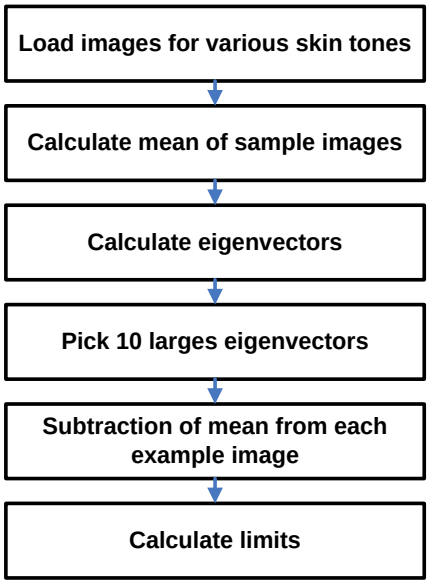


Fig. 5. Process Flow of Fuzzification.

E. Image Enhancement

This is the next steps of implementation after the binarization is carried out.In this process, the image is subjected for enhancement before passing it to the recognition algorithm. In this study hidden feature enhancements are being introduced which are not visible to human eyes however it will make a significant difference for the proposed system which will later contribute towards efficient authenticate a person. The final image will not be different from unprocessed images for human eyes.

V. Result and Discussion

This section discusses about the results being obtained after implementing the proposed logic discussed in prior section. The proposed system consists of nine metadata which is tabulated below in Table I.

TABLE I. Metadata of the Dataset

Sl. No.	Name	Description
1	Id	Identity of the person to whom hand belongs to
2	Age	Age of the person
3	Gender	Gender of the person
4	skinColor	Skin color of the person described by four types
5	Accessories	If the person is wearing a ring
6	nainPolish	If person has nail polish
7	aspectOfHand	Side of hand and identifies left/right hand
8	imageOfHand	Filename for image of hand
9	Irregularities	If the person is handicap person

The proposed system is scripted in python over 64 bit machine. 11K hands dataset has been chosen in this study since the dataset contain more images per person, higher resolution and other additional details [31][37]. The prime reason for selection of this dataset is because of its possession of maximum number of hand images for more than hundred number of subjects along with the presence of metadata. The contribution of the parameters used are: The parameter of minmax scaler Y' performs normalization, while R', G', B' is used for obtain pixel intensity information. From the algorithm of normalization, the effective control of RGB is carried out by parameters C_{max} and C_{min}, while a logical condition is maintained in order to yield Hue H and Saturation Values S. Since it is observed that number of images per subject is not same for all, so there is a need to carry out analysis of the frequency of it in order to ensure that the data is balanced. The frequency of images is plotted per subject and the plot is exhibited in Fig. 5.

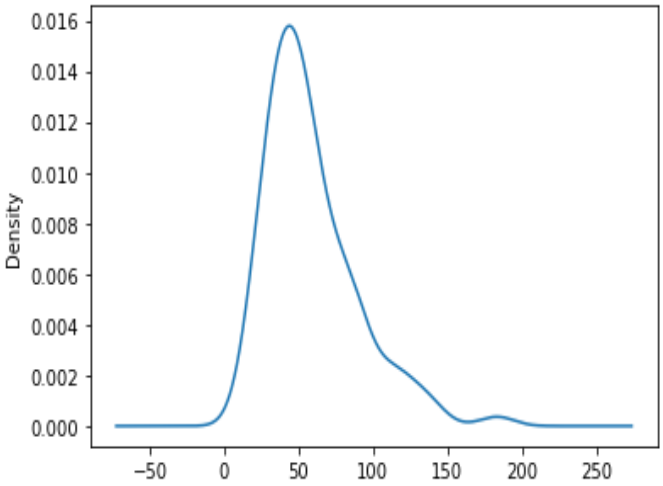


Fig. 6. Frequency Distribution of Images.

From Fig. 6, it can be observed that the data is not imbalanced as frequency doesn't deviate too much from mean/central tendency. As it can be observed, mean number of images per hand is 58.6. Minimum number of images per hand is 14 whereas maximum is 187. The detailed numerical description of the same is shown in following Table II as follow:

TABLE II. STATISTICAL DESCRIPTION

Items	Value
Count	189.00000
Mean	58.603175
Standard Deviation	30.841327
Minimum	14.00000
Maximum	187.00000

The visual analysis of the proposed system is also carried out for the images corresponding to subjects with handicap situation as shown in Fig. 7. Analysis with this dataset gives that opportunity.

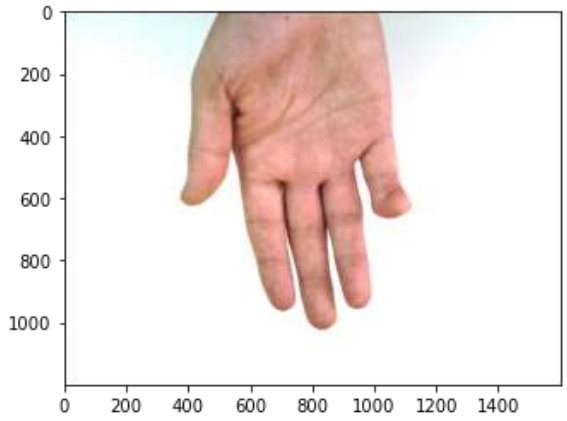


Fig. 7. Hand Image of Handicap Subject.

When the data if analyzed further, following conclusions can be drawn.

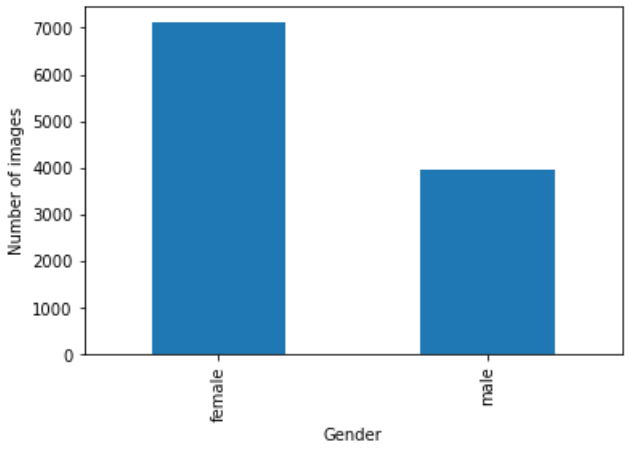


Fig. 8. Bar Graph of Female vs Male Hands.

There are around 7000 images of female hands whereas there are only around 4000 images of male hands as shown in Fig. 8. This will not pose many problems as there no significantly notable difference and gender recognition is not being perused in this study. Further, the proposed system also extracts information about Kernel Density Estimate (KDE) which is deployed for analyzing the probability density associated with the continuous variable.

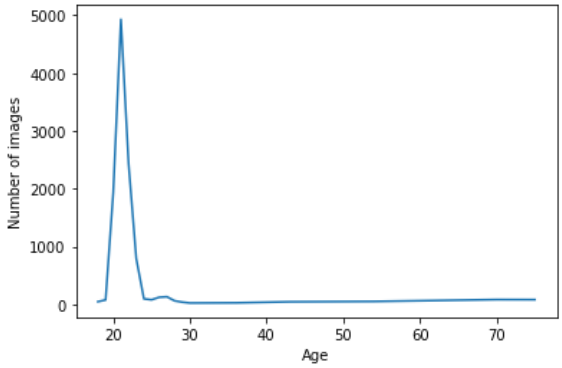


Fig. 9. KDE Plot for Age vs Number of Images.

In above figure we can observe that most of the hand images are from people of age between 20 and 30 although, there are some images of the 70-year-old person as shown in Fig. 9.

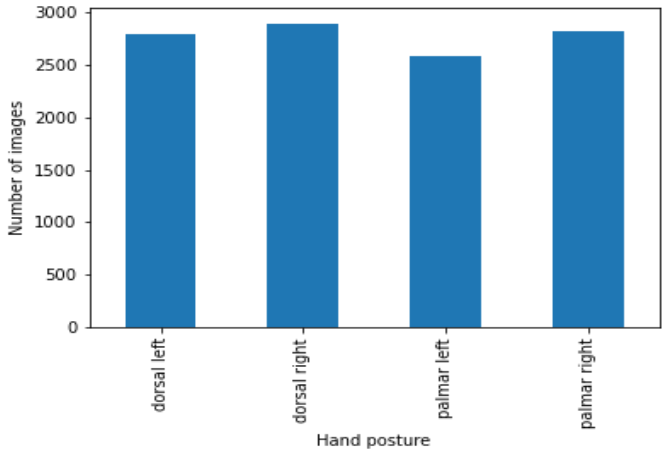


Fig. 10. Bar Graph of the Hand Poses.

As it can be observed above as shown Fig. 10, the images are evenly spread across among all four hand poses. Apart from this, it should be noted that image binarization is an important step in entire preprocessing algorithm. For every pixel within an image, the proposed system calculate whether it should be white in resulting image or black. Initially HSV values are considered. Based on all the images, considering all skin tones following minimum and maximum values are set by considering pixels belonging to skin of hand. Table III highlights the minimum and maximum values of HSV.

If any pixel value fall between these two maximum and minimum values then the pixel is made white by the algorithm. When the pixel is not within this range it could be possible that it is not within the range due to shadow and glaring effects. In order to avoid this phenomenon, proposed system uses fuzzy logic. In proposed fuzzification, the probability of the skin being part of actual skin is calculated by the help of eigenvectors. Eigenvectors are used to calculate a shape. They are nothing but set of vectors which represent the shape of the hand. The proposed system checks if the pixel falls within a shape of the hand in that particular image. In order to understand this system, the visual outcome exhibited in Fig. 6 represents the problems.

Fig. 11 shows the evident problems where middle finger and ringer finger are found to join with each other. This problem will eventually adversely affect the recognition process in later part. Hence, fuzzification significant contributes to address this problem. The fuzzy logic mechanism used in proposed system is very similar compared to the work of Chowdhury et al. [38] in which a very similar fuzzy logic algorithm is used to detect faces; however in this work, the equivalent process is being amended to detect hands. Any noises induced in this step will be removed by contour extraction as the algorithm considers only the larges contour among all. The next part of the visual analysis is that of image enhancement.

TABLE III. NUMERICAL OUTCOMES OF HSV

Items	Hue	Saturation	Value
Minimum	20	10	60
Maximum	40	165	255

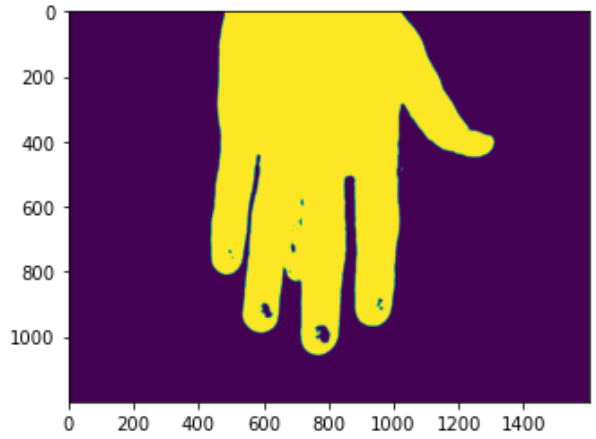


Fig. 11. Binarization without Fuzzy Logic.

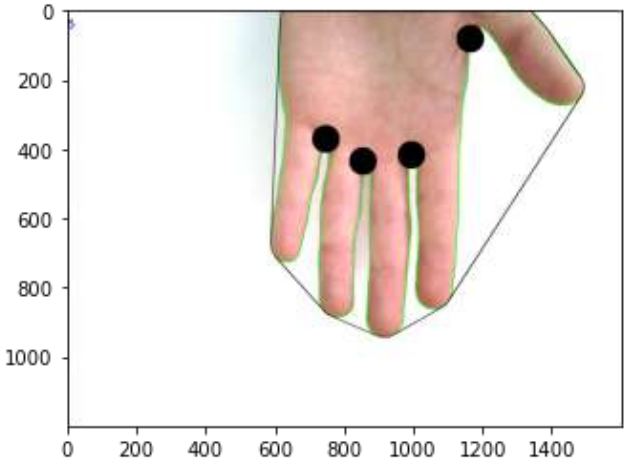


Fig. 12. Feature Enhancement Visible to Human Eyes.

Fig. 12 highlights the visual outcomes of enhancement operation

As it can be observed in above figure, majority the above features are enhanced. The contour is drawn around the hand, the out hull is drawn forming a polygon connecting the top most points of the hand, and finally the valleys in the contour are enhanced and these are the base of fingers. These are important features in hand geometry recognition.

The final stage of analysis is to carry out benchmarking of the proposed system by comparing with the existing system. For this purpose, the proposed system considers accuracy as performance parameters. Performance metrics of an authentication system is measured by Recognition accuracy. Formula for same has been provided below in expression (12).

$$Accuracy = \frac{Number\ of\ successful\ trials}{Total\ nuber\ of\ trials} * 100 \quad (12)$$

Successful trial means if a person is authenticated the system should recognize him as authenticated or else it will recognize him as not authenticated. However, if it authorizes an unauthorized person and vice versa, it will be considered as failure case. Total number of trials is summation of the success cases and failure cases. Comparison is carried out with the existing work of Afifi et al. [31] who has presented two schemes side-by-side e.g. CNN with SVM and CNN integrated with Linear Binary Pattern (LBP) and SVM. The analysis of accuracy was carried out for three use cases of palm side with dark, medium, and fair skin tone as shown in Table IV.

TABLE IV. NUMERICAL OUTCOME OF PROPOSED STUDY

Method	Palmer side – Dark skin	Palmer side – Medium skin	Palmer side – Fair skin
CNN - SVM	94.8	92.9	93.3
CNN – LBP - SVM	96.0	95.3	95.6
Proposed method	98.91	97.95	97.92

The graphical outcome of the above mentioned tabulated data is showcased in the Fig. 13.

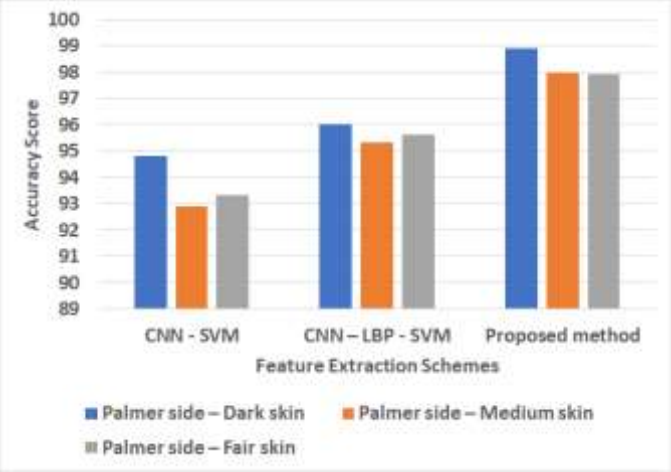


Fig. 13. Comparative Analysis Outcome.

The existing study of Afifi et al. [31] was able to produce 96% accuracy in most favorable conditions however the present work is able to show 98.91% accuracy which is a near-perfect system.

VI. CONCLUSION

This paper has presented a discussion of a novel feature extraction mechanism where the input image is considered along with its meta-data. After subjecting to blurring, color space conversion, binarization, and image enhancement, the study outcome shows better accuracy achievement for proposed system. The experimental outcome of the proposed study shows that it has achieved comparatively higher accuracy in the range of 97.95-98.91% for three different forms of biometric images. This value of accuracy is significantly improved version when compared to existing approaches of feature extraction using machine learning approaches. Following are the contribution / novelty of proposed study viz.

- the proposed study presents a simplified approach unlike the complex and iterative approaches presented by existing literatures,
- the proposed scheme introduces a fuzzification mechanism in order to control the possible cases of outliers arising from challenging illumination condition between the fingers,
- without usage of training mechanism, unlike the evolving methods in existing literatures, the proposed system managed to offer higher accuracy that is numerically proven.

The limitation of the paper is that it doesn't offer any privilege for performing authentication using certain use-cases. It is because the development of the model is basically a computational framework which is meant to be integrated with any form of biometrics where the extracted features will be considered for authentication. It can be widely applied to any form of hand-based biometric system.

The future work of the proposed system can be extended towards developing an authentication system. The extracted features can be further trained and stored which can be used as a metadata while performing authentication with a query image.

REFERENCES

[1] Rui Z, Yan Z. A survey on biometric authentication: Toward secure and privacy-preserving identification. IEEE Access. 2018 Dec 27;7:5994-6009.

[2] "What is Biometrics Security", https://www.kaspersky.com/resource-center/definitions/biometrics, Retrived on 16-08-2021.

[3] K. Přihodová and M. Hub, "Biometric Privacy through Hand Geometry-A Survey," 2019 International Conference on Information and Digital Technologies (IDT), 2019, pp. 395-401, doi: 10.1109/DT.2019.8813660.

[4] Sidlauskas DP, Tamer S. Hand geometry recognition. InHandbook of Biometrics 2008 (pp. 91-107). Springer, Boston, MA.

[5] C. Dhiman and D. K. Vishwakarma, "A Robust Framework for Abnormal Human Action Recognition Using $\boldsymbol{\mathcal{R}}$ -Transform and Zernike Moments in Depth Videos," in IEEE Sensors Journal, vol. 19, no. 13, pp. 5195-5203, 1 July1, 2019, doi: 10.1109/JSEN.2019.2903645.

[6] Y. Lu, S. Yoon, S. Wu and D. S. Park, "Pyramid Histogram of Double Competitive Pattern for Finger Vein Recognition," in IEEE Access, vol. 6, pp. 56445-56456, 2018, doi: 10.1109/ACCESS.2018.2872493.

[7] P. Chen et al., "Design of Low-Cost Personal Identification System That Uses Combined Palm Vein and Palmprint Biometric Features," in IEEE Access, vol. 7, pp. 15922-15931, 2019, doi: 10.1109/ACCESS.2019.2894393.

[8] J. Lu, V. E. Liong and J. Zhou, "Simultaneous Local Binary Feature Learning and Encoding for Homogeneous and Heterogeneous Face Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 8, pp. 1979-1993, 1 Aug. 2018, doi: 10.1109/TPAMI.2017.2737538.

[9] L. Fei, G. Lu, W. Jia, S. Teng and D. Zhang, "Feature Extraction Methods for Palmprint Recognition: A Survey and Evaluation," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 49, no. 2, pp. 346-363, Feb. 2019, doi: 10.1109/TSMC.2018.2795609.

[10] L. Fei, B. Zhang, W. Jia, J. Wen and D. Zhang, "Feature Extraction for 3-D Palmprint Recognition: A Survey," in IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 3, pp. 645-656, March 2020, doi: 10.1109/TIM.2020.2964076.

[11] B. Attallah, A. Serir, Youssef Chahir. Feature extraction in palmprint recognition using spiral of moment skewness and kurtosis algorithm.

Pattern Analysis and Applications, Springer Verlag,2019, 22 (3), pp.1197–1205. 10.1007/s10044-018-0712-5.

[12] S. Bakheet and A. Al-Hamadi, "Robust hand gesture recognition using multiple shape-oriented visual cues", EURASIP Journal on Image and Video Processing, 2021.

[13] P. D. Deshpande, P. Mukherji and A. S. Tavildar, "Accuracy enhancement of biometric recognition using iterative weights optimization algorithm", EURASIP Journal on Information Security, 2019.

[14] Y. Zhang, Y. Li and J. Su, "Iterative learning control for image feature extraction with multiple-image blends", EURASIP Journal on Image and Video Processing, 2018.

[15] S. Yang, L. Gong and D. Qiao, "Image offset density distribution model and recognition of hand knuckle", EURASIP Journal on Image and Video Processing, 2019.

[16] F. Liu, S. Jiang, B. Kang and T. Hou, "A Recognition System for Partially Occluded Dorsal Hand Vein Using Improved Biometric Graph Matching," in IEEE Access, vol. 8, pp. 74525-74534, 2020, doi: 10.1109/ACCESS.2020.2988714.

[17] S. Zhao and B. Zhang, "Learning Salient and Discriminative Descriptor for Palmprint Feature Extraction and Identification," in IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 12, pp. 5219-5230, Dec. 2020, doi: 10.1109/TNNLS.2020.2964799.

[18] P. Gupta and P. Gupta, "Multibiometric Authentication System Using Slap Fingerprints, Palm Dorsal Vein, and Hand Geometry," in IEEE Transactions on Industrial Electronics, vol. 65, no. 12, pp. 9777-9784, Dec. 2018, doi: 10.1109/TIE.2018.2823686.

[19] Z. Izakian, M. S. Mesgari, R. Weibel, "A feature extraction based trajectory segmentation approach based on multiple movement parameters", Elsevier-Engineering Applications of Artificial Intelligence, Vol.88, 2020.

[20] S. Ryu, J. Suh, S. Baek, S. Hong and J. Kim, "Feature-Based Hand Gesture Recognition Using an FMCW Radar and its Temporal Feature Analysis," in IEEE Sensors Journal, vol. 18, no. 18, pp. 7593-7602, 15 Sept.15, 2018, doi: 10.1109/JSEN.2018.2859815.

[21] M. Xin and Y. Wang, "Research on image classification model based on deep convolution neural network", EURASIP Journal on Image and Video Processing, 2019.

[22] C. Du, L. Zhang, X. Sun, J. Wang and J. Sheng, "Enhanced Multi-Channel Feature Synthesis for Hand Gesture Recognition Based on CNN With a Channel and Spatial Attention Mechanism," in IEEE Access, vol. 8, pp. 144610-144620, 2020, doi: 10.1109/ACCESS.2020.3010063.

[23] M. K. Abd-Ellah, A. I. Awad, A. A. M. Khalaf, and H. F. A. Hamed, "Two-phase multi-model automatic brain tumour diagnosis system from magnetic resonance images using convolutional neural networks", EURASIP Journal on Image and Video Processing, 2018.

[24] G. Li, R. Zhang, M. Ritchie and H. Griffiths, "Sparsity-Driven Micro-Doppler Feature Extraction for Dynamic Hand Gesture Recognition," in IEEE Transactions on Aerospace and Electronic Systems, vol. 54, no. 2, pp. 655-665, April 2018, doi: 10.1109/TAES.2017.2761229.

[25] M. J. J. Ghrabat, G. Ma, I. Y. Maolood, S. S. Alresheedi and Z. A. Abduljabbar, "An effective image retrieval based on optimized genetic algorithm utilized a novel SVM-based convolutional neural network classifier", Springer-Human-Centric Computing and Information Sciences, 2019.

[26] Z. Pan, J. Wang, Z. Shen, X. Chen and M. Li, "Multi-Layer Convolutional Features Concatenation With Semantic Feature Selector for Vein Recognition," in IEEE Access, vol. 7, pp. 90608-90619, 2019, doi: 10.1109/ACCESS.2019.2927230.

[27] C. Wu et al., "sEMG Measurement Position and Feature Optimization Strategy for Gesture Recognition Based on ANOVA and Neural Networks," in IEEE Access, vol. 8, pp. 56290-56299, 2020, doi: 10.1109/ACCESS.2020.2982405.

[28] C. Yoo, S. Ji, Y. Shin, S. Kim and S. Ko, "Fast and Accurate 3D Hand Pose Estimation via Recurrent Neural Network for Capturing Hand Articulations," in IEEE Access, vol. 8, pp. 114010-114019, 2020, doi: 10.1109/ACCESS.2020.3001637.

[29] G. Yuan, X. Liu, Q. Yan, S. Qiao, Z. Wang and L. Yuan, "Hand Gesture Recognition Using Deep Feature Fusion Network Based on Wearable Sensors," in IEEE Sensors Journal, vol. 21, no. 1, pp. 539-547, 1 Jan.1, 2021, doi: 10.1109/JSEN.2020.3014276.

[30] A. Wibisono and P. Mursanto, "Multi Region-Based Feature Connected Layer (RB-FCL) of deep learning models for bone age assessment", Springer-Journal of Big Data, 2020

[31] Mahmoud Afifi, "11K Hands: Gender recognition and biometric identification using a large dataset of hand images." Multimedia Tools and Applications, 2019.

[32] Pitaloka DA, Wulandari A, Basaruddin T, Liliana DY. Enhancing CNN with preprocessing stage in automatic emotion recognition. Procedia computer science. 2017 Jan 1;116:523-9.

[33] Ji H, Lu W, Shen L. Backbone Based Feature Enhancement for Object Detection. InProceedings of the Asian Conference on Computer Vision 2020.

[34] Kumar N, Nachamai M. Noise removal and filtering techniques used in medical images. Orient J. Comp. Sci and Technol. 2017 Mar;10(1).

[35] Oldal LG, Kovács A. Hand geometry and palmprint-based authentication using image processing. In2020 IEEE 18th International Symposium on Intelligent Systems and Informatics (SISY) 2020 Sep 17 (pp. 125-130). IEEE.

[36] Kolkur S, Kalbande D, Shimpi P, Bapat C, Jatakia J. Human skin detection using RGB, HSV and YCbCr color models. arXiv preprint arXiv:1708.02694. 2017 Aug 9.

[37] “11k Hands”, https://sites.google.com/view/11khands, Retrived on 16-08-2021.

[38] Chowdhury A, Tripathy SS. Human skin detection and face recognition using fuzzy logic and eigenface. In2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE) 2014 Mar 6 (pp. 1-4). IEEE.

Categorical Vehicle Classification and Tracking using Deep Neural Networks

Deependra Sharma, Zainul Abdin Jaffery
Department of Electrical Engineering
Jamia Millia Islamia, New Delhi, India

***Abstract*—The classification and tracking of vehicles is a crucial component of modern transportation infrastructure. Transport authorities make significant investments in it since it is one of the most critical transportation facilities for collecting and analyzing traffic data to optimize route utilization, increase transportation safety, and build future transportation plans. Numerous novel traffic evaluation and monitoring systems have been developed as a result of recent improvements in fast computing technologies. However, still the camera-based systems lag in accuracy as mostly the systems are constructed using limited traffic datasets that do not adequately account for weather conditions, camera viewpoints, and highway layouts, forcing the system to make trade-offs in terms of the number of actual detections. This research offers a categorical vehicle classification and tracking system based on deep neural networks to overcome these difficulties. The capabilities of generative adversarial networks framework to compensate for weather variability, Gaussian models to look for roadway configurations, single shot multibox detector for categorical vehicle detections with high precision and boosted efficient binary local image descriptor for tracking multiple vehicle objects are all incorporated into the research. The study also includes the publication of a high-quality traffic dataset with four different perspectives in various environments. The proposed approach has been applied on the published dataset and the performance has been evaluated. The results verify that using the proposed flow of approach one can attain higher detection and tracking accuracy.**

Keywords*—Vehicle classification; generative adversarial networks; single shot multibox detector; vehicle tracking; deep neural networks

I. Introduction

With a rising count of vehicles on road, and those in a huge variety, resulting in traffic congestion and a slew of related difficulties, it is necessary to address these issues [1]. It motivates us to consider an intelligent and smart traffic monitoring system that could assist traffic agencies in addressing issues such as routing traffic based on the density of vehicle movement on the road, collecting traffic data like count of vehicles, vehicle type, and vehicle motion parameters, and managing roadside assistance in the event of an accident or other anomalous incident. It conducts traffic analysis using the acquired data to optimize the use of highway networks, forecast future transportation demands, and enhance transportation safety [2]. The primary functions of an intelligent and intelligent traffic monitoring system are vehicle categorization and tracking on a category basis. Due to the substantial technological problems associated with the same, several research topics have been studied, resulting in the creation of numerous vehicle categorization, and tracking systems. Classifying vehicles and maintaining their trajectories properly in a variety of environmental circumstances is critical for efficient traffic operation and transportation planning.

The scientific advancements have resulted in the development of several novel vehicle categorization systems. Three types of categorical vehicle classification systems may be found in use today: in-road, over-road, and side-road. Each category of vehicle classification is further divided into subcategories depending on the sensors utilized, the techniques used to utilize the sensors, and the processes used to classify cars [3]. While both in-road and side-road approaches are capable of accurate categorical vehicle classification, they differ significantly in terms of sensor types, hardware configurations, configuration process, parameterization, operational requirements, and even expenses, making it even more difficult to determine the most suitable solution for a given vehicle in the first instance. These techniques have limitations when more than one vehicle is in the same location at the same time [4]. So, these techniques can't be utilized for tracking the vehicles.

To circumvent the restrictions, over-the-road-based methods for category vehicle classification and tracking are used. Camera-based systems are the most popular technology for over-road-based systems [5] [6]. The cameras are mounted at a height sufficient to cover the road's wide field of vision and can span several lanes. There are two primary obstacles to attaining our aim that are linked with camera-based systems. To begin, their performance is significantly impacted by weather and lighting conditions, resulting in blurred, hazy, and rainy observations in collected pictures. The same findings are made in captured pictures when automobiles are travelling at high speeds on the road. Second, a higher viewing angle allows for consideration of more distant road surfaces, however, the vehicle's object size changes significantly, and the accuracy of detection of tiny objects located distant from the road suffers because of the shift. We focus on above two difficulties in this work to provide a feasible solution, and we demonstrate how to adapt the category vehicle recognition findings to multiple object tracking.

II. Related Work

A. Image Restoration

Images restoration problems such as image deblurring, dehazing and deraining being all focused at creating an

accurate representation of a clear final picture out of an insufficiently clear input image. Numerous studies have been conducted in this area. A multi-layer perceptron technique for deblurring that eliminates noise and artefacts [7]. To cope with outliers, a CNN based on the single value dissemination is used [8]. Certain techniques [9], [10] begin by estimating blur kernels with convolutional neural networks and subsequently deblur images using traditional restoration methods. Many edge adaptive neural networks have been developed for the purpose of recovering clear images instantly [11], [12]. Recent deep learning-based approaches for image dehazing [13], [14] estimate transmission maps first and subsequently restore clear images using conventional methodologies [15]. Typically, traditional methods for image deraining are created using the statistical characteristics of rainy streaks [16-19]. The author in [20] built neural network for removing rain and/or dirt from pictures. Having been developed with the aid of the ResNet [21], [22] built deep network for image deraining. The author in [23] introduced Generative Adversarial Network (GAN) architecture for generating realistic pictures from random noise. Numerous techniques for visual tasks have been developed because of this framework [24–27]. The authors in [28-31] have also utilized the GAN framework to low-level vision issues. We chose to apply the capabilities of the GAN framework physics model [32] for picture restoration jobs due to the positive findings.

B. Detection of Vehicles

Now, vehicle detection can be accomplished using both standard machine vision techniques and sophisticated deep learning techniques. Traditionally, machine vision techniques employ a vehicle's motion to distinguish it from a fixed backdrop picture. This approach may be classified into three categories [33] as background subtraction [34], frame subtraction on a continual basis [35], and optical flow [36]. Variance is determined by applying the frame subtraction technique, which compares pixel data of two or three successive frames. Additionally, threshold separates the shifting foreground region [35]. By employing this technique and reducing noise, the vehicle's halt may also be recognized [37]. When the video's backdrop picture being stationary, background data is used to build the model [37]. Following that, it is possible to segment the moving object as well as the frame pictures by comparing each frame image to the backdrop model. Optical flow approach being exploited to detect a motion area in frames. The resulting optical flow field encodes the direction of motion and speed of each pixel [36]. While the classic machine vision approach detects the vehicle more quickly, it does not perform well in case the image brightness varies, there being a continuous motion in backdrop, or there are vehicles moving with low speed or some complicated sceneries. Vehicle identification using deep convolutional neural networks [52] may be classified into two broad groups. The two-stage technique begins by generating a candidate box for the item using multiple methods and then classifying it using a CNN. Second, a single-stage technique could not produce candidate box but instead turns object bounding box placement problem straight transform it into a regression problem that can be processed. Region-CNN (R-CNN) [38] employs a two-stage technique that utilizes selective search of region [39] in image. CNN image input must be fixed size, and the network's deeper structure needs a lengthy training period and uses a significant amount of storage capacity. SPP NET [40], which is based on concept of spatial pyramid matching, enables the network to accept pictures of varying sizes and provide fixed outputs. Among the one-stage techniques, the Single Shot Multibox Detector (SSMD) [41] and You Only Look Once (YOLO) [42] frameworks are most important. For many categories, SSD for single shot detectors (YOLO) that is significantly faster than the preceding state-of-the-art and as accurate as slower techniques that undertake explicit area recommendations and pooling, such as the Faster R-CNN [43]. SSMD's central idea is to forecast category scores and box offsets for a specific set of default bounding boxes by applying tiny convolutional filters on feature maps. We chose to use the SSD framework [43] for categorical vehicle identification and classification tasks due to the positive findings.

C. Tracking of Vehicles

Aspects of the functioning of an intelligent traffic system that need advanced vehicle object identification applications, such as multiple object tracking, are also crucial [44]. DBT (Detection-Based Tracking) and Detection-Free Tracking (DFT) are the two most common methods of initializing objects in multi-object tracking systems (DFT). To detect moving objects in video frames, the DBT method first uses background modelling to detect them before tracking them. However, the DFT technique is only capable of initializing the tracking object and cannot deal with the addition of new objects or the removal of current ones. Multi-object tracking algorithms must consider the similarity of items inside a frame, as well the associated problem of objects across frames, when developing their algorithms. The normalized cross-correlation function may be used to determine the similarity of objects inside a frame. As shown in [45], the Bhattacharyya distance is being used to calculate the distance between two objects based on the colour histograms of their respective images. When connecting inter-frame items, it is critical to specify that each item may appear on no more than one track at a time and that each track may include no more than one object. It is now possible to fix this issue by using either detection-level exclusion or trajectory-level exclusion. SIFT and ORB feature points were used for object tracking to overcome the difficulties caused by size and illumination changes in moving objects in [46] and [47], however this approach is slow and requires many feature points. The feature point detection technique Boosted Efficient Binary Local Image Descriptor (BEBLID) is proposed for use in this study [48]. BEBLID is considerably faster than SIFT and ORB in extracting feature points.

D. Our Contributions Comprise the following Items

- On the foundation of this work, a large-scale dataset of vehicle movement on roads has been developed, which may offer many distinct category vehicle objects that have been thoroughly annotated under diverse situations taken by high-mounted cameras. It is possible to utilize the dataset to test the performance of a variety of vehicle detection methods.
- For recovering blurred, hazy, or rainy images recorded in road scenes, a method based on the GAN framework

for image restoration has been developed. This approach is utilized to increase the accuracy of vehicle detection in road scenes.

- A technique based on convolutional neural networks, i.e., SSMD, is implemented for category vehicle detection.
- A system for tracking and analyzing several vehicles is presented for road situations. The BEBLID method extracts and matches the detected object's feature points.

Findings of this investigation will be discussed in further detail in the following sections. Section 3 introduces the vehicle dataset that will be utilized in this work. During Section 4, you'll learn about the general procedure of the suggested system. Section 5 shows the results of the experiments as well as the relevant analyses. Section 6 provides a comprehensive summary of the complete method.

III. VEHICLE DATASET

Because of concerns about copyright, privacy, and security, traffic dataset is rarely made public owing to the widespread use of traffic surveillance cameras on highways across the world. With images of highway sceneries and typical road scenes, the KITTI benchmark dataset [31] aids in the solution of issues such as 3D object identification and tracking, which are commonly encountered in automated vehicle driving applications. The Tsinghua-Tencent Traffic-Sign Dataset [32] contains pictures captured by automobile cameras in a variety of lighting and weather situations, however there are no cars identified. The Stanford Car Dataset [33] and the Comprehensive Cars Dataset [34] are vehicle datasets captured by non-monitoring cameras and featuring a bright car look; they are used in research and development. The datasets are captured by security cameras; one such dataset is BV Dataset [35], which is an example. Even though this dataset categorizes vehicles into 6 categories, shooting angle being positive, and the vehicle object is too tiny for each image, making the generalization impossible for CNN training. A dataset called Traffic and Congestions [36] comprises photos of cars on roads collected by security cameras, however most of the images have some degree of occlusion in them. This dataset has a small number of images and contains no information on the vehicle's classification, making it less helpful. As a result, only a few datasets have pertinent annotations, and there are only a few images of traffic scenes available. This section provides an overview of the vehicle dataset from the standpoint of road surveillance footage that we created. Dataset available on link: https://drive.google.com/drive/folders/1vYwLPkZZ2OX1cIIPQZA4SgB3dum7vPwV?usp=sharing. The video in the dataset is taken from the DND road in Delhi, India as shown in Fig. 1. The road monitoring camera was put on the side of the road and built at a height of 10 meters with a fixed angle of view. The photos taken from this vantage point span a large portion of the road in the distance and include cars of all types. The pictures in the dataset were taken from four surveillance cameras at different times of day and under varied lighting situations to provide a diverse range of photographs. The vehicles in this dataset are divided into three categories: two-wheelers, Light Motor Vehicles (LMV), which include three-wheelers, automobiles, minivans, and other similar vehicles, and Heavy Motor Vehicles (HMV), which include buses, trucks, and other similar vehicles (Fig. 2). The Table I details out the information about the dataset published.

An initial training set and a second test set are included in this dataset, which is separated into two sections. Two-wheelers accounted for 28.45 percent of all vehicles in our dataset, while light motor vehicles (LMV) accounted for 61.34 percent and heavy motor vehicles (HMV) accounted for 10.21 percent. On average, each image has 3.64 instances of annotated instances. Comparing our dataset to the current vehicle datasets, ours has a greater number of categorized vehicle pictures, adequate lighting conditions, and comprehensive annotations.

(a) View 1 (b) View 2 (c)View 3 (d) View 4

Fig. 1. Different views of the Dataset Collected.

Fig. 2. Dataset with Three Categories of Vehicles. (a) Two-Wheeler, (b) LMV and (c) HMV.

TABLE I. VEHICLE DATASET DETAIL

Image format	Size	Total number of images	Total number of annotated instances	Average annotated instances per image[$]
RGB	1280X720	10502	38228	3.64

[$]Total number of instances/Total number of images

IV. METHODOLOGY

The technique of the categorical vehicle classification and tracking system is described in detail in this section. First, the video data from the road traffic scenario is imported into the system. Second, the GAN framework is used to recover the pictures that have been captured. After that, the road area is excavated. The SSMD deep learning object detection technique is being used to recognize presence of vehicles belonging to three different categories in a road traffic environment. Finally, BEBLID feature extraction is carried out on the identified vehicle box to complete the tracking of numerous vehicle objects. In the proposed technique, the essential components of picture restoration, vehicle detection, propagating object states into future frames, linking current detections with existing objects, and controlling the lifespan of tracked objects are all discussed in detail. Diagram of the methodology's building blocks is depicted in Fig. 3.

A. Image Restoration

As previously stated, weather and lighting circumstances have a significant impact on the performance of camera-based systems, resulting in blurring, hazing, and precipitation observations in the captured pictures. High-speed vehicle movement on the road is observed in captured images, and the same or similar observations can be deduced from those images. The former scenario is caused by environmental changes and is thus less likely to occur, but the latter situation occurs almost without fail, necessitating the need for restoration. To achieve precise vehicle detection, it is necessary to repair the images to eliminate the issues that have arisen. Following a study of the literature on picture restoration approaches, we were encouraged by the positive results to apply the capabilities of the GAN framework physics model [32] to image restoration problems in our own research.

1) Image Restoration with GAN: An image restoration task is to predict a clear picture x from an input image y that as been provided. Fundamentally, the estimated x should be compatible with the input y under the picture creation paradigm, which is as follows:

$$y = H(x) \quad (1)$$

The operator H is used to transfer the unknown outcome x to the seen picture y. Depending on the situation, the blur, haze, or rain operation may be used. It is required to apply extra constraints on x to regularize it since the estimation of x from y is not well-posed. In the maximum a posteriori (MAP) paradigm, one frequently used method is predicated on the assumption that x may be solved by,

$$x^* = arg \max_x p(x|y) = arg \max_x p(y|x)p(x) \quad (2)$$

In the above equation, $p(y|x)$ and $p(x)$ are probability density functions, which are referred to as the likelihood term and image prior in the scientific literature, respectively. The mapping functions between *x* and *y* are directly learned using mathematical approaches,

$$x^* = G(y) \quad (3)$$

G is the mapping function in this case. In the case of the function *G*, it can be considered an inverse operator of *H*. If the mapping function can be predicted accurately, *G(y)* should be near to the ground truth, theoretically speaking.

The adversarial learning method used by the GAN algorithm is used to learn a generative model. It trains a generative network and a discriminative network at the same time by optimizing, among other things.

$$\min_G \max_D E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (4)$$

in which *z* represents random noise, *x* represents a genuine picture, and *D* represents a discriminative network are used. For the sake of convenience, we will also refer to a generative network as *G*. As part of the training process, the generator generates samples (*G(z)*) that may be used to deceive the discriminator, while the discriminator learns to discriminate between actual data and samples generated by the generator. A binary classifier is used as the discriminator. If the observed image y serves as the input to the generator, then the adversarial loss is,

$$\max_D E_{x\sim p_{data}(x)}[\log D(x)] + E_{y\sim data(z)}[\log(1 - D(G(y)))] \quad (5)$$

The value of (5) is near to zero if the distribution of the produced picture *G(y)* differs considerably from the distribution of the clear image, and it is greater if the distribution differs significantly from the clear image. It is possible to address the image restoration difficulty by doing the negative log procedure,

$$x^* = arg \min_x \rho(x, y) + \varphi(x) \quad (6)$$

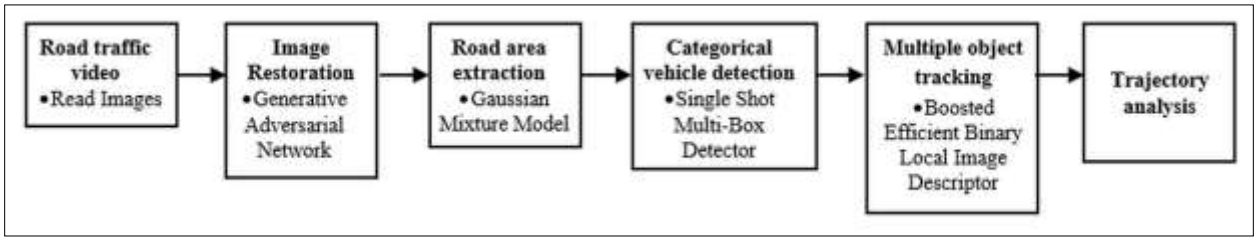


Fig. 3. Block Diagram of the Proposed System Methodology.

If we consider the data term to ensure that the recovered image x and the input image y are consistent under the appropriate image degradation model, then we get $\rho(x, y)$. The regularisation of the recovered image x is denoted by $\varphi(x)$ and models the characteristics of the recovered image, respectively. In vision tasks, the function $\varphi(x)$ functions as a discriminator, with the value of the function being considerably smaller if x is clear and much bigger otherwise. In other words, maximizing the goal function as Eq. 3 will result in a decrease in the value of x. As a result, the predicted intermediate picture will be significantly more detailed. Accordingly, in order to regularize the solution space of picture restoration, adversarial loss can be employed as a precursor to the restoration. Fig. 4 depicts the major components of the GAN method, which include two discriminative networks, one generative network, and one picture degradation model [32], as well as their interactions.

Let x_i and y_i indicate the clear and blurred images, respectively. The generative network derives the mapping function G from the input y_i and creates the intermediate restored image $G(y_i)$. The physics model for regenerating the image $\tilde{y}_i$ for various operations is as follows: for image deblurring,

$$\tilde{y}_i = k_i \otimes G(y_i) \tag{7}$$

where k_i being the kernel for blur, and $\otimes$ represents convolution operator. For image dehazing and deraining,

$$\tilde{y}_i = G(x_i)t_i + A_i(1 - t_i) \tag{8}$$

where A_i representing an atmospheric factor and t_i being the transmission map. The discriminative network D_g is used to determine if the distributions of the generator G outputs are comparable to those of the ground truth images. It is required to categorize using the discriminative network D_h whether the regenerated result $\tilde{y}_i$ is consistent with the observed image y_i. All the networks are taught in a collaborative manner from beginning to end.

During training, we rely on an Adam optimizer, which starts with a learning rate of 0.0002, with the method outlined in [24] being used. To get our results, we choose a batch size of one and a slope of 0.2 for the Leaky-ReLU. We use the same weight initialization strategy [24] uses. We must first get the generator G to create $G(y_i)$ and y_i. We may utilise the relevant physics model parameters to employ the generator, as we know the training data as well as the physics model parameters $\tilde{y}_i$. The discriminators D_g and D_h accept input data sets $\{x_i, G(y_i)\}$ and $\{y_i, \tilde{y}_i\}$ respectively. We update the discriminators using a history of produced pictures (rather than the most recent generative networks' images) according to the methods discussed in [24]. The generator and the discriminators have a one-to-one update ratio set between them.

B. Excavation of the Road Area

The next section covers the procedure for removing the road surface. We developed it using an image processing approach called the Gaussian mixture model, which results in superior vehicle detection results when combined with the deep learning object detection method, as shown in Fig. 2. The video picture of traffic on the road has a wide field of vision. In this investigation, the cars are the primary centre of attention, and the road area is the zone of interest in the resulting image. Meanwhile, depending on the camera's view angle, road area being focused for certain range of the image's horizontal and vertical planes. We were able to extract the road segments from the video using this function. In a traffic scenario, a perfect background is not always accessible and may always be modified in crucial circumstances by the introduction or removal of items from the picture, as well as the presence of objects that are either slow moving or immobile. The Gaussian mixture model (GMM) was used to account for all these factors correctly. According to the method, background is visible more frequently than foreground and model variance is small [49].

The recent history of the intensity values of each pixel X_1, ..., X_t is modeled by a mixture of K Gaussian distribution. The probability of observing the current pixel value is given by the formula:

$$P(X_t) = \sum_{k=1}^{K} w_{k,t} * \eta(X_t, \mu_{k,t}, \Sigma_{k,t}) \tag{9}$$

where K gives the number of Gaussian distributions, $w_{k,t}$ is the weight of the k^{th} Gaussian in the mixture at time t having mean $\mu_{k,t}$ and covariance matrix $\Sigma_{k,t}$ and η is a Gaussian probability density function which is given by

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} exp^{-\frac{1}{2}(X_t - \mu)^T \Sigma^{-1} (X_t - \mu)} \tag{10}$$

where n is the dimension of the colour space and is the number of colours in the colour space. As soon as the parameters have been initialized, the K Gaussians are sorted in the order of the ratio 1/(k). Due to the fact that backgrounds are more prevalent in scenes than moving objects, as well as the fact that their values are almost constant, it follows that a backdrop pixel equates to a high weight with low variation. The first B Gaussian distributions that surpass a specific threshold T_1 are kept for use as a background distribution. For example,

$$B = \arg\min_b \left(\sum_{k=1}^{b} w_k > T_1 \right) \tag{11}$$

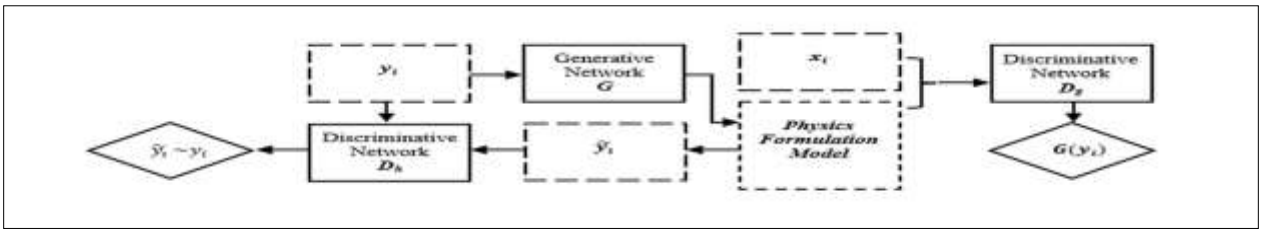


Fig. 4. Major Components of the GAN Framework.

(a) View 1 (b) View 2 (c) View 3 (d) View 4

Fig. 5. Road Area Extracted for all Four Views.

Distributed data that is part of the foreground represents other distributions. Until a match is found, the process repeats as the system computes and compares every new X_t value to the K Gaussian distributions. A pixel's value follows a Gaussian distribution if it is 2.5 standard deviations away from that distribution's mean. The background image is smoothed using a Gaussian filter once the road section has been extracted as the background picture. The MeanShift method smoothes the input image's colour. The final step is to finish filling the holes and carrying out morphological procedures in order to get most of the road surface. To extract the road regions, we made use of a variety of landscapes and have the results in Fig. 5.

C. Categorical Vehicle Detection using SSMD

Here is a description of the object detection approach that was employed in this study. The SSMD network was utilised in the development of the categorical vehicle detection framework and its deployment. The SSD approach's final detections are created by feeding bounding-boxes and scores of object class occurrences into a fixed-size feed-forward convolutional network followed by a non-maximum suppression phase. Addition of an auxiliary structure to the base network, such as the VGG-16, results in detections that have the following important characteristics:

1) Maps of multi-scale feature for identifying anomalies: At end of the truncated base network, convolutional feature layers are added to complete the network. These layers get smaller and smaller as time goes on, and they allow for predictions of detections at various sizes.

2) Convolutional neural network prediction techniques: A sequence of convolutional filters is associated with each feature layer, and it creates a discrete set of detection results. The three-dimensional tiny kernels provide either a score for each category or an offset in the shape relative to the default box coordinates and are the essential element used for the prediction of parameters in a feature layer of size mxn with channels. For each kernel location, it generates a number as an output. When it comes to figuring out the bounding box offset output values, it is crucial to first understand the differences between measurements made on various feature maps.

3) Box and aspect ratio defaults: In the design of feature map cells, each is equipped with default bounding boxes, even if many feature maps are employed above the cell. Due to the tiling of the feature map's boxes, with the position of each box in relation to its associated cell fixed, the boxes' arrangements in the feature map are fixed. We predict the offsets, class scores, and the box shapes for each feature map cell. From there, we calculate the class scores and four offsets to get the final bounding box, as seen in the illustration. The (c + 4)k filters being applied around each spot in the feature map amount to (c + 4)kmn outputs for a m x n feature map.

a) Training: For SSD training to be effective, the ground truth information must be allotted to certain detector outputs in the fixed set of detector outputs. Once a decision has been made on this assignment, it is applied completely to the loss function and back propagation. Additionally, you must pick the set of default boxes and scales that you will use for the data augmentation and the hard negative mining and methods.

i) Training method for matching: For training, we need to discover the ground truth boxes and train the network according to that discovery. For each ground truth box we create, we are using a preset box that is predefined with a variety of attributes, such as box size, box aspect ratio, and box placement. For every ground truth box, we compare it to the best-overlapping default box. Any boxes that meet the requirements are then matched to ground truth in which the jaccard overlap is over a certain level (0.5). By contrast, the learning challenge is made easier since the network may make predictions about a large number of default boxes that overlap, instead of needing to select one single box as the biggest overlapper.

ii) Loss function: The training aim is to be able to deal with a variety of vehicle types. We'll define an indication of matching a box in the i-th category to a box in the j-th category as $x_{ij}^{p} = \{1,0\}$. $\sum_i x_{ij}^{p} \geq 1$ holds under the matching strategy shown above. The weighted sum of the localization loss (loc) and the confidence loss (conf) are the overall objective loss function:

$$L(x, c, l, g) = \frac{1}{N}(L_{conf}(x, c) + \alpha L_{loc}(x, l, g)) \quad (12)$$

where N is the number of matching default boxes, and the weight term has been adjusted to one via cross validation. If N equals 0, the loss is set to zero. In a localization test, the localization loss is the difference between the expected box (l) parameters and the ground truth box (g) values.

iii) Scales and aspect ratios for default boxes: To manage diverse object scales, feature maps from many distinct layers in a single network are used for prediction, with parameters shared across all object scales. This allows the network to handle several object scales at the same time. In addition, it has been depicted that feature maps from the lower layers could help to enhance the quality of semantic segmentation since the lower layers capture finer features of the input

objects. For detection, we make use of both the bottom and higher feature maps. With the tiling of default boxes, we may train individual feature maps to be sensitive to objects of different sizes and shapes over time. Assume that we wish to make predictions using m feature maps. The following formula is used to determine scale of default boxes for every feature map:

$$s_k = s_{min} + \frac{s_{max} - s_{min}}{m-1}(k-1), k \in [1, m] \quad (13)$$

Where s_{min} equals 0.2 and s_{max} equals 0.9, the lowest layer has a scale of 0.2, the topmost layer has a scale of 0.9, and all levels in between are evenly spaced. We impose various aspect ratios on the default boxes, denoted by the variables $a_r \in \{1,2,3,1/2,1/3\}$. We can determine the width $w_k^a = s_k\sqrt{a_r}$ and height $h_k^a = s_k\sqrt{a_r}$ of each default box. The centre of each default box is set to $\left(\frac{i+0.5}{|f_k|}, \frac{j+0.5}{|f_k|}\right)$, where $|f_k|$ denotes the size of the k-th square feature map, $i, j \in [0, |f_k|]$.

iv) Hard negative mining: we rank the default boxes according to their largest confidence loss and choose just those at the top of the list, ensuring that the ratio of negatives to positives is no more than 3:1. This resulted in a speedier optimization process and more uniform training.

v) Enhancement of data: To make the model more robust to a broad range of input object sizes and shapes, each training image is randomly chosen using one of the following methods:

- Utilize the whole original input image.
- Sample a patch with values of 0.1, 0.3, 0.5, 0.7, or 0.9 to obtain the least feasible jaccard overlap with the objects.
- Take a sample of a patch at random.

Each sampled patch is between [0.1 and 1] of the original image's size, with an aspect ratio of between 1/2 and 2. Following the preceding sampling step, each sampled patch is given a fixed size, and the patches are then horizontally flipped with a probability of 50%.

D. Multiple Vehicle Object Tracking

This section describes how numerous vehicle objects are tracked using the object box discovered in the preceding section. During this stage, the BEBLID algorithm was employed to extract vehicle characteristics, and good results were achieved. The BEBLID method surpasses the competition by a considerable margin in terms of computing performance and matching costs. This algorithm is a superior alternative to other image description algorithms that have been previously described in the literature. Feature computations for the BEBLID algorithm are based on differences in grey values between a pair of box image regions, with the integral image serving as a basis for computations for the BEBLID algorithm features based on differences in grey values between a pair of box image regions. The technique takes use of AdaBoost to train a descriptor on an imbalanced data set to handle the challenge of highly asymmetric image matching. Binarization in a descriptor is achieved by minimizing the amount of new similarity loss in which all weak learners share a common weight. The coordinate system must be established by assuming the feature point to be at the centre of a circle and using the centroid of the point region to represent the coordinate system's x-axis. Thus, when the image is rotated, the coordinate system may be adjusted to match the image's rotation, resulting in rotation consistency in the feature point descriptor. When viewed from a different angle, a consistent point can be made. After getting the binarization, the feature points are matched using the XOR operation, which improves the overall efficiency of the matching process.

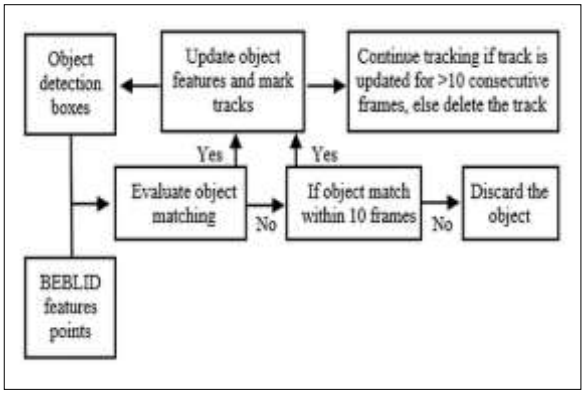


Fig. 6. Multiple Vehicle Object Tracking Method.

Fig. 6 illustrates the tracking method. When the number of matching points collected reaches a predefined threshold, the point is regarded successfully matched, and the object's matching box is painted around it. The following information relates to the source of the prediction box: Purification of feature points is performed using the Maximum Likelihood Estimator Sample Consensus (MLESAC) algorithm, which can exclude incorrect noise points caused by matching errors, and estimation of the homography matrix is performed using the MLESAC algorithm, which is capable of excluding incorrect noise points caused by matching errors. The estimated homography matrix and the location of the original object detection box are transformed into a perspective to get a matching prediction box for the original object detection box. Both the prediction box in the first frame and the detection box in the second frame must fulfil the centre point's criterion for the smallest distance between them to match the same item effectively. To be more specific, we define a threshold T equal to the greatest pixel change between the observed centre point of the vehicle object box and the vehicle object box's centre point when it moves between two subsequent video frames. The difference between two successive frames of the same vehicle in terms of positional movement is less than the threshold T. When the centre point of the vehicle object box crosses T in two subsequent frames, the vehicles in those two frames become unrelated, and the data connection fails. The threshold T value is proportional to the size of the vehicle object box, taking scale shift into vehicle. The thresholds for each vehicle object box are set to a variety of values. This definition is sufficiently flexible to accommodate vehicle mobility and a variety of different video input sizes. When T = box height/0.25 is used, the height of the vehicle object box is utilized as the input parameter for the calculation. We discard

any trajectory that has not been updated in ten consecutive frames, which is suitable for a camera scene with a wide-angle image collection along the route under investigation. If the prediction box does not match the item in future frames, it is determined that the object is absent from the video scene and the prediction box is removed. The method outlined above results in the collection of global object identification and tracking trajectories from the viewpoint of the whole road surveillance video.

E. Analysis of Trajectories

This section discusses both the analysis of moving objects' trajectories and the gathering of data on numerous items in a traffic flow. The majority of roadways are split into two lanes, separated by isolation barriers. We identify the vehicle's orientation in the world coordinate system based on its tracking trajectory and mark it as approaching or fleeing the camera. A straight line is drawn across the traffic scene image to serve as a detection line for the purpose of calculating vehicle classification data. The detection line must be centred on the 1/2 point of the traffic image's high side. Concurrently, the road's traffic flow in both directions is counted. The object's memory is accessed when the object's trajectory crosses the detection line. The number of objects in different orientations and categories over a certain time may be calculated at the end of the operation.

V. Simulation and Results

Many measures have been developed in the past for evaluating the systems performance quantitatively. The proper one depends heavily on the application, and the search for a single, universal evaluation criterion is currently underway. On one side, it being ideal to condense results into a single number that can be compared directly. On the other side, one could not want to lose knowledge about the algorithms' specific faults and present a large number of performance estimations, which makes a clear voting impossible. So, we would be evaluating the performances with more than one parameter.

A. For Image Restoration

1) Peak signal to noise ratio(PSNR): Considering a reference image f and a test image g, which have a resolution of MxN, the PSNR score among f and g being calculated as:

$$PSNR(f,g) = 10log_{10}((\max \text{pixel value})^2/MSE(f,g) \quad (14)$$

$$where, MSE(f,g) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(f_{ij} - g_{ij})^2 \quad (15)$$

The PSNR score increases as the mean squared error (MSE) decreases; this indicates that a greater PSNR value results in a higher image quality.

2) Structural similarity index (SSIM): The SSIM being a well-known quality statistic that is used to compare two images. It is thought to be connected to the human visual system's perception of quality. The SSIM score being calculated as:

$$SSIM(f,g) = l(f,g)c(f,g)s(f,g) \quad (16)$$

$$where, l(f,g) = \frac{2\mu_f\mu_g+C_1}{\mu_f^2+\mu_g^2+C_1} \quad (17)$$

$$c(f,g) = \frac{2\sigma_f\sigma_g+C_2}{\sigma_f^2+\sigma_g^2+C_2} \quad (18)$$

$$s(f,g) = \frac{\sigma_{fg}+C_3}{\sigma_f\sigma_g+C_3} \quad (19)$$

l: luminance, ***c***: contrast and ***s***: structural comparison function Few results of GAN framework for image restoration are shown in Fig. 7.

The images are randomly selected, and their performance is quantified in terms of PSNR and SSIM. The average of the two parameters' scores, is shown in Table II.

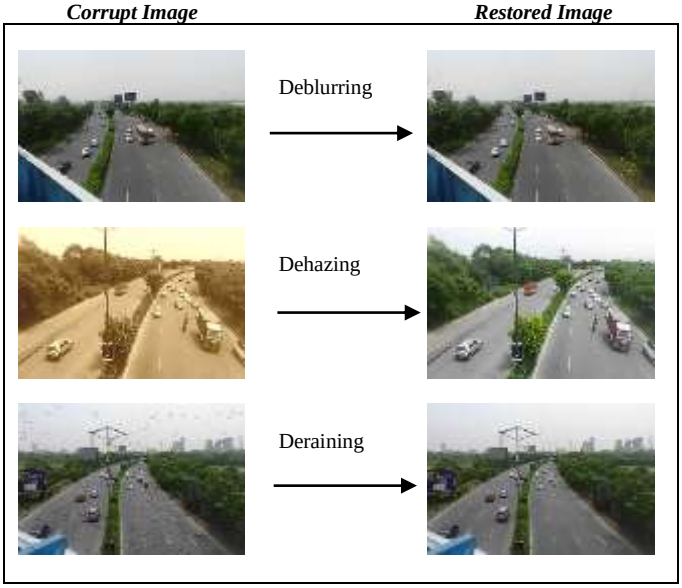


Fig. 7. Few Results of GAN Framework for Image Restoration.

TABLE II. Performance Evaluation of Image Restoration Method

Parameter [50]	Input	Average Scores		
		Deblurring	***Dehazing***	***Deraining***
PSNR	20.42	27.44	25.61	24.86
SSIM	0.5691	0.8811	0.9187	0.8367

B. For Vehicle Detection

It was necessary to use the test set to compute the mean average precision (mAP); mAP is an acronym for Average Precision (AP), which is defined as calculating the area under the precision-recall curve for a given total number of object class instances [43]. The experiment is divided into three classes, which include two-wheelers, light motor vehicles, and heavy motor vehicles. The mean of 11 points for each potential threshold in the category's precision/recall curve is described for each category by AP. We utilized a series of criteria [0, 0.1, 0.2,..., 1] to measure our results. For recall values larger than each threshold (in this experiment, the barrier is 0.25), there will be a matching maximum precision value, denoted by *pmax(recall)*. The precisions listed above are computed, and AP is the average of these 11 maximum precisions (recall). This number was used to describe the overall quality of our model.

$$AP = \frac{1}{11}\sum_{recall=0}^{1} p_{max}(recall), recall \in [0, 0.1, \dots, 1] \quad (20)$$

$$mAP = \frac{\sum AP}{class\ number} \quad (21)$$

The calculation of precision, recall and IoU (Intersection over union) is as follows:

$$Precision = \frac{TP}{TP+FP} \quad (22)$$

$$Recall = \frac{TP}{TP+FN} \quad (23)$$

$$IoU = \frac{area\ of\ overlap}{area\ of\ union} \quad (24)$$

in which TP, FN, and FP denote the number of true positives, false negatives, and false positives, respectively We used the following formulas to compute the parameter scores for both categories:

1) When the dataset was sent directly into the object detection algorithm, that is, when no image restoration procedure was used to restore the image.

2) When a picture is restored using the GAN framework, a dataset is fed into the object detection algorithm.

Tables III and IV provide the results of the parameters for each of the two categories. There is a 13.7 percent difference in the two-category results for the metric mAP when comparing them. This improvement figure clearly demonstrates that restoring the pictures has a significant influence on the quality of object identification and, indirectly, on the accuracy of tracking while tracking objects.

Few results of SSMD approach for categorical vehicle detection id depicted in Fig. 8.

C. Multiple Vehicle Object Tracking

The performance evaluation for multiple vehicle object tracking is done through following parameters [51]:

1) Multiple Object Tracking Accuracy (MOTA): This parameter takes into account three different types of errors: false positives, missed targets, and identity changes. For improved tracking accuracy, a high MOTA value is preferred. It is calculated as:

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t GT_t} \quad (25)$$

The frame index is t, and the count of ground truth objects is GT. MOTA could be negative if count of mistakes produced by tracker is more than total object count in the scene. MOTA score being solid indicator of tracking system's overall performance.

2) Multiple Object Tracking Precision (MOTP): Refers to average difference between all true positives and their ground truth objectives. For improved tracking, a high MOTP value is preferred. Average dissimilarity among all true positives and their matching ground truth targets is Multiple Object Tracking Precision. This being calculated as, for bounding box overlap:

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t} \quad (26)$$

$d_{t,i}$ is the bounding box overlap of target i with its assigned ground truth object, and c_t is count of matches in frame t. Average overlap among all properly matched hypotheses and their corresponding objects being given by MOTP, which spans among t_d: 50% and 100%.

3) False Alarms per Frame (FAF): It reflects per-frame amount of false alarms. A lower value of FAF is desirable for better tracking.

4) Mostly Tracked (MT): It indicates the number of paths that have been mainly tracked. i.e. the target has had the same label for at least 80% of its existence. A high value of MT parameter is desirable for better tracking.

5) Mostly Lost (ML): It indicates the amount of trajectories that have been lost for the most part. i.e. the target being not monitored for at least 20% of the time it is alive. A lower value of ML parameter is desirable for better tracking.

6) False Positive (FP): It reflects number of false detections. A lower value of FP parameter is desirable for better tracking.

7) False Negative (FN): It reflects number of missed detections. A lower value of FN parameter is desirable for better tracking.

8) IDsw: The amount of times an ID changes to a formerly tracked object. A lower value of IDsw parameter is desirable for better tracking.

9) Frag: The amount of times a track is fragmented due to a miss detection. A lower value of Frag parameter is desirable for better tracking.

TABLE III. PERFORMANCE EVALUATION OF VEHICLE DETECTION METHOD[1]

Para-meter	AP(%)			Precision	Recall	Average IoU (%)	mAP (%)
	Two-wheeler	*LMV*	*HMV*				
Scores	68.4	72.6	71.1	0.66	0.71	62.41	70.7

TABLE IV. PERFORMANCE EVALUATION OF VEHICLE DETECTION METHOD[2]

Para-meter	AP(%)			Precision	Recall	Average IoU (%)	mAP (%)
	Two-wheeler	*LMV*	*HMV*				
Scores	84.7	87.5	87.1	0.86	0.88	73.64	84.4

(a) View 1 (b) View 2 (c)View 3 (d) View 4

Fig. 8. Few Results of SSMD Approach for Categorical Vehicle Detection.

TABLE V. PERFORMANCE EVALUATION OF MULTIPLE VEHICLE OBJECT TRACKING METHOD

Parameter	MOTA(↑)	MOTP(↑)	FAF(↓)	MT(↑)	ML(↓)	FP(↓)	FN(↓)	IDsw(↓)	Frag(↓)
Scores	36.3	72.9	1.4%	13.4%	33.4%	140	304	35	28

(a) View 1 (b) View 2 (c)View 3 (d) View 4

Fig. 9. Few Results of Trajectory Estimation for Multiple Vehicle Object Tracking.

The score of the various tracking parameters is depicted in Table V. Trajectory estimation done on the dataset is depicted in Fig. 9. It summarizes the movement of vehicles with direction information and maps the future state predictions.

VI. CONCLUSION

This research developed from the standpoint of surveillance cameras, a dataset of vehicle objects and presented a technique for image restoration, object detection, and tracking for road traffic video scenes. The use of the GAN framework for picture restoration, as well as the GMM for road area extraction, resulted in a more effective detection system. The annotated road vehicle object dataset was used to train the SSMD object identification algorithm, which resulted in the development of an end-to-end vehicle detection model. The location of the object in the image being evaluated by the BEBLID feature extraction method based on results of the object detection technique and image data. The trajectory of the vehicle might thus be determined by tracking the binary characteristics of many objects. Lastly, the vehicle trajectories were examined to obtain information on the road traffic scene, such as driving direction as well as vehicle category and traffic density. Testing findings confirmed that suggested vehicle identification and tracking approach for road traffic scene has good performance and is practicable, as demonstrated by the outcomes of the experiments. The method described in this paper being low in cost and high in stability when compared to the traditional method of monitoring vehicle traffic by hardware. It also requires no large-scale construction or installation work on existing monitoring equipment, which is a significant advantage over the traditional method.

REFERENCES

[1] M. Won, T. Park, and S. H. Son, “Toward mitigating phantom jam using vehicle-to-vehicle communication,” IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 5, 2016, pp. 1313–1324.

[2] M. Won, S. Sahu, and K.-J. Park, “DeepWiTraffic: Low cost WiFi-based traffic monitoring system using deep learning,” arXiv:1812.08208 preprint, 2018.

[3] Myounggyu Won, “Intelligent Traffic Monitoring Systems for Vehicle Classification: A Survey,” IEEE Access, vol. 8, 2020, pp. 73340-73358.

[4] K. Kanistras, G. Martins, M. J. Rutherford, and K. P. Valavanis, “Survey of unmanned aerial vehicles (UAVs) for traffic monitoring,” Handbook of unmanned aerial vehicles, pp. 2643–2666, 2015.

[5] Z. Chen, T. Ellis, and S. A. Velastin, “Vehicle detection, tracking and classification in urban traffic,” in International IEEE Conference on Intelligent Transportation Systems, 2012, pp. 951–956.

[6] C. M. Bautista, C. A. Dy, M. I. Mañalac, R. A. Orbe, and M. Cordel, “Convolutional neural network for vehicle detection in low resolution traffic videos,” in IEEE Region 10 Symposium (TENSYMP), 2016, pp. 277–281.

[7] C. J. Schuler, H. C. Burger, S. Harmeling, and B. Sch¨olkopf, “A machine learning approach for non-blind image deconvolution,” in CVPR, 2013, pp. 1067–1074.

[8] L. Xu, J. S. J. Ren, C. Liu, and J. Jia, “Deep convolutional neural network for image deconvolution,” in NIPS, 2014, pp. 1790– 1798.

[9] J. Sun, W. Cao, Z. Xu, and J. Ponce, “Learning a convolutional neural network for non-uniform motion blur removal,” in CVPR, 2015, pp. 769–777.

[10] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Sch¨olkopf, “Learning to deblur,” IEEE TPAMI, vol. 38, no. 7, 2016, pp. 1439–1451.

[11] M. Hradis, J. Kotera, P. Zemc´ık, and F. Sroubek, “Convolutional neural networks for direct text deblurring,” in BMVC, 2015, pp. 6.1–6.13.

[12] S. Nah, T. Hyun Kim, and K. Mu Lee, “Deep multi-scale convolutional neural network for dynamic scene deblurring,” in CVPR, 2017, pp. 3883–3891.

[13] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, “Single image dehazing via multi-scale convolutional neural networks,” in ECCV, 2016, pp. 154–169.

[14] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, “Dehazenet: An end-to-end system for single image haze removal,” IEEE TIP, vol. 25, no. 11, 2016, pp. 5187–5198.

[15] K. He, J. Sun, and X. Tang, “Single image haze removal using dark channel prior,” in CVPR, 2009, pp. 1956–1963.

[16] L.-W. Kang, C.-W. Lin, and Y.-H. Fu, “Automatic single-imagebased rain streaks removal via image decomposition,” IEEE TIP, vol. 21, no. 4, 2012, pp. 1742–1755.

[17] Y. Luo, Y. Xu, and H. Ji, "Removing rain from a single image via discriminative sparse coding," in ICCV, 2015, pp. 3397–3405.

[18] Y.-L. Chen and C.-T. Hsu, "A generalized low-rank appearance model for spatio-temporally correlated rain streaks," in ICCV, 2013, pp. 1968–1975.

[19] Y. Li, R. T. Tan, X. Guo, J. Lu, and M. S. Brown, "Rain streak removal using layer priors," in CVPR, 2016, pp. 2736–2744.

[20] D. Eigen, D. Krishnan, and R. Fergus, "Restoring an image taken through a window covered with dirt or rain," in ICCV, 2013, pp. 633–640.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016, pp. 770–778.

[22] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," in CVPR, 2017, pp. 3855–3863.

[23] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in NIPS, 2014, pp. 2672–2680.

[24] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired imageto-image translation using cycle-consistent adversarial networks," in ICCV, 2017, pp. 2223–2232.

[25] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in ICML, 2017, pp. 1857–1865.

[26] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in CoRR, 2017, pp. 2849–2857.

[27] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, and Z. Wang, "Multi-class generative adversarial networks with the L2 loss function," CoRR, vol. abs/1611.04076, 2016.

[28] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in CVPR, 2017, pp. 4681–4690.

[29] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in CVPR, 2018, pp. 8183–8192.

[30] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Husz´ar, "Amortised MAP inference for image super-resolution," in ICLR, 2017.

[31] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in ICCV, 2017, pp. 251–260.

[32] Jinshan Pan, Jiangxin Dong, Yang Liu, Jiawei Zhang, Jimmy Ren, Jinhui Tang, Yu-Wing Tai, and Ming-Hsuan Yang, "Physics-Based Generative Adversarial Models for Image Restoration and Beyond," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[33] Al-Smadi M., Abdulrahim K., and Salam R.A., "Traffic surveillance: A review of vision based vehicle detection, recognition and tracking," International Journal of Applied Engineering Research, vol. 11, no. 1, 2016, pp. 713–726.

[34] Radhakrishnan M., "Video object extraction by using background subtraction techniques for sports applications," Digital Image Processing, vol. 5, no. 9, 2013, pp. 91–97.

[35] Qiu-Lin L.I., and Jia-Feng H.E, "Vehicles detection based on three-frame-difference method and cross-entropy threshold method," Computer Engineering, vol. 37, no. 4, 2011, 172–174.

[36] Liu Y., Yao L., Shi Q., and Ding J., "Optical flow based urban road vehicle tracking," IEEE Conference on Computational Intelligence and Security, 2013.

[37] Park K., Lee D., and Park Y., "Video-based detection of street-parking violation," IEEE Conference on Image Processing," vol. 1, 2007, pp. 152–156.

[38] Girshick R., Donahue J., Darrell T., and Malik J., "Rich feature hierarchies for accurate object detection and semantic segmentation," IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[39] Uijlings J.R.R., Van de Sande K.E.A., Gevers T., and Smeulders A.W.M., "Selective search for object recognition," International Journal of Computer Vision, vol. 104, no. 2, 2013, pp. 154–171.

[40] Kaiming H., Xiangyu Z., Shaoqing R., and Jian S., "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 37, no. 9, 2014, pp. 1904–16.

[41] Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C.Y., and Berg A.C., "SSD: Single shot multibox detector," European conference on computer vision, 2016, pp. 21–37.

[42] Redmon J., Divvala S., Girshick R., and Farhadi A., "You Only Look Once: Unified, real-time object detection," IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[43] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg, "SSD: Single Shot MultiBox Detector," 2016.

[44] Luo W., Xing J., Milan A., Zhang X., Liu W., Zhao X., and Kim T.K., "Multiple object tracking: A literature review," arXiv:1409.7618 preprint, 2014.

[45] Xing J., Ai H., and Lao S., "Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses," IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1200–1207.

[46] Zhou H., Yuan Y., and Shi C., "Object tracking using sift features and mean shift," Computer Vision & Image Understanding, vol. 113, no. 3, 2009, pp. 345–352.

[47] Rublee E., Rabaud V., Konolige K., and Bradski G.R., "ORB: An Efficient Alternative to SIFT or SURF," International Conference on Computer Vision, 2011.

[48] Iago Su´areza, Ghesn Sfeira, Jos´e M. Buenaposadac, and Luis Baumela, "BEBLID: Boosted Efficient Binary Local Image Descriptor," Pattern Recognition Letters, 2020.

[49] Zezhi Chen, Tim Ellis, and Sergio A Velastin, "Vehicle Detection, Tracking and Classification in Urban Traffic," IEEE Conference on Intelligent Transportation Systems, Alaska, USA, 2012.

[50] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity", IEEE Transactions on Image Processing, vol. 13, no. 4, 2004, pp. 600-612.

[51] K. Bernardin and R. Stiefel hagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," Image and Video Processing, 2008.

[52] D. Sharma, Z. A. Jaffery and N. Ahmad, "Categorical vehicle classification using Deep Neural Networks," International Conference on Power Electronics, Control and Automation, 2019, pp. 1-6.

Goal-oriented Email Stream Classifier with A Multi-agent System Approach

Wenny Hojas-Mazo[1], Mailyn Moreno-Espino[2]
José Vicente Berná Martínez[3], Francisco Maciá Pérez[4], Iren Lorenzo Fonseca[5]
Technological University of Havana "José Antonio Echeverria" – CUJAE, Havana, Cuba[1, 2]
Department of Computer Science and Technology, University of Alicante, Alicante, Spain[3, 4, 5]

***Abstract*—Now-a-days, email is often one of the most widely used means of communication despite the rise of other communication methods such as instant messaging or communication via social networks. The need to automate the email stream management increases for reasons such as multi-folder categorization, and spam email classification. There are solutions based on email content, capable of contemplating elements such as the text subjective nature, adverse effects of concept drift, among others. This paper presents an email stream classifier with a goal-oriented approach to client and server environment. The i* language was the basis for designing the proposed email stream classifier. The email environment was represented with the early requirements model and the proposed classifier with the late requirements model. The classifier was implemented following a multi-agent system approach supported by JADE agent platform and Implementation_JADE pattern. The behavior of agents was taking from an existing classifier. The multi-agent classifier was evaluated using functional, efficacy and performance tests, which compared the existing classifier with the multi-agent approach. The results obtained were satisfactory in all the tests. The performance of multi-agent approach was better than the existing classifier due to the use of multi-threads.**

Keywords*—Email stream classification; goal-oriented requirements; i*; multi-agent system

I. Introduction

Email is one of the most widely used services by Internet users. Moreover, the growth in the number of users makes this service grow as well. [1]. Every user can receive around 40-50 emails per day [2]; but other professional users may receive hundreds or thousands per day. Users spend a lot of time processing the emails they receive on a daily basis. This implies that email management is a major problem in organisations and that it is therefore important to have tools, preferably intelligent ones, to solve this problem. [1]. There are many types of tools for automatic mail management; one of them is the automatic email classifier [3, 4]. An automatic email stream classifier allows for quick and agile classification of emails into discrete sets of predefined categories [1]; for example, to classify an incoming email into professional or personal, spam or desirable, phishing or normal.

There are two levels where the email classification is applied: user and server. Email classification can be considered a goal that serves as a means to satisfy other goals. An example of this is the email filtering that the mail user agent performs in the client application or the spam email detector on the server. This includes other actors who relate to each other to achieve proper functioning, an aspect that could be represented through social modelling [5].

Email is one of the communication media through which most problems and security incidents occur due to spam and phishing [1]. According to [6] up to 80% of the emails sent worldwide are created by spam [6]. The adverse effect caused by spam emails has resulted in the economic loss of billions of dollars annually [7]. Several approaches have been proposed for spam detection [8]. To evaluate the performance of the filters, it has been published diverse corpus [9], different measures [10] and evaluation methods [11] have been used.

Moreover, the classifier, to be deployed in real environments, covers various aspects such as: the email pre-processing, the features selection, the concept drift detection and the classification itself that depends on the other aspects mentioned above. Proposals facing the challenge of increasing the adaptive capacity of email classification solutions tend to focus on specific modules [12, 13, 14, 15, 6]. However, these solutions do not provide a representation that relates the objectives to be achieved with each of the aspects to achieve the email classification. These relationships can lay the foundations for reactive, proactive and social behaviors that allow the classifier to increase his ability to adapt.

The main contribution of this work is a goal-oriented email stream classifier for client and server environment with a multi-agent system approach. Email environment requirements and proposed classifier were modeled with early and late requirements of i*. The email environment was represented with the early requirements and the proposed classifier with the late requirements. The classifier was implemented following a multi-agent system approach supported by JADE agent platform and Implementation_JADE pattern.

The paper is organized as follows: Section 2 offer background and an overview of related work. Section 3 presents proposed solution. Section 4 describes proposed solution evaluation. Section 5 gives conclusions.

II. Background and Related Work

The architecture of the email system consists of two kinds of subsystems [16]: Mail User Agents (MUAs) and the Message Transfer Agents (MTA). MUA is a client application that allows to the users to manage emails of their mailboxes. It can be desktop application (e.g. Thunderbird) or web-based (e.g. Gmail) and includes functionalities such as to compose, to

display, to organize and to filter messages. MTA, informally known as email servers, move the messages from the source to the destination sometimes through Internet, or if the recipient's server has been reached, to the Mail Deliver Agent (MDA). An example of email architecture [4] is showing in Fig. 1.

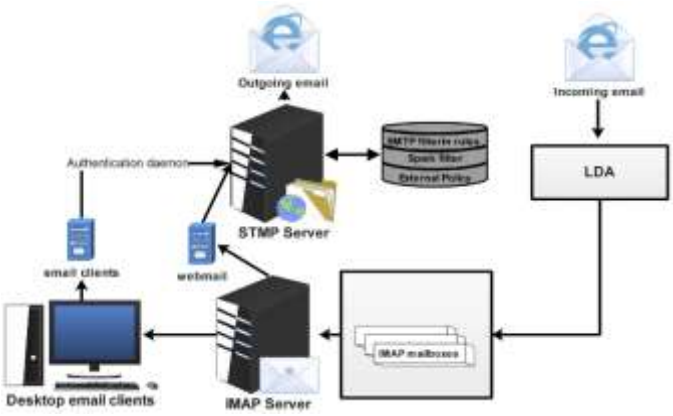


Fig. 1. Example of Email Architecture.

Other components complement the MTA such as user repository and content filter. User repository manages user information such as their username, password and profile information. Content filter evaluates incoming email to determine the probability that the messages are legitimate. This evaluation is supporting with filters such as antivirus filter and spam filter.

Moreover, the email stream classifiers, to be deployed in real environments, cover various aspects such as: the email pre-processing, the features selection, the concept drift detection and the classification itself that depends on the other aspects mentioned above. For example, features selection is intended to identify only those features with the highest discriminatory capacity to improve classifier performance [6]. Proposals, that face the challenge of increasing the adaptive capacity, tend to focus on specific modules [12, 13, 14, 15]. However, these solutions do not provide a representation that relates the objectives to be achieved with each of the aspects to achieve the classification of emails. These relationships can lay the foundations for reactive, proactive and social behaviors that allow the classifier to increase his ability to adapt. This requires that the solutions have adaptive capacities [17] to reduce the negative effects caused by noise, concept drift and the constant appearance of new instances. Requirements capture is an important action for developing an email stream classifier with these characteristics.

Requirements are the first stage when developing a software [18] and might be considered as one of the most important stages [19]. In the requirements analysis stage, the analysts detect the needs of the stakeholders to generate a system description document. At this stage, the goals, functionalities, and constraints (why is the system necessary) of the system are elicited in order to implement the established requirements [20]. Social modelling [21] is an option to capture these requirements.

Social modeling is focusing in system social dimension, it environment and adopts requirements with a goal-oriented approach by seeing intentions and reasons behind a behavior [21]. A detailed analysis of goals reveals desires, which shows expectations or troubles. A goal-oriented model may help to manage changes and allows evaluating alternative solutions by showing strengths and weaknesses [22, 23]. Goals provide criteria, support formal reasoning schemes during requirements engineering and guides to evaluate possible solutions [22, 23] and have been widely used and discussed in literature. Several examples of how to use goal-oriented models and how to apply them in real projects can be found in [21].

We can use the modeling language i* to introduces aspects of social modeling on requirements stage, this modeling follows a goal-oriented approach [24]. In i*, actors are seen as intentional, i.e. they have abilities, goals, beliefs and obligations. Thus, the analysis of each actor focuses on capturing their objectives, considering the relationships between the human actors and the future software system. This analysis allows setting the strategic interests of actors [25].

When we use i*, the requirement stage is divided in two other steeps [24]: step 1 Early Requirements and step 2 Late Requirements. Early requirements identify the actors involved in the context of the problem, their needs and their intentions. Late Requirements models what the futures software system should do and do this description using the most clear form as possible [26]. i* uses the Strategic Dependence model (SD) and the Strategic Rational model (SR) [24], each one with a different level of abstraction. In the SD model, dependency-relations existing among social actors are represented. In the SR model dependencies among objects within an actor are represente. Strength points of goal-oriented approach can be exploiting by agent-oriented [21].

III. Proposed Solution

During the email stream classifier development, three fundamental activities were performed. The first and second activities are focusing in to model in i* by using the tool TAOM4E [27]. The activity third consists in to implement the email stream classifier with a multi-agent system approach.

A. Activity 1

Use early requirements to model the requirements of the email context. The following steps were performed and Table I describes social actors identified:

1) Identify and model the social actors that are involved in the business context. All the actors modeled are show in Table I.

2) Represent the dependencies between actors using the SD model.

3) Identify and represent the objectives of each actor through the RS model.

In Fig. 2 are represented the SD and SR models obtained during the step of early requirements.

TABLE I. SOCIAL ACTORS IDENTIFIED IN EARLY REQUIREMENTS

Social actor	Description
User	Represents the people that use a MUA to manage emails.
MUA	Is a client application that allows to the users manage emails. It can be desktop application (e.g. Thunderbird) or web-based (e.g. Gmail).
MTA	This server application receives emails from MUA, or from another MTA, transfers the mail to another MTA (e.g. using Internet) and if the recipient's server has been reached, transfers the email to the MDA. Postfix is an extended example.
MDA	This server program stores the mail received form servers´s MTA into the mailbox. An example is Dovecot.
Content Filter	Evaluates incoming email to determine the probability that the messages are legitimate. An example is Amavis.
Antivirus Filter	A server program to recognize virus so as to prevent its delivery.
Spam Filter	A server program to recognize spam so as to prevent its delivery.
Internet	A vast collection of different networks that use certain common protocols and provide certain common services.
User Repository	Manages user information such as their username, password and profile information. It can be an SQL database, an LDAP or so on.

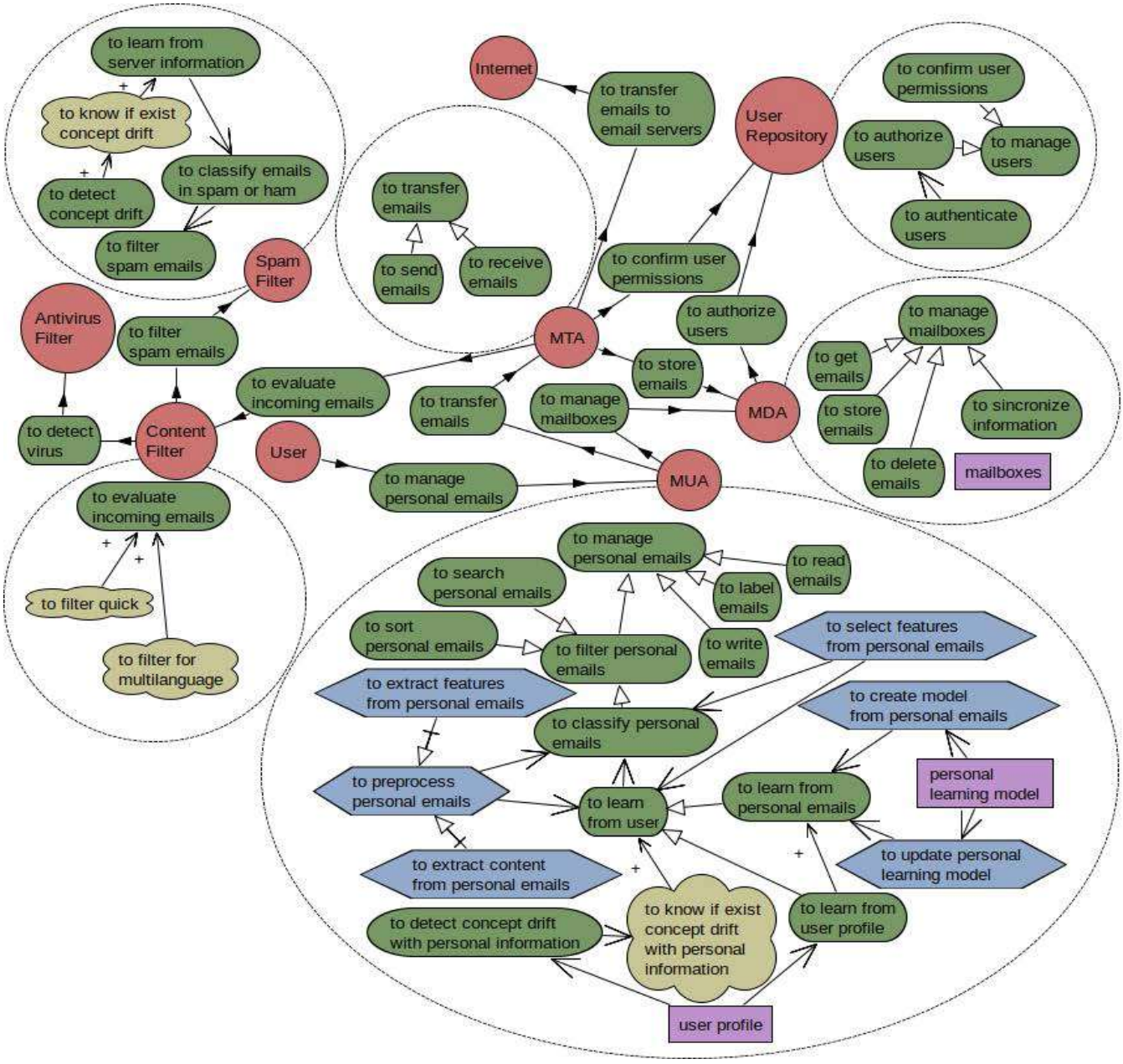


Fig. 2. SD and SR Models of Early Requirements.

B. Activity 2

To model the email stream classifier with late requirements we use the following steps:

1) Modelling social actors within the context of the system.

2) Representing dependencies among actors in an SD model.

3) Representing the goals of each actor in SR models.

Table II shows the social actors identified during late requirements. In Fig. 3 are represented the SD and SR models obtained during late requirements.

C. Activity 3

To implement the email stream classifier with a multi-agent system approach.

In this activity, a Multi-Agent System (MAS) based in late requirement modeling with i* to classify email (spam or ham) is proposed. The MAS is useful for problem resolution in distributed environments [28]. MAS agents are always active and organize so that their behavior emerges from the bottom up. This makes it easier to change the organizational structure when appropriate, or to expand its use, which enhances reusability. The agents of the proposed MAS are the actors represented as systems in the modeling of the late requirements, which can change their behavior without having a negative impact on the rest of the system.

The implementation of the solution was based on an existing spam email classifier, the JADE agent platform [29] and the Implementation_JADE pattern [30]. The classifier was used to take the behavior that each of the agents identified in the late requirements diagram would have. The JADE platform supports the development of the MAS, which provides advantages such as [31]: simplifying the development of a MAS, guaranteeing compliance with FIPA standards; high cohesion and low coupling between the modules; simple liability; independent threads to simultaneously perform different functions; allows programmers to focus on the specific parts of your problem; and adaptive and decision-making capacity. The Implementation_JADE pattern [30] encapsulates the processes and functionalities that can be implemented with the JADE platform and implements and solves those functionalities that are essential so that developers can make use of this agent technology more easily.

TABLE II. SOCIAL ACTORS IDENTIFIED IN LATE REQUIREMENTS

Social actor	Description
MUA	Mail User Agent is a client application that allows to the users manages emails.
Spam Filter	A server program to recognize spam so as to prevent its delivery.
Classifier	A system that classifies emails at MUA and server level.
Preprocesser	A system that preprocesses emails to extract features.
Features Selector	A system that selects the most relevant features.

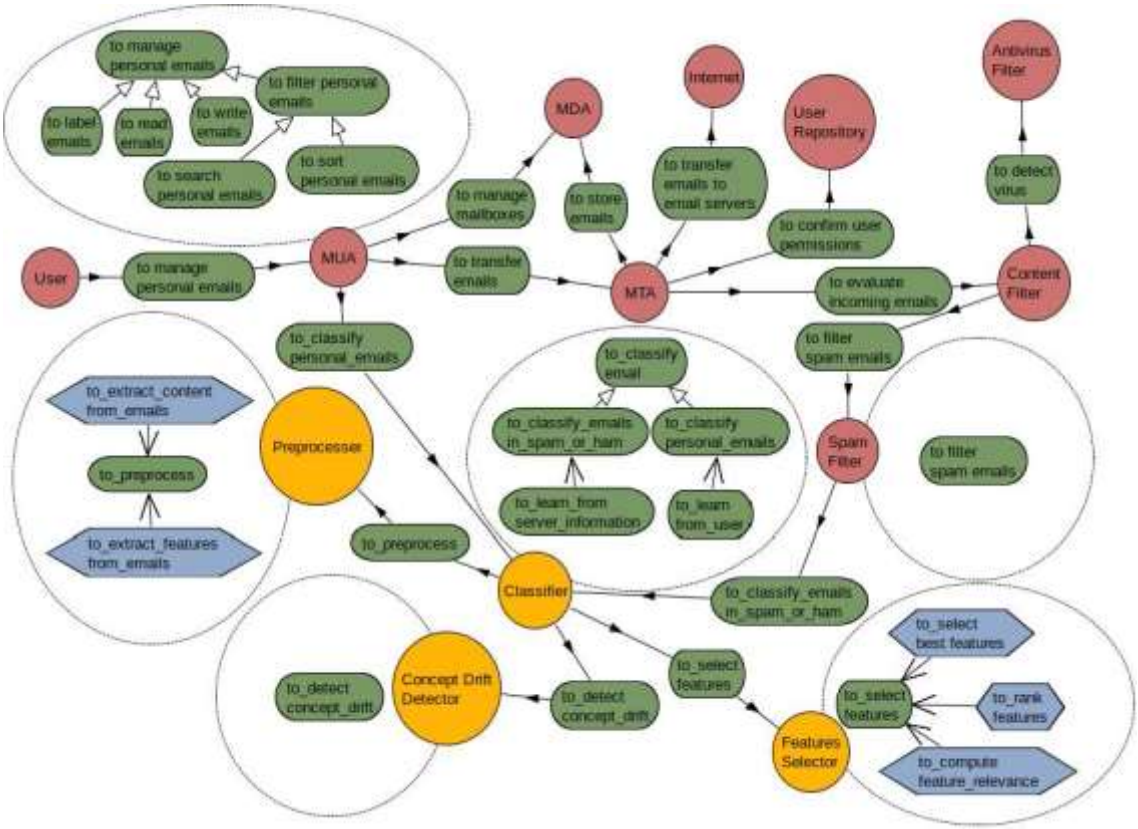


Fig. 3. SD and SR Models of Late Requirements.

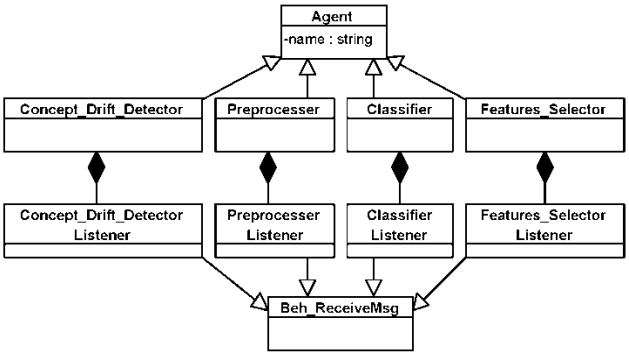


Fig. 4. Class Diagram of Multi-Agent System.

Fig. 4 represents the classes that make up the MAS. The Classifier, Concept_Drift_Detector, Features_Selector, and Preprocesser classes represent agents, while the listener classes represent the fundamental behavior of each agent. In the Listeners the message will be received and later an action will be executed according to what was received in the message or another behavior could also be executed. This implementation makes it easier for agents to add new behaviors to subsequent versions and to fully exploit their capabilities in the system.

IV. Evaluation of Classifier with Multi-Agent System Approach

To evaluate the solution, functional, efficacy and performance tests were carried out, taking as a reference the base spam classifier for the proposed MAS. The functional and efficacy tests consist of classifying 35 emails with the reference classifier and by the proposed MAS, with the aim of verifying that the results of both classifications coincide and therefore that the MAS works correctly. The results for the existed classifier (Classifier) and the proposed MAS (MAS), with respect to accuracy, True Positive Rate (TPR) and False Positive Rate (FPR), are shown in Fig. 5. The coincidence of the results both systems show the correct functioning of the proposed MAS.

Moreover, performance tests consist of obtaining the times in milliseconds (ms) that it takes to classify different amounts of emails, the reference classifier and the proposed MAS, in order to see how the processing time behaves. The results of these tests are shown in Fig. 6 and a decrease in the processing time is evident by the proposed MAS with respect to the reference classifier. This decrease in time is due to the use of multi-threads incorporated by JADE.

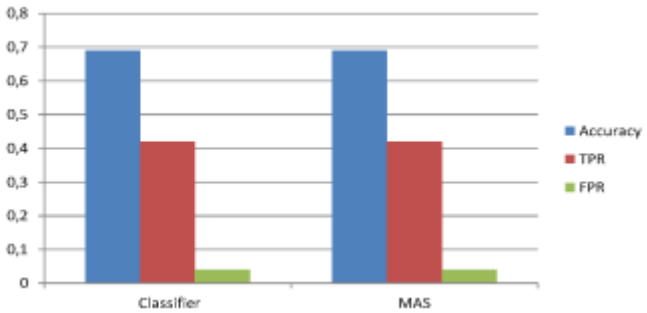


Fig. 5. Efficiency Test Results.

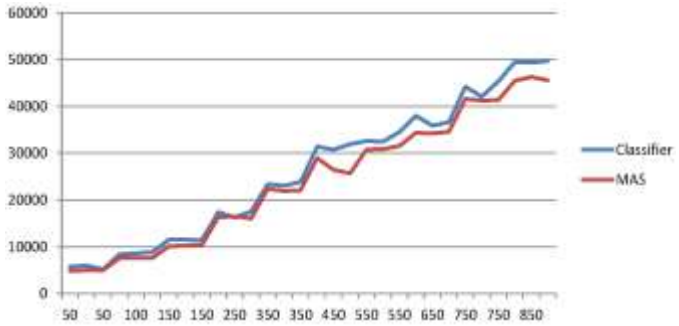


Fig. 6. Performance Tests Results.

V. Conclusion and Future Work

With aspects of social modeling and the language i* was representing actors, goals, tasks, resources and dependency-relations existing among actors of the email environment and designing an email stream classifier by following a goal-oriented approach. The email stream classifier was implemented with a multi-agent system approach. This will allows establishing bases to future achieve reactive, proactive and social behaviors, which allows the classifier to increase his adaptability. The results obtained in functional, efficacy and performance tests were satisfactory. The performance of multi-agent approach was better than the existing classifier due to the use of multi-threads incorporated by JADE. In future work, it is recommended to incorporate behaviors into the multi-agent system of existing solutions with better results and carry out a more exhaustive evaluation of the classifier.

Acknowledgment

This work was performed as part of the Smart University Project financed by the University of Alicante.

References

[1] Mujtaba, G., Shuib, L., Raj, R. G., Majeed, N., & Al-Garadi, M. A. (2017). Email classification research trends: Review and open issues. IEEE Access, Vol. 5, pp. 9044-9064.

[2] Team, R. (2020). Email statistics report, 2020-2024. Technical report, The Radicati Group, Inc. Palo Alto, CA, USA.

[3] Bhowmick, A. & Hazarika, S. M. (2018). E-mail spam filtering: A review of techniques and trends. Kalam, A., Das, S., & Sharma, K., editors, Advances in Electronics, Communication and Computing, Springer Singapore, Singapore, pp. 583-590.

[4] Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, Vol. 5, pp. 1–23.

[5] Luh, R., Marschalek, S., Kaiser, M., Janicke, H., & Schrittwieser, S. (2017). Semantics-aware detection of targeted attacks: a survey. Journal of Computer Virology and Hacking Techniques, 13(1), 47-85.

[6] Sanghani, G. & Kotecha, K. (2019). Incremental personalized e-mail spam filter using novel tfdcr feature selection with dynamic feature update. Expert Systems With Applications, Vol. 115, pp. 287-299.

[7] Rao, J. M. & Reiley, D. H. (2012). The economics of spam. Journal of Economic Perspectives, Vol. 26, pp. 87-110.

[8] Hussain, N., Turab Mirza, H., Rasool, G., Hussain, I., & Kaleem, M. (2019). Spam review detection techniques: A systematic literature review. Applied Sciences, 9(5), 987.

[9] Asdaghi, F., & Soleimani, A. (2019). An effective feature selection method for web spam detection. Knowledge-Based Systems, 166, 198-206.

[10] Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. Journal of Big Data, 2(1), 23.

[11] Nandhini, S., & KS, J. M. (2020, February). Performance Evaluation of Machine Learning Algorithms for Email Spam Detection. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-4). IEEE.

[12] Bahgat, E. M., Rady, S., Gad, W., & Moawad, I. F. (2018). Efficient email classification approach based on semantic methods. Ain Shams Engineering Journal, Vol. 9, pp. 3259-3269.

[13] Barddal, J. P., Gomes, H. M., Enembreck, F., & Pfahringer, B. (2017). A survey on feature drift adaptation: Definition, benchmark, challenges and future directions. Journal of Systems and Software, Vol. 127, pp. 278-294.

[14] Diale, M., Celik, T., & Van-Der-Walt, C. (2019). Unsupervised feature learning for spam email filtering. Computers and Electrical Engineering, Vol. 74, pp. 89-104.

[15] Méndez, J. R., Cotos-Yañez, T. R., & Ruano-Ordás, D. (2019). A new semantic-based feature selection method for spam filtering. Applied Soft Computing Journal, Vol. 76, pp. 89-104.

[16] Tanenbaum, A. S., & Wetherall, D. J. (2011). Computer networks fifth edition. In Pearson Education, Inc.

[17] Idris, I., Selamat, A., Nguyen, N. T., Omatu, S., Krejcar, O., Kuca, K., & Penhaker, M. (2015). A combined negative selection algorithm–particle swarm optimization for an email spam detection system. Engineering Applications of Artificial Intelligence, 39, 33-44.

[18] Jacobson, I., Booch, G., & Rumbaugh, J. (2012). The Unified Software Development Process. Prentice Hall.

[19] Pressman, R. S. (2010). Software engineering: a practitioner's approach. McGrawHill Higher Education.

[20] Haidar, H., Kolp, M., & Wautelet, Y. (2017). Goal-oriented requirement engineering for agile software product lines: an overview. Louvain School of Management Research Institute Working Paper Series, Louvain, Belgium, 1-36.

[21] Yu, E., Giorgini, P., Maiden, N., & Mylopoulos, J. (2011). Social Modeling for Requirements Engineering. The MIT Press.

[22] Horkoff, J., Aydemir, F. B., Cardoso, E., Li, T., Maté, A., Paja, E., ... & Giorgini, P. (2019). Goal-oriented requirements engineering: an extended systematic mapping study. Requirements Engineering, 24(2), 133-160.

[23] Cailliau, A., & Van Lamsweerde, A. (2012, September). A probabilistic framework for goal-oriented risk analysis. In 2012 20th IEEE International Requirements Engineering Conference (RE) (pp. 201-210). IEEE.

[24] Eric, S. Y. (2009). Social Modeling and i. In Conceptual modeling: Foundations and applications (pp. 99-121). Springer, Berlin, Heidelberg.

[25] AlhajHassan, S., Odeh, M., & Green, S. (2016, October). Aligning systems of systems engineering with goal-oriented approaches using the i∗ framework. In 2016 IEEE International Symposium on Systems Engineering (ISSE) (pp. 1-7). IEEE.

[26] Danesh, M. H., & Yu, E. (2014, June). Modeling enterprise capabilities with i*: reasoning on alternatives. In International Conference on Advanced Information Systems Engineering (pp. 112-123). Springer, Cham.

[27] Bertolini, D., Novikau, A., Susi, A., & Perini, A. (2006). Taom4e: an eclipse ready tool for agent-oriented modeling. issue on the development process. Technical report, Fondazione Bruno Kessler-irst.

[28] Jennings, N. (2000). On agent-based software engineering. Artificial Intelligence, Vol. 117, No. 2, pp. 277-296.

[29] Bellifemine, F. L., Caire, G., & Greenwood, D. (2007). Developing Multi-Agent Systems with JADE. Wiley.

[30] Moreno-Espino, M., Carrasco-Bustamante, A., Rosete-Suárez, A., & Delgado-Dapena, M. D. (2013). Patrones de implementación para incluir comportamientos proactivos. Polibits, Vol. 47, pp. 75-88.

[31] Khamis, M. A., & Nagi, K. (2013). Designing multi-agent unit tests using systematic test design patterns-(extended version). Engineering Applications of Artificial Intelligence, 26(9), 2128-2142.

Problem based Learning: An Experience of Evaluation based on Indicators, Case of Electronic Business in Professional Career of Systems Engineering

César Baluarte-Araya, Ernesto Suarez-Lopez, Oscar Ramirez-Valdez
Academic Department of Systems and Informatics Engineering
Universidad Nacional de San Agustín de Arequipa, Arequipa, Perú

***Abstract*—It is a reality that universities place great emphasis on formative research in the training of their students in order to increase their knowledge, skills, attitudes and achieve competences. This paper aims to show the experience of applying the Problem-Based Learning (PBL) methodology to assess learning based on indicators that have been determined from the criteria considered to correspond to the competences of the course under study Business Electronic, at the Professional School of Systems Engineering (EPIS) of the Universidad Nacional de San Agustín de Arequipa (UNSA), Arequipa-Peru, with theory and laboratory practice taught by two teachers. The objective is to apply an evaluation strategy for the development of competences with active didactics to the engineering training course. The methodology used is Problem-Based Learning applied to a formative research project based on real problems common to many organizations. In the semester, the students in groups solve the problems stated, and then they deliver a deliverable report and formative research report of each problem that is scored through a rubric. The teachers make contributions and provide feedback in the report for the improvement and experience that the student is acquiring in their training. The results obtained show that the objectives are achieved, increasing knowledge, skills, attitudes and adequate evaluation in the training of students, as well as the development of the competences of the course, as well as achieving the results of the student; showing that the application of PBL with Formative Research would provide good results for other courses of the professional career, allowing continuous improvement in the teaching-learning process. The conclusion is that an adequate assessment of learning based on indicators with an active didactic strategy, effectively planned and adequately applied to real-life problems, makes it possible to achieve the expected student results.**

Keywords*—Problem-based learning; formative research; competences; evaluation; performance indicator; skills; deliverable report

I. Introduction

At the university higher education level, universities try to achieve accreditation to guarantee the quality of the service offered. In Peru, accreditation is available through the National System of Evaluation, Accreditation and Certification of Educational Quality (SINEACE) [1] for all areas of knowledge; the area of engineering tries to adapt to other international standards such as [2] and others recognised by SINEACE such as [3], [4], [5].

Universities are applying formative research in order to allow their students to achieve a suitable level of research in their professional training as part of their competences to increase their knowledge and thus apply it to solving problems in reality. Thus, the Universidad Nacional de San Agustín de Arequipa [6], Arequipa - Peru, its Educational Model is based on the competency-based model for the professional training of students, and take into account that Professional Schools, such as the Professional School of Systems Engineering [7], teachers must include Formative Research in the syllabus of the course, in order to eliminate gaps in knowledge, skills and attitudes in students, and thus be able to use didactic strategies [8],[9],[10] adapting them to the nature of the course in order to achieve the training objectives and competences.

Since 2019, the EPIS has begun the accreditation process through the Accreditation Board for Engineering and Technology (ABET). And the stage of self-assessment was completed in 2020. Consequently, the authors of this article proposed to apply active teaching strategies as an action of continuous improvement in the development of their subjects. It is necessary for the student to develop greater knowledge, other skills, attitudes and training competences of the courses, as well as to achieve a higher level of research. The experience is developed in the subject Electronic Business (NE) which corresponds to the VIII semester of the Study Plan.

The objective pursued is to apply the evaluation based on indicators to each Deliverable Report and Formative Research Report, by using the PBL as an active strategy in the solution of laboratory problems in the NE course, by the working groups formed in the course within the teaching-learning process.

The research carried out is descriptive and the methodology is based on the stages of the scientific method for problem solving [11], which states that the student has access to develop skills, as we would say greater knowledge, procedures and values.

As a result of working in groups and applying PBL, the students developed each of the problems posed. The students

recognise that they have been evaluated in each report in the fairest and most objective way by the indicators that make up each criterion and the criteria that make up each competence of the course.

The conclusions reached are that the Problem-Based Learning methodology is appropriate for its application because of the nature of the course, which allows autonomous learning, increases knowledge of the problem area, strengthens the development of skills and attitudes, enables students to perform well, and achieves the expected student outcomes of the course; it also strengthens research training.

The article is organized in the Section II of Related Works - Context of the Experience, in Section III the Methodology is treated, in Section IV the Design of the Project of PBL and Formative Research is shown, in Section V the Method of Work is shown, in Section VI the Results of the work are shown, in Section VII the Discussion that is made is touched, that they worked or investigated others and of the result that is obtained that contributes in the formation of the student, in Section VIII the conclusions of the work are shown, in Section IX the future works that can be carried out are shown.

II. Related Work: Context of the Experience

A. Challenges in the Current Situation

The author of the present work; as well as the teachers who develop the courses with him; set out to accept the challenge of formally applying the Formative Research together with the Active Didactic Strategies from 2019 in the courses that it is his turn to develop in the EPIS.

In the current situation the evaluation of work or projects is done based on the criteria that make up the evaluation rubric, not performing a more detailed, real, fair, and evolution and monitoring that should be done to each student or group of students work that allows a better feedback to them of the deliverable reports for continuous improvement. Having thus the problem defined and to be treated for its solution in the present work.

B. Formative Research

In this regard [12] states that formative research nurtures research by generating future researchers, and defines it [13] as the pedagogical strategies carried out to train in and for research based on the appropriate activities of the scientific method, and the opinion that focuses on the strategy of learning by discovery and its construction [14], with the student at the centre of the process [15], with the teacher playing the role of advisor in the demand for and rigorousness of research.

On this subject [13] emphasises that the pedagogical function focuses attention on Formative Research, which is derived from the university's mission to generate both theoretical knowledge and knowledge on the application of knowledge to solve problems of reality. The problem of the teaching-research relationship must be addressed, contemplating the role that research must play in the learning process. It also reflects that Formative Research is linked to the concept of training, placing us in the field of teaching strategies that of shaping, structuring something that is built throughout the process, which in this case refers to the training of students who are trained through the various activities developed to understand and advance through Formative Research in deepening scientific research in their professional training. In the educational process of formative research, the actors directly involved are teachers and students.

We have previous experience in the EPIS in a first level of formative research in a speciality course in the courses of Writing Research Articles and Reports (RAII), Research Methods and Writing, (MIR), having as deliverables the elaboration of posters and documentary research articles, reports of results, and Formative Research Report; this allows to develop in the student the acquisition of new knowledge, skills and training competences.

Thus, in order to achieve a higher level of research, it is determined that in the NE course taught in the 8th semester of the Study Plan, the Problem-Based Learning (PBL) methodology is applied to deal with the different problems close to reality and having as a result the Deliverable and Formative Research Report of the problems solved by each working group of students.

For the present work, the NE course is taken as a case study, applying PBL in the laboratory sessions using techniques, methodologies and tools for the resolution of each problem dealt with, producing the respective Deliverable Report and Formative Research Report using the established template, which will be evaluated with the relevant rubric.

In Formative Research, by applying active learning strategies, learning is participatory, student-centred and employing active methodologies such as PBL. The author in [16] considers in Formative Research Management establishing mechanisms for integrating the research process into the teaching-learning process by designing methodological mechanisms. Likewise [17] analyses the role of Formative Research in the development of undergraduate scientific competences, how students perceive the contribution and impact of the professional training process in the development of their research competences. It is worth mentioning that since 2014 there have been the first results of the experience in the EPIS of the start in Formative Research with [18-19] evidencing and validating in [20] the achievement of the students of delivering as a result articles or posters of documentary research in events, of managing to develop the soft skills and also the competences of the course.

C. Student Competence Development

A curriculum based on competences implies that these are adequately developed in the students, considering that the advantages of PBL related to the acquisition, development and duration of competences by students trained with the methodology as referred to by [21] for different degrees or contexts are well known, which we would say we contemplate in Table I for better visualisation.

Very interesting is what is referred to by [22] when quoting (Vargas, S. 2010) where "competences, in addition to intellectual aspects, incorporate skills and attitudes ...", the aim is to master knowledge focused in a global and integral way of the person.

TABLE I. COMPETENCIES ADDRESSED BY RESEARCHERS

Competencies	Researcher(s)	Year
Development of critical thinking skills	Koh, Khoo, Wong y Koh Şendağa y Odabaşi Tiwari, Lai, So y Yuen	2008 2009 2006
The highest degree of self-efficacy	Rajab	2007
Self-perceived competencies	Cohen-Schotanus, Muijtjens, Schönrock-Adema, Geertsma y van der Vleuten	2008
Cognitive and social competencies	Koh et al.	2008
Intrinsic goal orientation Assessment of the tasks Processing strategies Metacognitive self-regulation Regulation of effort Peer learning	Sungur y Tekkaya	2006
Long-term impact of measures on competencies acquisition	Schmidt, Vermeulen y van der Molen	2006
Clinical reasoning	Scaffa y Wooster	2004

Source: Own elaboration taken from [21].

In other words, competence involves the development of knowledge, skills and attitudes in a specific area of knowledge, which, with the application of PBL, increases procedural skills, valuing achievements, learning from experiences when carrying out research [25] in the students' professional training.

Furthermore, as proposed by [23] in a methodology with control points and evaluation mechanisms, the results obtained demonstrate progress in the competences for solving real-life problems, managing projects, determining and discovering shortcomings, and allowing weaknesses in the teaching process of future engineering professionals to be corrected; noting that better results can be obtained throughout the curriculum with the support of the teachers.

D. Problem based Learning

In [20] referring to [2] who mentions that future graduates must be able to work in a global context, in multidisciplinary teams, solve problems and that what is learned is constantly changing, therefore, their training requires constant updating.

An appropriate concept contemplated by [24] when referring to Barrows (1986) who defines PBL as "a learning method based on the principle of using problems as a starting point for the acquisition and integration of new knowledge".

PBL as a didactic strategy used in universities is widely accepted in the area of engineering, it advocates cooperative or collaborative activities of active interaction between students and the student with the teacher who acts as a facilitator of learning; working in small groups of 4 to 8 members, oriented to the student who is the centre of attention of the strategy [14] [25], dealing with problems of reality in a systematic way seeking their solution through analysis, observation and evaluation, strengthening their knowledge of the area being treated.

PBL was born in the mid-1960s [23] for medical professionals to develop competencies and skills to master real-world problems. There are different methods as referred to in [14]. Morales and Landa [24] consider the following steps to be followed for its application: (a) analyse the problem scenario, (b) brainstorm, (c) list what is known, (d) list what is not known, (e) list what needs to be done to solve the problem, (f) define the problem, (g) obtain and process information and (h) present results.

There are many experiences related to the application of PBL in different areas, such as [15] in the subject of systems analysis, in basic sciences such as Physics [26], in Engineering such as Chemistry [27]. Experiences of applying PBL, such as those of [28] at the University of Murcia, Spain, within the process of convergence towards a European Higher Education Area (EHEA), as well as [29] in Computer Science, and [30] in the area of Mathematical Sciences.

In order to work with PBL, we have the approach of [24] and [31] that propose the steps for students to work with PBL, providing activities that can be evaluated, such as: team and individual work, written report, individual and team contribution, self-evaluation.

PBL has fundamental characteristics that come from the model developed at McMaster as referred to in [24] which among others are:

- Learning is student-centred.
- Learning takes place in small groups of students.
- Teachers are facilitators or guides.
- Problems form the focus of organisation and stimulus for learning.
- Problems are a vehicle for the development of problem-solving skills.
- New information is acquired through self-directed learning.

Producing significant changes in students by promoting, developing collaborative work culture, interpersonal skills, student participation, valuing teamwork, interdisciplinary work, problem identification skills, critical thinking, formative assessment.

E. Assessment for Learning

In this regard, there are the experiments referred to in [21] by the studies carried out by (De Grave, Boshuizen and Schmidt, 1996; De Grave, Schmidt and Boshuizen, 2001) where they refer that students obtain better scores when they analyse the problem before gathering information, as opposed to those who only do so on the basis of the material provided.

It is convenient to refer to [22] when he states that learning assessment is an integral and continuous process of diagnosis and training, with criticality, retrospection, introspection and projection of the learning process; considering the basic questions of What, How, When, When, Who to assess, which allows determining the type of assessment according to the context where it is developed. As a reference, we also have the contribution of [32], whose questions break down the basic questions into sub-subheadings.

Table II shows the basic criteria for quality evaluation from the point of view of [33].

TABLE II. CRITERIA FOR QUALITY ASSESSMENT

Criteria	It means to the students
Transparency	Clarity of criteria and levels of demand, allows him to orient his work, self-evaluate his own learning pace.
Validity	It assures you that the assessment system assesses the learning outcomes it is supposed to assess and not others.
Fiability	It is not enough to value what must be valued, but it must also be done in the right way.
Democratic	Participate actively in the teaching-learning process.
Global	That it is an essential part of their learning process, integrated into the curriculum and a task that is continuous and not just for a special moment.
Formative	It will serve you as motivation, orientation of your work pace.

Source: Available at [37].

In this regard, it is important to note what [2] says from the point of view of programme accreditation "Assessment of student learning, with a focus on continuous improvement, is key to ensuring the quality of our educational programs and preparing our graduates to enter a global workforce. The cumulative result of student learning in our curricula and co-curricular activities enables the career and professional accomplishments of our graduates. In an era of accountability and transparency, outcomes assessment has become an international standard of quality".

Bearing in mind that EPIS teachers are immersed in the application of active didactic strategies to assess student results that are immersed in the ABET accreditation model for the development of student competences.

In the evaluation of learning when applying PBL in the teaching-learning process by competences, instruments are applied according to the context of the reality of the problems to be solved, in the present work the rubric [34, 35, 36, 37] was used to evaluate through indicators that determine in the results the achievement of knowledge, the level of development of competences, attitudes, using rating scales, in an objective and consistent way of the activities.

The author's knowledge in active didactic learning strategies is given by the constant study, adaptation and application of changes in the study materials, the forms of course development and the forms of evaluation that lead to demonstrate how the student achieves the expected results and develops the competencies of both the course and his professional training profile, and not only achieve an overall grade to pass a course. Taking into account that the results of its application are a contribution for the researchers of the scientific community, and through its analysis allows continuous improvement within the teaching-learning process.

III. METHODOLOGY

The methodological design used is quasi-experimental, not working with a control group. The research is applied to solve the problem of the qualification of the results through the Deliverable Report and Formative Research Report. The research developed uses the PBL methodology, the rubric instrument, the scoring tool for the indicators of each criterion in each subject competence.

There are many definitions for indicators. UNE 66.175 (2003) defines it as "Data or set of data that help to objectively measure the evolution of a process or an activity".

IV. DESIGN OF THE PLB AND FORMATIVE RESEARCH

The main characteristic of PBL is the use of a set of problems that are developed and given to students in small groups to promote learning, so the subject was adapted to the methodological principles of PBL, contemplating the application of the evaluation system based on indicators from [38].

The design of the project considered aspects such as relevance, dealing with real-life problems, helping students to increase their knowledge, applying methodologies, methods, techniques and tools related to the course, including formative research activities, allowing students to prepare structured reports for objective and fair evaluation.

In this regard, in [23] it is stated what should be taken into account in the design of the PBL, and most authors agree that there are a series of basic steps that may vary depending on: a) number of students, b) time available, c) objectives to be achieved, d) available bibliography, e) available resources of the teacher and the entity.

A. The Project

The project to be developed has the following characteristics:

- It contains problems of real-life organisations.
- The work team is made up of 4 or 3 students.
- The problems will be developed throughout the academic semester.

B. The Problem

According to [14], a good problem formulation or approach should consider 3 variables: a) Relevance, so that students understand the importance of the problem, b) Coverage, to guide students to search for, discover and analyse the information on the topic of study, c) Complexity, as there is no single solution, several hypotheses should be put forward, testing and documenting them.

A series of problems were adapted in the course to respond to the objectives, the development of the contents of each subject, the development of competences, the reinforcement of skills and the achievement of student results.

Considering the different scenarios of the type of real organisations, both business and governmental, where the student assumes the different roles when participating in the work team to develop the problems.

The characteristics of well-structured problems are referred to by [38] citing Romero M. and J. García-Sevilla (2008) and contemplated by citing (Ching and Chia, 2005), which are very valuable in their contribution.

C. Project Problem Issues

The topics to be covered in the problems are:

- Virtual stores.
- Customer Relationship Management CRM.
- Supply Chain Management SCM.
- E Marketplace.
- E Learning.
- E Employee.
- E Government.
- M Business.

The development of the problems in the laboratory sessions of the NE course was determined either by groups of students made up of 4 or 3 members, taking into account in its conformation: the willingness to meet, communication, work schedules for the sessions, taking into account the other courses of study of the team members, coincidence of personal interrelation; which gives very good results.

D. Tracking

In the development of the course during the semester, there are sessions for the delivery of the Deliverable Report and Formative Research, where the result of the problem, its exposition and evaluation are reflected, providing feedback so that the team of students can make corrections and take into account the opinions and suggestions that will be useful for the development of future problems.

The role played by the teacher is that of a guide by giving guidelines, suggestions, orienting, suggesting sources of information and being willing to collaborate with the needs of the students in the team.

E. Evaluation

If we start from the concept and function of an indicator, which normally occurs at the level of an institution, as contemplated by [39], and if we adapt and apply them to the level of a subject performance indicator that allows us to evaluate the problems posed to students.

The most general characteristics of these indicators are: objective, relevant, useful, precise, congruent, contextualised; that they provide data to generate information, make the changes to which they give rise and be useful for reporting on the changes that arise between inputs and results, point out the strong and weak points in the criteria, compare the results of the most and least successful experiences shown in each Deliverable Report, analyse the results for changes in the development of the course in future versions, discover other aspects related to the actions of the students that improve the application of PBL to the course.

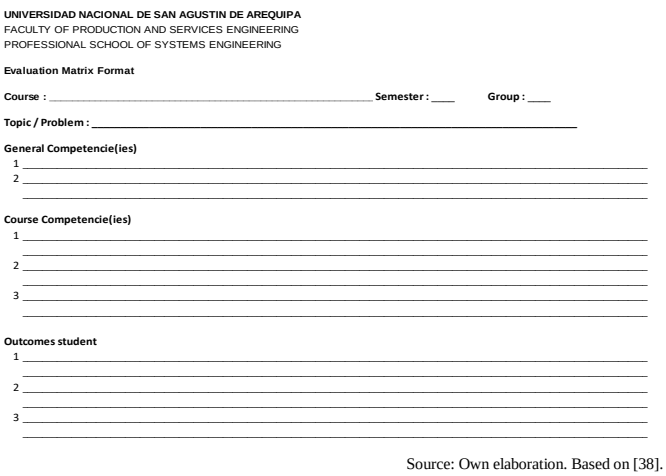
In this case, the indicators will be useful when they are used to evaluate the results of the students' work with the aim of improving the teaching process and therefore the development of their competences.

For the year 2020, based on the Educational Model for competence-based vocational training, assessment is envisaged for the area of engineering using the rubric as an instrument for assessing competences and student results in the NE course. Considering that with the rubric a better result is possible, achieving the objectives and measuring performance. Taking into account what is stated by [34] as a point to carry out learning and assessment work, and also by [35] that provides the provision for the student to be constant as to how far to go with their learning of the topics and what would be their desired maximum level; with the purpose of:

- Achieving the student outcomes determined for the subject.
- Achieve the development of the competences defined for the subject.
- Strengthen the development of soft skills, attitudes and values in the student.
- Establish achievable goals, deliverable reports and formative research.
- Develop the problems by applying the PBL established for the subjects.
- Follow-up and feedback by the teacher from the Deliverables Reports and Formative Research in the evaluations of the same for continuous improvement.
- Increase the interpersonal and social relations of the members of the team or working group.
- Manage the activities and time for the development of the problem, obtaining the results, the elaboration of the Deliverable.

Fig. 1 shows the format of the evaluation matrix in its first part of the header.

Fig. 2 shows the second part of the evaluation matrix, the detail of the format, which is the basis for the elaboration of the evaluation matrix in an electronic spreadsheet.



UNIVERSIDAD NACIONAL DE SAN AGUSTIN DE AREQUIPA
FACULTY OF PRODUCTION AND SERVICES ENGINEERING
PROFESSIONAL SCHOOL OF SYSTEMS ENGINEERING

Evaluation Matrix Format

Course : _____ **Semester :** ___ **Group :** ___

Topic / Problem : _____

General Competencie(ies)

1 _____
2 _____
_____

Course Competencie(ies)

1 _____
_____
2 _____
_____
3 _____
_____

Outcomes student

1 _____
_____
2 _____
_____
3 _____
_____

Source: Own elaboration. Based on [38].

Fig. 1. Header of the Evaluation Matrix Format.

Source: Own elaboration. Based on [38].

Fig. 2. Detail of the Evaluation Matrix Format.

F. Grading

For the grading of the Deliverables and Formative Research Reports, the indicator-based evaluation system proposed by [38] was followed. The initial structure was adapted to include the element "student outcomes", which is considered important and better complements its definition, structure and application to the evaluation of problems or projects; namely: a) General Competence(s) (from the graduate profile), b) Course Competence(s), c) Student Outcomes, d) Process(es), e) Criterion(s), f) Performance Indicator(s), g) Indicator(s) Rating Scale; having that the rating scales are applied according to the nature of the indicator in each criterion and each competence involved in the subject.

The Grading Scales Table for Assessment proposed in [38] is used, in input scales 5 or 6 values can be adopted according to the determined grading level, which is shown in Fig. 3.

G. Conversion - Rubric Grading Scale Equivalence

Converts from the grading scale to the scale of the rubric established to be reflected in the student's results grade and the semester evaluation periods.

UNIVERSIDAD NACIONAL DE SAN AGUSTIN DE AREQUIPA
FACULTY OF PRODUCTION AND SERVICES ENGINEERING
PROFESSIONAL SCHOOL OF SYSTEMS ENGINEERING

Table of Qualification Scales for Evaluation

No. Entries						
2	Not present	Present				
	0	1				
3	Not present	Regular	Good			
	0	1	2			
4	Not present	Regular	Good	Very Good		
	Not present	Regular	Good	Excellent		
	0	1	2	3		
5	Not present	Regular	Good	Very Good	Excellent	
	Not present	Insufficient	Regular	Good	Excellent	
	0	1	2	3	4	
	0	2	6	8	10	
	0	5	10	15	20	
6	Not present	Insufficient	Regular	Good	Very Good	Excellent
	0	1	2	3	4	5
	0	2	4	6	8	10
	0	4	8	12	16	20

Note: Qualification scales can be assembled according to the nature of the criterion or indicator.

Source: [38].

Fig. 3. Table of Qualification Scales for Evaluation.

V. Method of Work

A. Conceptual Design

There is a good experience of one of the authors of this work [40] in applying Project Based Learning that serves as a reference to apply active strategies in the teaching-learning process in the NE course in the professional career of Systems Engineering.

The purpose is for the student to research, carry out searches, review analogous scenarios, examine diverse literature, select, organise, analyse, interpret data and information, and be able to propose alternative solutions to the real problem posed in an organisation, using the PBL strategy to reinforce and discover greater knowledge, develop soft and procedural skills, and evaluate the results obtained.

The research for development is based on the problem-solving stages of the scientific method, which together with the active didactic strategy of PBL, allows the student to develop skills such as those discussed in [38] where good results were obtained.

In order to obtain data on the students' perception, the survey technique is applied and the questionnaire is used as an instrument, the results are systematised and analysed so that the conclusions can be used to take actions for the continuous improvement of the learning process.

The stages determined for the development of the project are:

Stages of the project

1. Starting point

   Theme:

   - Dealing with Reality Problems in Electronics Business.

   Initial Question

   - ¿From the knowledge obtained and with the knowledge acquired from e-Business, will it be possible to solve the problems posed from the real context in the organisations?

2. Formation of Collaborative Teams

   - Team/group of 4 or 3 students.

3. Definition of the Final Product

   Product to be developed

   - Solving real context problems in organisations.

   What you need to know (learning objectives)

   - Apply Problem Based Learning to solve the problems posed.
   - To elaborate the report of each deliverable applying the template defined for it.
   - To carry out the presentation of the solution of the problem

4. Planning

- Determination of the competences of the course.
- Alignment of the competences to the learner outcomes.

General student competences:

C.x Competence r.
C.y Competence p.

Course competences Electronics Business:

a Competence a of the course.
b Competence b of the course.

Student outcomes to be assessed:

a) Criterion RE.7.2. Student outcome 7.2 of competence a.
b) Criterion RE.8.2. Student outcome 8.2 of competence b.

Competence Criteria and Indicators

Criterion 1 Indicator 1 Student Outcome 1.
Indicator 2 Student Outcome 1.
Criterion 2 Indicator 3 Student outcome 2.
Indicator n Student outcome m.

5. Organising Team/Group Work

Team or group work of 4 or 3 students.

6. Review / Development of Problems (Laboratory Guides)
7. Instruction in PBL methodology
In the first class session, it is explained.
8. Application of the PBL methodology

- Presentation of the problem situation.
- Laboratory guides to solve problems.
- Answering queries in the session by the means determined.
- Carrying out the activities and tasks of the problem.
- Presentation of deliverables.
- Oral presentation of results of the deliverable.

9. Elaboration of the Deliverable and Formative Research Report

Elaboration of the report according to the elaboration template.

10. Evaluation of the report and competences

- The teacher in the sessions evaluates the competences and the results of the student, to the group through the grading tool based on the rubric.
- Assessment of the Deliverable Report and Formative Research Report of each problem through the rubric-based grading tool.

11. Initial Evaluation of the Application of the PBL Methodology

At the conclusion of the first development phase of the semester.

12. Final Evaluation of the Application of the PBL methodology

At the end of the third phase of the semester.
Apply the questionnaire to receive the student's perceptions.

Methodologies, Techniques and Instruments to be Used

Methodologies, Techniques and Instruments

- Problem Based Learning Methodology - PBL.
  - Problem solving, in sessions established within the development of the course.
- Survey technique.
  - Questionnaire - of students' perception of PBL.
  - Questionnaire - students' perception of the application of PBL to the e-Business course. o Questionnaire - students' perception of the application of PBL to the Electronics Business course.
  - Questionnaire - on tutor's perception of the tutor's work.

Evidence

- Digital files of lab work.
- Moodle, Virtual Classroom as a repository of course work.

Assessment Instruments

- Written report grading tool.
- Assessment rubrics.
  - Evaluation of Competences and Student Outcomes.

Student self-assessment, by general appreciation.

B. Participants

The course of NE in the professional career of Systems Engineering is developed in the eighth semester (4th year), with 5 hours per week, 3 of them theoretical and 2 of laboratory, the semester is developed in 17 weeks; participating in the evaluation of the Deliverable Report and Formative Research, taking as a case the semester 2020 B with a total of 35 students; the theory and laboratory practices were in charge of two teachers, forming two groups of theory and ten laboratory subgroups.

C. Data Analysis Technique

From the evaluation of the Deliverable Report and Formative Research, they were graded with the grading tool generated for it contemplating the structure based on the

evaluation system based on indicators of [38], application of questionnaires to students to get appreciation data, systematization of data in the electronic spreadsheet EXCEL, analyzing the results achieving averages, tables, and graphs.

D. Instruments

The instruments used are:

- Laboratory Guides.
- Template for the elaboration of the Deliverable Report and Formative Research.
- Evaluation rubrics.
- Grading tool for the written report of the NE course.
- Student Achievement Appreciation Questionnaire.

E. Techniques

Evaluation techniques are used:

- The rubric.
- The Rating Scale.
- Survey.

F. Deliverables

The deliverables that correspond to each of the topics indicated in the point Project Problem Topics were determined, which are developed and delivered by each team or working group, and stored in the virtual classroom repository. The evaluation of which is graded with the grading tool and with immediate feedback in the report to strengthen the teaching-learning process, the development of competences and the achievement of student outcomes.

The structure of the Formative Inquiry and Deliverable Report is shown below:

Cover page

Index

1. Plan the treatment of the problem
2. Organising the work of the team/group
3. Problem
4. Theoretical framework
5. Comparative of the selection of aspects
6. Working in a group, collaboratively with colleagues and avoiding working alone
7. Generation of possible solutions (Alternatives)
8. Presentation of the solution
9. Prototype or Situational Analysis
10. Lessons Learned
11. Conclusions
12. References
13. Annexes
14. Report
15. Self-evaluation.

VI. Results

The results of the Evaluation of the Deliverable and Formative Research Report (EDRFR) are shown below; with a defined and used structure, based on the defined rubric of the NE course. A sample is shown in Fig. 4, where the general competences (of the student or of the graduate profile), the course competences related to the general ones, the criteria, the indicators of the criteria with the rating scale determined for each one of them, and their rating for the proposed problem of the working groups can be seen.

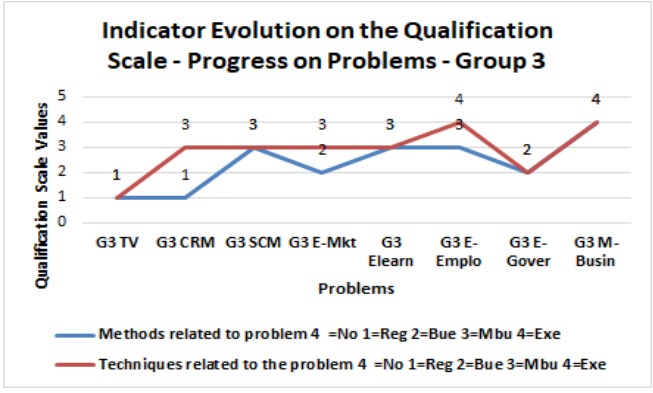
Fig. 5 shows the evolution of two of the 66 indicators in the group's progress in developing each problem during the semester.

Fig. 6 shows the evolution of the Students' Results related to the course competence a, of the General Competence r. Where it can be seen how groups 1, 2 and 4 have a growing tendency to obtain the maximum grade close to 20, and group 5 remains at a level with a grade value of 15 between 12 and 15.

Fig. 7 shows the group grades for the Student Outcomes related to competences a and b of the course, of the General Competences r and p. Here it can be seen that groups 1, 2 and 10 have an average grade of 18 on the vigesimal scale for competence a, and group 5 has an average grade of 14.

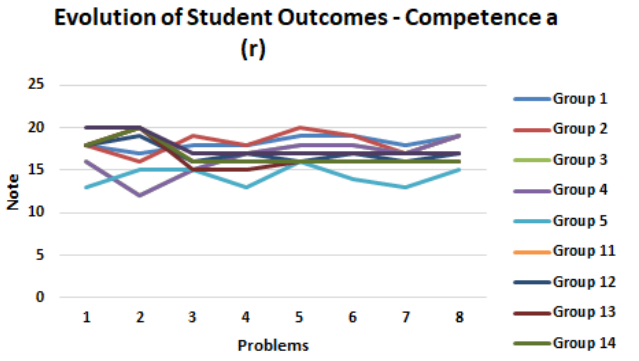
Source: Own elaboration.

Fig. 4. ADRFR – Report Qualification Tool.



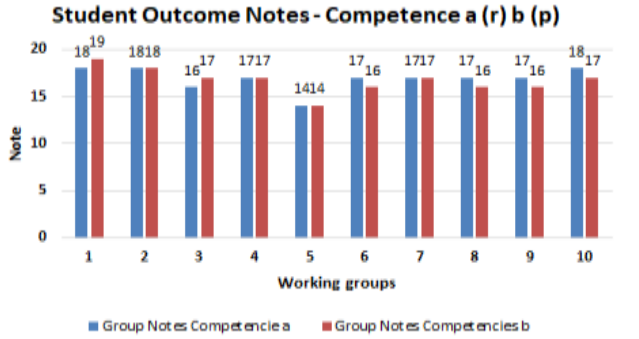
Source: Own elaboration.

Fig. 5. IEQS-PP-HG3 – Report Indicator Evolution.



Source: Own elaboration.

Fig. 6. ESO - Report Evolution Outcomes – Competence a.

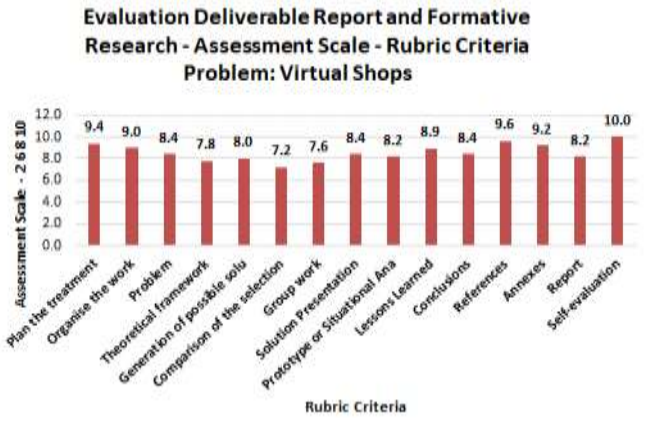


Source: Own elaboration.

Fig. 7. SON – Report Student Outcome Note – Competence a and b.

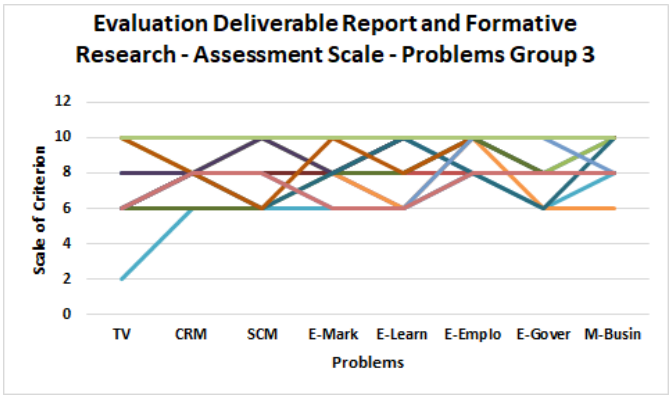
Fig. 8 shows the good results as an average of the evaluation of one problem: Virtual Shops, taking into account the levels of the assessment scale of the rubric criteria, which are also used for the other 7 problems.

Fig. 9 shows the evolution of working group 3 in the development of each of the problems during the academic semester.



Source: Own elaboration.

Fig. 8. EDRFR – Evaluation Deliverable Report and Formative Research-Assessment Scale-Rubric Criteria-Problem:Virtual Shops.



Source: Own elaboration.

Fig. 9. EDRFR – AS – PG3 - Evaluation Deliverable Report and Formative Research-Assessment Scale-Problems Group 3.

VII. Discussion

It is important in engineering education to consider the statement by [28] that "University education is the last educational step before professional practice..." Therefore, in order to exercise a profession, the graduate must be competent. This means that in order to solve real-life problems, he/she must possess knowledge, soft and procedural skills, ingenuity, among others. Being of the opinion that PBL and the assessment process greatly favour the development of complex competences.

The application of PBL was born for the development of competencies and skills in the training of professionals in the area of medicine for the mastery of real problems in the mid-1960s; however, it is still applied with relevant adaptations as discussed by [41] in their study in the field of Clinical Laboratory Medicine, where the problem is that there is a large gap between theoretical concepts and clinical laboratory practice, and in actual clinical practice. Faced with this they reorganized the cases of PBL in hematology studies, defining the processes of design, preparation, implementation and evaluation of PBL. Concluding that it incorporates a strong way to safeguard the gap between theory and practice, so students make good use of their medical science training toward solving complicated problems. The results show that PBL encouraged students to interact, pose questions, state hypotheses, find out answers, and evaluate their ability to prove their points.

Having that [42] in their work on formative research provide some reflections that are necessary to take into account such as: they seek that in their processes students become familiar with scientific and research culture; training them in research skills; interaction, dialogue and research practice are mandatory; formative research is positioned in academic and formative activities in universities. Concluding that formative research; referring to (Arakaki, 2009); is fundamental for the professional formation of the student with critical, reflective and proactive thinking, by stimulating the ability to raise and respond to unresolved problems or scenarios of the environment.

When referring [24] to teachers in higher education, very few have training in pedagogy and teach as they were taught; or as we say, take the best university lecturer as a reference point; and active methodologies are not applied and assessment is limited to checking the memorisation of facts and information. In view of this, in order to apply PBL, changes are required in the role of the teacher, who becomes a facilitator so that students build their knowledge and apply it in real situations.

Thus, [21] mentions that there is experimental verification when students start from the analysis of a problem prior to the documentary study, they achieve much higher marks in the learning evaluations, compared to what others do with the same material; in the present work the verification is made through the activities and the results obtained.

There are results on the development of research competences by applying research strategies from [17] in research training. The author in [13] points out the importance of formative research as a means of pedagogical function in the production of knowledge.

In [43], the result of the research is directed to university teachers who wish to apply didactic strategies that affect the book in explaining the didactic discipline, being the didactic strategies focused on meeting the objectives set in a teaching-learning context.

In [44] the application of PBL in the university level course they decided to rethink the way of working because the experience showed them that something was failing in the learning of students in the last 10 years, they proposed to change and go beyond a simple evaluation. This contributes to the fact that when faced with a similar situation, the problem should be analyzed and the feasibility of using didactic teaching and learning strategies applied to the context of reality should be assessed.

In the research work it has been achieved to apply PBL as an active didactic strategy in the development of the NE course that contributes to the integral formation of the student, achieving the development of competences, allowing a better development in their formation and future work performance.

Also the PBL besides being applied in courses of a curriculum is used in other areas that are related to professional training and where the student participates as being the experience of [45] concluding that students participate in the planning of projects, development of skills, and that the PBL bridges the gap between learning and application; and thus the student effectively faces real world problems. Having that in our country, Peru, with the University Law No. 30220 of 2014 universities are obliged to work the University Social Responsibility that contributes to sustainable development and the welfare of society.

As an experience we have the results of the evaluation grades of the last three years, which shows how the qualification from indicators allows a fairer and more objective evaluation and growth of each deliverable of the students, when applying PBL from 2019, which are shown in Table III for better visualisation.

TABLE III. GRADE POINT AVERAGE

Course	**Grade Point Average**		
	2018	***2019***	***2020***
Electronics Business Theme: Mobile Business	15.17	16.15	17.4

VIII. CONCLUSION

The following conclusions have been reached:

- The course objectives have been achieved.
- The student outcomes determined for the course are achieved.
- Students achieve the development of the competences of the course and therefore favour the achievement of the competences of the syllabus.
- Problem-Based Learning as an active methodology allows: enhancing autonomous learning, increasing knowledge on topics by understanding them, applying the knowledge obtained, motivation and teamwork, strengthening the development of procedural skills and abilities, improving students' academic performance.
- Formative research in the student's professional training process is transversal in the courses of the curriculum, as in the case of the Database and Electronic Business courses, where the teachers of this work intervene, achieving good results.
- The students manage to have the appropriate evaluation from the qualification of each indicator involved in each problem, being the most objective, real and fair, going from less to more.
- The students recognise that what they have achieved in the development of the problems in the course will serve as a guideline for them to develop in future organisations.
- Providing the students with feedback in the reports of deliverables reviewed, graded and observations made by the teacher, allows them to continuously improve in the work group.
- The student at the end of the course has been able to carry out documentary research on the topics of the theoretical framework to increase their knowledge, and to conveniently write the Deliverable Report and Formative Research.

IX. FUTURE WORK

- To carry out comparative research on the application of PBL in the courses of the Curriculum.
- Conduct comprehensive formative research in other Curriculum courses for the development of students' competences.

ACKNOWLEDGMENT

Our thanks to the Universidad Nacional de San Agustín de Arequipa for allowing and supporting the development of

Formative Research, strengthening the formative area and the application of proposals for continuous improvement based on active didactic strategies that will benefit the students.

REFERENCES

[1] SINEACE, Sistema Nacional de Evaluación, Acreditación y Certificación de la Calidad Educativa https://www.gob.pe/sineace/

[2] ABET. Why ABET Accreditation Matters. https://www.abet.org/accreditation/what-is-accreditation/why-abet-accreditation-matters/. Ultimo acceso junio 2021. https://www.abet.org/ assessment/

[3] ICACIT, Instituto de Calidad y Acreditación de Programas de Computación, Ingeniería, Tecnología https://www.icacit.org.pe/web/es/

[4] CNA, Consejo Nacional de Acreditación, Colombia. https://www.cna.gov.co/1741/articles-186359_pregrado_2013.pdf

[5] AcreditAcción, Agencia Acreditadora https://acreditaccion.cl/

[6] Universidad Nacional de San Agustín de Arequipa. http://www.unsa.edu.pe.

[7] Escuela Profesional de Ingeniería de Sistemas. http://www.episunsa.edu.pe.

[8] Subdirección de Currículum y Evaluación, Dirección de Desarrollo Académico, Vicerrectoría Académica de Pregrado, Universidad Tecnológica de Chile INACAP. (2017), pp. 21-24. Manual de Estrategias Didácticas: Orientaciones para su selección. Santiago, Chile: Ediciones INACAP.

[9] Rodriguez Cruz, Reyna, Compendio de estrategias bajo el enfoque por competencias, Instituto Tecnológico de Sonora ITESCA, México, 2007, pp. 26-29. http://www.itesca.edu.mx/documentos/desarrollo_academico/compendio_de_estrategias_didacticas.pdf.

[10] Ma. Cristina Sánchez Martínez, Marcos Aguilar Venegas, José Luis Martínez Durán and José Luis Sánchez Ríos, Estrategias didácticas en entornos de aprendizaje enriquecidos con tecnología (antes del Covid-19), UNIVERSIDAD AUTÓNOMA METROPOLITANA-XOCHIMILCO, México, 2020, pp. 11-15.

[11] GARZA-RIVERA, RG. El rol de la física en la formación del ingeniero. Ingenierías, 2001, vol. IV, No. 13, pp. 48-54.

[12] Patiño, R. A., Melgarejo, Z.A. and Valero G.M. (2019). Percepción de los egresados contables sobre la investigación formativa. Revista Activos, 16(30), 101-125. DOI: https://doi.org/10.15332/25005278.5062, Colombia.

[13] Restrepo Gómez, Bernardo, INVESTIGACIÓN FORMATIVA E INVESTIGACIÓN PRODUCTIVA DE CONOCIMIENTO EN LA UNIVERSIDAD, Nómadas (Col), núm. 18, mayo, 2003, pp. 195-202, ISSN: 0121-7550, Universidad Central, Bogotá, Colombia.

[14] Restrepo Gómez, Bernardo, Aprendizaje basado en problemas (ABP): una innovación didáctica para la enseñanza universitaria, Educación y Educadores, vol. 8, 2005, pp. 9-19, Universidad de La Sabana, Cundinamarca, Colombia. (DOI): http://dx.doi.org/10.18687/LACCEI2018.1.1.97

[15] Medina-Rojas, F., Nuñez-Santa, J.M., Sánchez-Medina, I.I. and Cabrera-Medina, J.M., Implementación del ABP, PBL y método SCRUM en cursos académicos para desarrollar sistemas informáticos enfocados en fortalecer la región, Revista Educación en Ingeniería, 12(24), pp. 52-57, Julio, 2017, Bogotá. ISSN 1900-8260 DOI: http://dx.doi.org/10.26507/rei.v12n24.758

[16] Mejía Murillo, Carmen, Manual de Procesos de Investigación Formativa, Universidad Herminio Valdizán, Perú, 2016, pp. 7-9.

[17] Pinto Santos, Alba and Cortés Peña, Omar, ¿Qué Piensan los Estudiantes Universitarios Frente a la Formación Investigativa?, REDU. Revista de Docencia Universitaria, 2017, 15(2), 57-75.

[18] Baluarte C., Vidal E., Delgado L. and Castro E.; Integrando Habilidades Blandas: Redacción, Comunicación y Ética en la Currícula de la Escuela Profesional de Ingeniería de Sistemas – UNSA; 15th LACCEI International Multi-Conference for Engineering, Education, and Technology: "Global Partnerships for Development and Engineering Education", 19-21 July 2017, Boca Raton Fl, United States. Digital Object Identifier (DOI): http://dx.doi.org/10.18687/LACCEI2017.1.1.141 ISBN: 978-0-9993443-0-9 ISSN: 2414-6390.

[19] Castro E., Vidal E. and Baluarte C., Integrando la Comprensión de la Responsabilidad Ética y Profesional en una Carrera de Ingeniería: Experiencia y Lecciones Aprendidas, 14th LACCEI International Multi-Conference for Engineering, Education, and Technology: "Engineering Innovations for Global Sustainability", 20-22 July 2016, San José, Costa Rica, RP139.

[20] Baluarte Araya, César., Vidal Duarte, Elizabeth and Castro Gutierrez, Eveling, Validación de las Habilidades Blandas en los cursos de la Currícula de la Escuela Profesional de Ingeniería de Sistemas-UNSA, 16th LACCEI International Multi-Conference for Engineering, Education, and Technology: "Innovation in Education and Inclusion", 19-21 July 2018, Lima, Perú. Digital Object Identifier (DOI): http://dx.doi.org/10.18687/LACCEI2018.1.1.97 ISBN: 978-0-9993443-1-6 ISSN: 2414-6390.

[21] Briones, Elena and Vera, Jesús, Aprendizaje Basado en Problemas (ABP): Percepción de carga de trabajo y satisfacción con la metodología, V Congreso Mundial de Estilos de Aprendizaje, Santander, España, 2012. https://dialnet.unirioja.es/servlet/articulo?codigo=4640627

[22] Acebedo, M.J. (2016). La evaluación del aprendizaje en la perspectiva de las competencias. Revista TEMAS, 3(11), pp. 203–226.

[23] Fernández, F.H and Duarte, J., El aprendizaje basado en problemas como estrategia para el desarrollo de competencias específicas en estudiantes de ingeniería, Formación Universitaria, 6(5), 2013, pp. 29-38.

[24] P. Morales and V. Landa, "Aprendizaje Basado en Problemas". Theoria, Vol. 13, 2004, pp. 145-157. [On line]. Disponible: http://biblioteca.udgvirtual.udg.mx/jspui/handle/123456789/574.

[25] C. Gamboa, "Apuntes sobre Investigación Formativa"; versión No. 2, Colombia, 2013. [On line]. Disponible: http://idead.ut.edu.co/Aplicativos/PortafoliosV2/Autoformacion/materiales/documentos/u2/Apuntes_sobre_investigacion_formativa.pdf.

[26] Espinoza Suarez, Silvia, Effectiveness Of Problem-Based Learning In The Academic Performance of course Physics I, 16th LACCEI International Multi-Conference for Engineering, Education, and Technology: "Innovation in Education and Inclusion", 19-21 July 2018, Lima, Perú. (DOI): http://dx.doi.org/10.18687/LACCEI2018.1.1.226.

[27] Cañas Cano María, Aprendizaje Basado en Problemas (ABP), competencias y la enseñanza de química para Ingenieros, 16th LACCEI International Multi-Conference for Engineering, Education, and Technology: "Innovation in Education and Inclusion", 19-21 July 2018, Lima, Perú. (DOI): http://dx.doi.org/10.18687/LACCEI2018.1.1.66

[28] Vizcarro C. and Juárez E., La metodología del Aprendizaje Basado en Problemas, ¿Qué es y cómo funciona el aprendizaje basado en problemas?, Universidad de Murcia, España, 2008, cap. 1. http://www.ub.edu/dikasteia/LIBRO_MURCIA.pdf

[29] Farinazzo Valeria, Sampaio Paulo, Cordeiro Antonio and Ferreira Bruno, Implementing a Data Network Infrastructure Course using a Problem based Learning Methodology, Journal of Information Systems Engineering and Management, 2018 - Volume 3 Issue 2, Article No: 10 https://doi.org/10.20897/jisem.201810.

[30] N. Bedregal-Alpaca, D. Tupacyupanqui-Jaen, M. Rodriguez-Quiroz, L. Delgado-Barra, K. Guevara-Puente and O. Sharhorodoska, "Problem-Based Learning with ICT Support: An experience in teaching-learning the concept of derivative," 2019 38th International Conference of the Chilean Computer Science Society (SCCC), Concepcion, Chile, 2019, pp. 1-7. DOI: 10.1109/SCCC49216.2019.8966396

[31] Dirección de Investigación y Desarrollo Educativo, Vicerrectoría Académica, Instituto Tecnológico y de Estudios Superiores de Monterrey. "El Aprendizaje Basado en Problemas como técnica didáctica", pp.14-18, [On line]. Disponible: http://www.sistema.itesm.mx/va/dide/inf-doc/estrategias/

[32] F. Bermejo, and M.J. Pedraja, La metodología del Aprendizaje Basado en Problemas, La evaluación de competencias en el ABP y el papel del portafolio, Universidad de Murcia, España, 2008, cap.5.

[33] S. Grau C. and Gómez L. Ma.C., "La evaluación, un proceso de cambio para el aprendizaje". En: Evaluación de los aprendizajes en el Espacio Europeo de Educación Superior / Salvador Grau Company, Cecilia Gómez Lucas (Coord.). Alcoy: Marfil, 2010. ISBN 978-84-268-1523-1, 17-32. http://rua.ua.es/dspace/handle/10045/14937.

[34] Octaedro, Rúbricas para la evaluación de competencias, España, 2013, pp. 8-23.

[35] Sáiz Manzanares, María Consuelo and Bol Arreba, Alfredo; Aprendizaje basado en la evaluación mediante rúbricas en educación superior, Elsevier, Suma Psicológica, SUMA PSICOL. 2014; 21(1):28-35, España, 2014.

[36] Ortega Andrade NA, Romero Ramírez MA and Guzmán Saldaña RME, Rubrica para evaluar la elaboración de un Proyecto de Investigación Basado en el Desarrollo de Competencias, Universidad Autónoma del estado de Hidalgo, México, 2014, vol.2, No. 4. https://www.uaeh.edu.mx/scige/boletin/icsa/n4/e6.html

[37] Ma.C. Sanchez Martinez, M. Aguilar Benegas, J.L. Martinez Durán, and J.L. Sánchez Ríos, Estrategias didácticas en entornos de aprendizaje enriquecidos con tecnología (antes del covid-19), Universidad Autonoma Metropolitana-Xochimilco, Mexico, 2020, pp.67-71.

[38] C. Baluarte-Araya, " Proposal of an Assessment System based on Indicators to Problem Based Learning – IEEE Conference Publication, Published in: 2020 39th International Conference of the Chilean Computer Science Society (SCCC), 16-20 Nov. 2020, Coquimbo, Chile. DOI: 10.1109/SCCC51225.2020.9281203.

[39] J.M. Muñoz Cantero, and M.P. Rios de Deus, Indicadores de evaluación de la investigación en educación superior, REVISTA GALEGO-PORTUGUESA DE PSICOLOXÍA E EDUCACIÓN, N° 8 (Vol. 10) Ano 7°-2003 ISSN: 1138-1663.

[40] C. Baluarte-Araya, "Project based Learning Application Experience in Engineering Courses: Database Case in the Professional Career of Systems Engineering", (IJACSA) International Journal of Advanced Computer Science and Applications, Volume 11 Issue 3, 2020, pp. 131-140. DOI: 10.14569/IJACSA.2020.0110316.

[41] Xiandong Li, Fei Xie, Xiaoqiang Li, Guangwu Li, Xu Chen, Jun Lv, Chunyan Peng, Development, application, and evaluation of a problem-based learning method in clinical laboratory education, Clinica Chimica Acta, Volume 510, 2020, Pages 681-684, ISSN 0009-8981. https://doi.org/10.1016/j.cca.2020.08.037. https://www.sciencedirect.com/science/article/pii/S0009898120304307.

[42] Jackeline Valencia, Jackeline Macias, Alejandro Valencia, Formative Research in Higher Education: Some Reflections, Procedia - Social and Behavioral Sciences, Volume 176, 2015, Pages 940-945, SSN 1877-0428. https://doi.org/10.1016/j.sbspro.2015.01.562. https://www.sciencedirect.com/science/article/pii/S1877042815005996.

[43] J. Flores, J. Avila, C. Rojas, F. Sáez, R. Acosta and C. Diaz, Estrategias Didácticas para el aprendizaje significativo en contextos universitarios, Unidad de Investigación y Desarrollo Docente, Dirección de Docencia, Universidad de Concepción, Chile, 2017, pp. 10-14.

[44] B. Santamarina, La evaluación de los estudiantes en la Educación Superior: Mas allá de la Evaluación, Universidad de Valencia, Servicio de Formación Permanente, España, 2007, pp. 49-53.

[45] Kuo-Huan Ting, Chung-Ting Cheng, Hou-Yi Ting, Introducing the problem/project based learning as a learning strategy in University Social Responsibility Program - A study of local revitalization of Coastal Area, Yong-An District of Kaohsiung City, Marine Policy, Volume 131, 2021, 104546, ISSN 0308-597X. https://doi.org/10.1016/j.marpol.2021.104546. https://www.sciencedirect.com/science/article/pii/S0308597X21001573.

Genetic Behaviour of Zika Virus and Identification of Motif

Pushpa Susant Mahapatro[1], Jatinderkumar R. Saini[2]*
Ph.D. Scholar, Symbiosis International (Deemed University), Pune, India[1]
Professor & Director, Symbiosis Institute of Computer Studies and Research[2]
Symbiosis International (Deemed University), Pune, India

***Abstract*—ZIKV is a mosquito-borne disease. It is known to cause neurological disorders and congenital disabilities in newborns. The Genome Sequence of the Zika virus is used for the study. The essential cell functionalities like circadian behavior and expression of genes are studied. Regulatory proteins are alternating functionalities during daytime and night time. Identifying motif is made by understanding the features of motifs, finding the count matrix, and formulating the profile matrix. The consensus string of the Zika virus is to be computed, and the score motif is to be calculated. Different techniques of motif finding like the Brute Force technique and Greedy Search techniques are proposed. In the Brute Force technique, each motif is selected, its score is to be calculated, and then the minimum score can be obtained. The Brute Force technique will take an enormous amount of time, but it is guaranteed to find a solution. The Greedy Search technique is not guaranteed to find motif like the Brute Force technique but can give a close answer in a realistic time. This paper presents the identification of motif in the Zika virus genome using programming techniques.**

***Keywords*—Circadian behaviour; consensus string; genome study; greedy search technique; motif search; regulatory proteins**

I. Introduction

Zika virus (ZIKV) was discovered in 1947. It is a mosquito-borne disease. In 2013, it spread in South Pacific. It continued to spread to all parts of America. Early studies revealed that the virus originated and remained in Africa for many decades. Zika virus study is interesting as it can solve many biological problems. Genome sequences are quite complex. It is not possible to explain by a probabilistic model. So low-order Markov models explain the properties quite well. The DNA k-mer frequencies in the genome sequence of the Zika virus provide an insight into the genome complexity. It is possible to study the k-mers with different lengths as segments of genome sequence from different animals are available. A study was carried out for k-spectra of different species from Archaea, Bacteria, and e-coil. It studied the modalities of distributions. Some species have multimodal spectra, whereas all other species have a unimodal k-mer spectrum [1].

As a patient falls ill, based on the medical signs and symptoms, Symptoms bases categorization is possible. A frequent association can be extracted based on association mining [2].

There is a need to extract a sequence of existing strains. Different disciplines are collaborating for combating the outbreak of the disease. The different methodologies used are RNA extraction and material validation. Also, genome sequencing, consensus variation, and sequence analysis are done to understand the whole genome sequencing. The scientific method does not need to separately authenticate the virus reagents. The data is available through public repositories. It helps to study the pathogenesis, neurotropism, prevention, and possible spread [3]. Microcephaly, primers, and probes in ZIKV is detected. ZIKV is transmitted from a viremic host to normal people. Non-African mosquitoes have more potential to transmit the disease compared to African mosquitoes. Mosquito to mosquito transmission is possible and was evaluated in a study. In Africa, aegypti is less susceptible to ZIKV [4]. The spread of ZIKV is associated with neurological complications. The ZIKV spread and mode of transmission is studied carefully. Serological tests of ZIKV react with antibodies by other viral infection. Viral nucleic acids are present in polymerase chain reaction testing. Also, virus isolation is done for confirmation. It confirms the presence of ZIKV. It combines with global aedes vector distribution. Mother to the foetus and sexual transmission is very common person-to-person transmission modes [5].

The detection of the infection becomes difficult due to its close association with flaviviruses. The potential of antibody-dependent enhancement also increases as the cross reactiveness to flaviviruses like the dengue virus and the West Nile virus. Serum samples were collected and tested. A study was conducted on the dengue virus and the West Nile virus to find whether these viruses can enhance or neutralize the ZIKV. The West Nile virus enhanced ZIKV, so it failed to neutralize [6].

People with ZIKV have a mild fever and fewer symptoms to identify the infection. Babies are born with birth defects for infected pregnant women. Deterministic models were designed, considering sexual transmission, mosquito-human transmission, and Wolbachia-infected male mosquito release. Disease-free equilibrium and its stability is studied. The study was performed on impact of parameters on reproduction. The intervals of the liberation of Wolbachia-infected male mosquitoes were studied. A bounded global solution was derived for the extinction of ZIKV. As per the numerical simulations, ZIKV may be destroyed when the amount of white noise reaches a threshold value. The wild mosquitoes may be extinct with the delivery of Wolbachia-infected mosquitoes [7].

ZIKV is a severe public health issue. Still, a little study is done to find the transmission of the Zika virus in sexual groups. A study to control the spread of virus between

*Corresponding Author

individuals by changing contact patterns is done. A heterosexual network-based model is designed based on the Costa Rica case study. A study is carried out to measure the effect of changing the degree of heterogeneity. It is measured by removing the sexual contact of persons with a limited number but a greater degree and at different places. A threshold time for Zika virus infection next to the peak time was devised [8].

II. Literature Review

The ZIKV genome is studied to better understand the evolution and spread of Zika virus infection in more than fifty countries of the world. The spread is caused due to infected mosquito bites and person-to-person transmission. This disease is better understood with molecular insights of ZIKV [9]. It can help to better combat the disease. In monkeys and humans, the neural progenitor cell growth is attenuated by infection. The DNA is damaged by the virus and activates DNA damage responses [10]. The biological cycle of the virus is studied to understand the behaviour of the virus during the daytime, night time, and replication process. With the introduction of ZIKV to the Americans, four mutations of ZIKV were reported. This represents direct eversions from earlier mutations during the spread from Africa to Asia. Studies were performed with and without mutation on the experimental infection of aedes aegypti mosquito and human cells. It was found that fitness is reduced for original mutation for urban human-amplified transmission, whereas the fitness was enhanced for new mutations increasing the risk. The findings include three adaptive mutations of ZIKV [11].

The adult with moderate immunocompetent features may get infected by ZIKV. It triggers and enhances antiviral responses and brain damage. The neuroendocrine functions, inflammation, and immune reaction for different pathogens can get modulated due to gut microbiota composition. The modification stimulated by ZIKV in the belly microbiome of immune-capable mice was studied. It was found that the infection caused a considerable decline in microbes like Actinobacteria and Firmicutes phyla; compared to healthy mice. A significant boost of Deferribacteres and Spirochaetes was identified. Intestinal harm and extreme white blood cells recruitment were caused due to modulation of microbiota induced by the Zika virus [12]. The birth defects are associated with utero exposure to ZIKV. In early childhood, the impact remained unclear. The study of neurodevelopment and impact of ZIKV to 24-month toddlers born to pregnant women infected with ZIKV was conducted. These women were pregnant during the 2016 ZIKV outbreak in America. There was no abnormal transfontanelle cerebral ultrasound finding before and after delivery. But later, the child had reduced brain activity and birth defects [13]. There are no approved vaccines, i.e., antiviral treatments, available. Using the dengue vaccine as a reference, a chimeric dengue/ZIKV named VacDZ was created. It is a live diluted inoculation to ZIKV. It reveals key markers of dilution of pathogenicity in interferon deficient adult mice. The vaccine shows an immune response to ZIKV. It neutralizes the virus and shows a successful shot against ZIKV in mice [14]. The mobilization of the health test center network to detect COVID-19 patients was prompted by the emergence of SARS-Co V-2. It started tracing the contacts; identify the hot spot area prone to active community transmission. The Brazilian public health system faced difficulties amid triple epidemics, i.e., dengue, chikungunya, and Zika virus. Various samples were collected from Brazil and tested. An inter-disciplinary response to health gained importance. A need to search for an effective vaccine became important as no vaccine is 100% effective to any virus [15]. The Literature Review is shown in Table I.

TABLE I. Literature Review

Sr. No.	Year	Ref.	Topic	Concept/ Theoretical Model	Paradigm/ Method	Context/ Setting/ Sample	Findings	Future Research
1	2016	[5]	Origin and spread of ZIKV	Paper illustrates the neurological complications and spread of Zika virus.	The Zika virus spread and its mode of transmission.	The virus is separated, the viral nucleic acid detection is done using polymerase chain reaction testing	Viruses mostly spread due to international travel. Person-to-person transmission both vertically and horizontally is possible.	Reasons for the spread of the virus
3	2018	[3]	Genome sequencing and variant analysis, phylogenetics, and profound sequencing of Zika virus strains	The existing strains of the Zika virus are validated.	RNA Extraction and Material Validation, Genome Sequencing, 3′ Race, Read Assembly, Code Availability, and many more are done.	Sequence Comparison of Various Stocks, Deep Sequencing for Minor Variants, Recombination Analysis,	Minor variants were detected. Separate authentication is not needed. It represents the viral population diversity of ZIKV.	To understand better immunology, dissemination, possible cure, and avoidance
4	2019	[1]	Models of Genomic DNA k-mer spectra	The genomic complexity is studied in whole genome sequences using frequencies of DNA k-mers	k-mer sequences and modalities of around 100 species of Archea, Bacteria, and viruses are studied. A few species have multimodal spectra, whereas others have unimodal	To study complicated Genomic sequences, a probabilistic model is not sufficient. A low-order Markov model is discovered for study.	Dinucleotide suppression like C+G and CpG happens in Multimodal spectra. Tetrapods were analyzed. The sensor system was identified in the Human genome	Another genome like Entamoeba histolytica is to be studied.

5	2019	[2]	Apriori-based Frequent Symptoms Association Mining in Medical Databases	Symptom-based categorization, the symptoms are described by patients to a medical consultant	A large volume of the data repository is used, and using association rule mining; frequent associations are extracted.	Using the symptoms, the database is scanned to count the minimum support of each candidate.	Patients are found to be infected by malaria as most have the same set of symptoms.	The classification algorithm suggested mining important information health care databases.
6	2019	[16]	Flavivirus replication stimulated by circadian clock components	The invulnerable reactions of viruses are regulated by the circadian clock. It also affects pathogen replication. The underlying molecular mechanisms is difficult to understand.	Some components like BMAL1 and REV-ERBα are identified as circadian components.	Replication of flaviviruses, dengue, and ZIKV is inhibited by REV-ERB	Highlights the circadian clock compo nent in regulating flavivirus replication	Fragment entry into hepatocytes and RNA reproduction to be performed
7	2020	[4]	Susceptibility of Enhanced Zika virus	ZIKV is transmitted from an infected host. Non-African mosquitoes have more potential to transmit.	Transmission from mouse to mosquito is assessed in immune-compromised	Day Post Mouse infection	The population outside Africa are more susceptible to the Zika virus as compared to native populations of aegypti in Africa	More sophisticated experiments to be performed
8	2020	[17]	ZIKV peripheral blood Gene expression responses	Congenital ZIKV is transmitted from mother to fetus. Tangential blood cells are unfettered to Zika virus and carry the infection	The gene expression of the Profile of diseased and healthy Peripheral blood mononuclear cells is compared. Samples were collected from expectant and non-pregnant women.	More expression by M1 shifted pro-inflammatory responses, less expression by M2-shifted anti-inflammatory	Shaping neonatal pathology in Pregnancy-induced immune dysregulation	Records participating in osteoclast diversity and cardiac myopathies
9	2020	[18]	Locating Motif in DNA sequences	Discovering motif or short replicating forms in genetic strings	Genomic monitoring systems of living creatures	Exhibited the motif hunt challenge	The actual motif is obtained from a set of candidate motif	Locate motif with better precision
10	2021	[6]	Enhancement of ZIKV infection by antibodies from West Nile	Owing to quiet antigenic similarity, exposure to ZIKV is difficult. It raises the potential of antibody-dependent development.	The capacity to improve or deactivate ZIKV infection is examined from the serum trials acquired with suggestive or symptomless WNV disease.	Sero-investigation information showed a 7% occurrence for WNV antibodies	WNV antibodies in the sera substantially improve Zika virus in Fc receptor constructive cells	Additional appropriate versions of ADE
11	2021	[7]	Spread of ZIKV mitigated due to release of Wolbachia diseased mosquitos	Deterministic and stochastic versions are created	Disease-free balance and its strength are explored. The effect of the factors on basic breeding is studied.	The stochastic prototype has a distinctive and bordered overall solution	The randomness may force the disease to be eliminated when the strength of the white noise is significant enough.	Wild mosquitos turn out to be destroyed as Wolbachia-diseased mosquitoes reduce the time
12	2021	[8]	Costa Rica analysis on effect of contact patterns of sexual groups	Diseased mosquitos and sexual relations are primary reasons for Zika spread	Altering contact patterns among persons to manage Zika virus propagation	Build heterosexual network-based version to find threshold time of Zika infection	Analysis of the impact of altering the degree of heterogeneity in the sexually active places at distinct period	Upper limit later than the peak time of ZIKV infected cases
13	2021	[11]	Spread of ZIKV due to mutational reversions and health restoration	Zika caused four alterations prior to propagation in the Americas,	Diseases of Zika virus with and without alterations reveal that the original transformations decreased strength for human-intensified spread.	Classification of transmission adaptive Zika virus alterations	Early alterations and drift when the virus was announced	Pandemic rise due to reversions restored fitness

14	2021	[12]	Immunocompetent rats stomach microbiota modulation induced by ZIKV	Neuroendocrine functions are modulated by stomach microbiota composition. It includes contamination, cellular and immunologic reactions	Next-generation sequencing is used to analyze stomach microbiota structure	Compared to uninfected mice, the Zika virus activated a substantial illness in the microbes in the infected mouse.	A significant rise in the microbes from the Spriorchaetaceae heredity in the stomach microbiota is observed.	Colon soft tissue homeostasis in adult immunocompetent rats is diseased by ZIKV.
15	2021	[13]	Exposure of ZIKV and neurodevelopment in toddlers normocephalic at birth	In French regions, newborn to Women pregnant in 2016 is studied. Impact of utero ZIKV exposure on neurodevelopment among toddlers	Inhabitants-based mother-child associate study, pregnancies overlapped with the 2016 ZIKV epidemic	Due to ZIKV infection during pregnancy, exposure of the toddler to ZIKV was identified. Different stages of questionnaires can be created for the assessment of Brain development	15.4% of ZIKV-revealed toddlers and 25.3% ZIKV-unexposed toddlers were found. ASQ result of 2SD cut-off (P = 0.10) is obtained.	Population-based cohorts of in utero ZIKV-exposed normocephalic newborns to be analyzed.
16	2021	[14]	Single-dose live-attenuated chimeric vaccine	VacDZ vaccine is proposed against ZIKV. PDK-53 dengue virus vaccine to be used as a backbone	It is studied as the key makers of mitigation. Plaque phenotype, heat sensitivity, mitigation of neurovirulence in suckling mice is studied.	VacDZ to be controlled as a conventional live virus inoculation, or as a DNA-introduced inoculation that emits living VacDZ.	Vaccine expressions induce an impervious defensive reaction to ZIKV in AG129 mice. It neutralizes antibodies and a strong Th1 reaction.	VacDZ to be an effective vaccine
17	2021	[15]	Spread of DENV, CHIKV, ZIKV, and SARS-CoV-2 in Brazil	Viruses like SARS-CoV-2 have triggered the deployment of the network. Detect COVID-19 affected role, locate interactions and classify regions	The incursion of SARS-CoV-2 during ongoing triple arboviral pandemics produced by dengue, ZIKV, and Chikungunya	Sample collected from Brazil.	Fitness methodology is an efficient way to handle to lessen the destructive impact triggered by pathogens.	Successful antivirals serum treatment for diseased patients and vaccination for non-infected

In the biological sequences, motifs are short repeating patterns. Motifs are least conserved, so; it is a challenging task to identify the motifs. Motifs are important to study the genetic behavior of the Zika Virus. The candidate motif of the dataset can be computed to identify the real motif using the computation method.

III. Materials and Methods

A. Data Collection: Zika Virus Genome

ZIKV is family of Flaviviridae. It is a type of virus family. The entire genome sequence is available at NCBI. The functionalities of DNA were studied and discussed by Watson and Crick [19]. The filename of the dataset is ZikaVirus.fasta. It is stored as a nucleotide sequence, and fasta defines the file format. The size of the dataset is 11 KB. The genome of the Zika virus is stored in the file.

B. Circadian behavior in Zika Virus

The daily activities of any virus or any living organism are controlled by an internal clock called the circadian clock. Animals also follow the daily routine work based on the circadian clock. The clock maintains a 24-hour activity cycle. When it starts malfunctioning due to disorder, then many organisms face genetic diseases. It is called a delayed sleep-phase syndrome. The circadian clock has its base at the molecular level. Because of the malfunction of the circadian clock, people become prone the many diseases. Heart attack is more common in the daytime, whereas asthma attack is more common in the night time.

Scientists Ron Konopka and Seymour Benzer identified abnormal circadian patterns in mutant flies and traced their causes. They found that the mutation in a single gene. Later after many years, a similar clock gene in mammals was discovered. Then many circadian genes were discovered. These genes display a high degree of evolutionary conservations across different species. Maintaining the circadian clock in a plant is very important as its entire life cycle depends on it. It is a matter of life and death for plants. More than a thousand plant genes are circadian. Such genes include the genes that control photosynthesis, photoreception, budding and flowering. Circadian behavior of the Zika virus is studied [16]. The immune system is regulated by the circadian clock. The immune system reacts to microbes, and pathogen replication is affected. BMAL1 and REV-ERBα are circadian components related to flaviviruses in dengue and Zika. The replication of flavivirus is regulated by the circadian clock.

C. Representations of Genes

DNA makes RNA which makes proteins. It is composed of four ribonucleotides, namely adenine, cytosine, guanine, and uracil. Thymine is replaced by Uracil in DNA. RNA transcript is translated to the amino acids sequence of a protein. These proteins regulate the function inside the cell.

DNA replication happens at the origin of replication called ori. Finding the position of ori is a complicated task even for biologists. The process of transcription and transpiration is also a complicated task happening inside the cells [20]. During transcription, all occurrences of Thymine (T) in DNA is replaced with Uracil (U). The RNA strand is then translated into an amino acid sequence. The RNA strand is partitioned into 3-mers. These 3-mers are called codons. Each codon takes the form of one of the 20 amino acids. During this, it follows the genetic code. Each of the 64 codons encodes an amino acid. Out of 64 codons, 3 codons are stop codons which halt the translation. For example, the DNA string "ATATCGAAA" transcribes into the RNA string "AUAUCGAAA" which translates into the amino acid "ISK".

Cells can transcribe different genes and can form RNA. The rates may be different for other genes. This is known as gene transcripts or gene expression. That is the reason why brain cells and skin cells behave in various manners. Both have different functionalities and vary greatly in their features. These variations help the cells to understand the time and keep track of it. Pregnancy-associated variations in reactions to ZIKV were identified using DNA expression of samples of different women. ZIKV infected pregnant showed pro-inflammatory responses [17].

IV. Regulatory Proteins

The dataset contains the nucleotide sequence of the Zika virus. The length of the string can be found using the python program. It was found to be 10780. Each cell in the plant keeps track of day and night. There are three master cells, which are called clock masters. These are CCA1, TOC1 and LHY. These genes are controlled by external factors like sunlight and the availability of nutrients in the soil. This helps the organism to adjust to the gene expression.

The regulatory protein TOC1 regulates the expression of LHY and CCA1. The expression of TOC1 is suppressed by LHY and CCA1. It basically works in a negative feedback loop. Sunlight activates the transcription of LHY and CCA1. This deactivates the TOC1 transcription. At night time, TOC1 peaks and starts promoting the transcription of LHY and CCA1. LHY and CCA1 repress the transcription of TOC1, and the loop continues. The Condon usage is controlled by biased nucleotide composition in the Zika virus [21].

The transcription regulates a gene by binding to a specific short DNA. It is called a regulatory motif [22, 23]. It is also called as the transcription binding site. It is the upstream gene region which is 600-1000 nucleotide long, also the start of the gene. CCA1 can bind to "AAAAAATCT" in the upstream region. It will be helpful for bioinformaticians if the regulatory motifs can be in the gene. An algorithm to find motif will be useful.

A. Importance of Motif

Motifs are short sequence patterns. It has a finite length. It is used to study the features of DNA, RNA, and Proteins. Transcription factor binding sites are represented using sequence motifs. Finding the motif sequences of motifs can help in understanding the transcription regulation [24, 25]. Motifs represent active sites of enzymes and proteins structures and stability. Study of DNA Arrays is done to identify the genes that are active during the daytime in plants. The upstream region of nearly 500 genes was extracted to find the circadian behaviour. The frequently appearing pattern in the upstream region was identified. Suppose it was found that "AAAATATCT" is the most frequent word that appears more than 40 times. It was named as an upstream region evening element. The gene loses its circadian behaviour if the gene is muted. In plants, the evening element is quite conserved. It is easy to find the evening element in the plant whereas in animals, finding the evening element is quite difficult because of many mutations. If a fly is infected with a bacterium, its immunity genes will get activated to fight with the bacterium. The immunity gene has elevated expression levels as the fly gets infected. The most common 12-mers is "TCGGGGATTTCC" in the upstream region of many genes. It is the binding of the transcription factor NF-kB that activates the various genes in flies. The biological challenge of finding a regulatory motif is to be converted into a computational problem.

Depending on the similarity with the ideal motif, it will score individual instances of motifs. An ideal motif is the transcription factor binding site that best binds to the transcription factor. An attempt is made to select a k-mer from each string, as the ideal motif is not known. Each motif is scored depending on their similarity to each other. A list of t DNA string DNA is taken. Each string is of length n. k-mer from each string is selected to form a collection of motifs. It represents a (t X k) motif matrix. The motif matrix of the Zika virus is formulated. The most frequent Nucleotide in each column is identified and denoted by upper case. By using different values of k-mers in each string, a different motif matrix from each DNA string is created. The most conserved motif matrix is to be obtained. It also means matrix with most uppercase characters or few lower-case characters. The goal is to compute a collection of k-mers that minimizes the score.

B. Finding the Count Matrix

A 4 X k count matrix can be created for a given Motifs. It is denoted by count (Motifs). It represents the count of each Nucleotide in each column of the motif matrix. The element (I, j) represents the count of Nucleotide I in column j of Motifs.

{'A': [45, 37, 39, 41, 42, 36, 33, 46, 45, 50, 41, 41, 42, 46, 38, 45, 46, 36, 39, 49, 38, 42, 40, 45, 36, 39, 42, 41, 40, 30, 36, 40, 36, 44, 40, 46, 36, 45, 40, 48, 37, 46, 41, 41, 47, 40, 44, 51, 38, 47, 41, 46, 29, 49, 41, 51, 39, 40, 39, 46, 33, 50, 52, 36, 50, 40, 45, 58, 47, 50],

'C': [23, 39, 34, 29, 24, 34, 37, 33, 36, 30, 27, 34, 23, 35, 41, 38, 34, 43, 36, 32, 46, 31, 37, 34, 35, 42, 38, 28, 40, 33, 38, 35, 43, 37, 34, 33, 32, 41, 26, 35, 36, 36, 31, 36, 41, 34, 39, 24, 32, 33, 32, 29, 49, 42, 31, 24, 35, 33, 32, 33, 42, 26, 26, 30, 32, 36, 34, 26, 33, 31],

'G': [41, 53, 46, 45, 49, 49, 42, 35, 46, 45, 48, 47, 57, 38, 45, 50, 38, 50, 43, 42, 33, 49, 40, 42, 56, 43, 41, 46, 41, 45, 48, 48, 39, 37, 41, 45, 42, 38, 57, 42, 47, 45, 50, 47, 41, 46, 38,

43, 56, 49, 49, 52, 50, 35, 48, 41, 40, 46, 47, 42, 52, 42, 40, 50, 48, 48, 48, 35, 45, 38],

'T': [45, 25, 35, 39, 39, 35, 42, 40, 27, 29, 38, 32, 32, 35, 30, 21, 36, 25, 36, 31, 37, 32, 37, 33, 27, 30, 33, 39, 33, 46, 32, 31, 36, 36, 39, 30, 44, 30, 31, 29, 34, 27, 32, 30, 25, 34, 33, 36, 28, 25, 32, 27, 26, 28, 34, 38, 40, 35, 36, 33, 27, 36, 36, 38, 24, 30, 27, 35, 29, 35]}.

C. Formulating the Profile Matrices

An ideal motif is a transcription factor binding site that binds the best to the transcription factor. A motif finding problem is would score instances of motifs depending on the similarity to the ideal motif as the ideal motif is not known to us. Our aim is to find a k-mer from each string of the array and find the score depending on similarity.

The most frequent Nucleotide in each column is identified and denoted in the upper case. If two nucleotides are most frequent, then randomly, one Nucleotide is selected. The motif matrix is represented as a string of motif matrices. The i-th row and j-th column can be accessed by using the motif[i][j]. A conserved matrix is a matrix with a smaller number of lower-case characters or more uppercase characters. A most conserved motif matrix is to be selected from several different motif matrices. From a given sample of DNA string, using different values of k-mer, a different motif matrix can be created. The score of the motif matrix is found by counting the number of lower-case letters in the motif matrix. Then we can find a set of k-mer that reduces the score. To find it, all elements of the count matrix is divided by the number of rows in the motif i.e., t. The resultant matrix is the Profile of the motif matrix. The element (I, j) is the i-th nucleotide frequency in the j-th column of the motif matrix. The sum of any column is 1 in the profile matrix.

D. Finding a Consensus String for Zika Virus

A consensus string for the Zika virus is derived by identifying the most common Nucleotide present in the column of the motif matrix. If two nucleotides have the same frequency, anyone is selected at random. If the motif is selected correctly, the consensus matrix provides a candidate regulatory motif.

The most frequent Nucleotide in each column i.e. the Consensus (Motifs) of the Zika virus genome is:

AGGGGGGAGAGGGAGGAGGACGAAGGAGGTGGCAG
ATAGAGAGGAGAAGGGGGAGAGGGAGAAGAGGAAA
. So, the consensus string of the Zika virus is known.

E. Score Motif

The score motif of the Zika virus can be calculated using the consensus matrix. The number of symbols in the j-th column that does not match with the symbol at position j of the consensus matrix is added. The score of the Zika virus genome is 7444.

V. Finding the Binding Sites

The motif finding problem is to be solved using a collection of strings. A set of k-mers for each string to be identified minimizes the score of the resulting motif. The input to the problem is the DNA string and an integer k. The output is k-mer collection motif for each DNA. The output will minimize the score motif for any choice of k-mers. A general problem-solving technique like Brute Force algorithm can find a solution that will take a lot of time to execute for a large genome. The brute force algorithm will consider each possible k-mers Motifs and gives a solution as motifs with the least score.

A. Comparing the Working of Brute Force Motif Finding

The Brute Force motif finding technique identifies all possible solutions. These algorithms may be easy to design. It will be guaranteed to find a solution as it will verify each and every possible solution and identify the best solution or the motif with the lowest score. These algorithms will take an enormous amount of time as it has to check all possible solutions to discard a motif. The number of candidate motif will be too large to verify.

In the brute force algorithm, n-k+1 choice of k-mers is possible. There is a number of ways to form motif are $(n-k+1)^t$. The algorithm can calculate the score in k X t steps. The running time of the algorithm is of the order $((n-k+1)^t)$ X k X t. This value is too high to be calculated using even the fastest computer. If the value of k is already known, then it may be a little easy, but this is not possible. So, another method needs to be explored.

VI. Results and Discussion

A. Use of Profile Matrix

Iterative procedures are used in many algorithms that select different alternatives during the iterations. Some of these iterations are correct, whereas some are not. The most attractive alternative is selected by greedy search algorithms. In a chess game, the Greedy search algorithm at every move tries to capture valuable piece. Greedy may not find the best solution but can quickly predict the approximate solution in many cases. So, The Greedy search is to be applied to biological problems to approximate a solution. So, this algorithm is applied for motif finding. A collection of k-mers from a DNA string is motif. The columns of the profile matrix are viewed as four-sided dice. Each Nucleotide {A, C, G, T} is present on each side. The first column of the profile matrix has the data (0.2922077922077922, 0.14935064935064934, 0.2662337662337662, 0.2922077922077922).

The sum of all probabilities is 1 for any column. So, it means that the probability of generating A is 0.2922077922077922, C is 0.14935064935064934, G is 0.2662337662337662, and T is 0.2922077922077922. The profile matrix for the Zika virus is given in the previous section. The probability of any selected string can be calculated using the entry in the i-th column of the Nucleotide. Say, for example, the probability of the series "ACGG" is found to be 0.006458989565084851. A higher probability k-mer is achieved when it is more like the consensus string, "AGGGGGGAGAGGGAGGAGGACGAAGGAGGTGGCA
GATAGAGAGGAGAAGGGGGAGAGGGAGAAGAGGAA
A".

Prob("AGGGGGGAGAGGGAGGAGGACGAAGGAGGTG GCAGATAGAGAGGAGAAGGGGGAGAGGGAGAAGAG GAAA") = 1.757001027053479e-36

B. The Search for Binding Sites

Search for Binding Sites or Greedy Motifs is done. The best motif is set to the first k-mer from each string in Deoxyribonucleic acid (DNA). The DNA string is represented using the abbreviation DNA. These strings will be helpful for study. It ranges over all possible k-mers in DNA[0]. It finds a value for each motif [0]. The algorithm builds a profile matrix for the k-mer. Then motif [1] is set equal to Profile most probable k-mer in DNA[1]. Greedy motif search is iterated by updating Profile. To generalize, to find k-mers motifs in the i-strings of DNA, greedy motif search constructs a profile matrix and sets motif[i] equal to Profile most probable k-mer from DNA[i]. k-mer from each string in DNA is obtained as a collection of strings. Greedy motif search compares whether the motif score is greater than the best scoring collection of motifs. If it is greater than the best score motif is updated, otherwise ignored. The execution moves to the beginning of the loop, and the next symbol in the DNA[0] is selected. The results of the Greedy motif search can be for different k-mer strings and summarized in Table II. The 1-mer and 2-mer string have less significance, so the results are demonstrated for 3-mer till 15-mer string.

TABLE II. RESULTS OF GREEDY MOTIF SEARCH FOR DIFFERENT K-MER STRINGS

Sr. No.	k-mer	Score	Snapshot of the k-mer string
1	3-mer	10	['ACA', 'GGT', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', , 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA' ', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGT A', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'AGA', 'GGA', 'GGA', 'GG GA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GCA', 'A GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'AGA', 'GGA', ' 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA',
2	4-mer	85	['TGAC', 'GGTT', 'GGAT', 'GGAT', 'GGTT', 'AAGA', 'GGAA', 'GGAC' , 'GGAA', 'GGAC', 'GGAA', 'GGAA', 'GGAT', 'GGAA', 'GGAC', 'GGAC ', 'GGAT', 'GGAA', 'GGAT', 'GGAA', 'GGAC', 'GGAA', 'GGAA', 'GGA A', 'GGAA', 'GGAT', 'GGAA', 'GGAA', 'GGTT', 'GGAT', 'GGAA', 'GG AC', 'GGAC', 'GGAA', 'TGAA', 'TGAC', 'GGAA', 'GGAA', 'GGAA', 'G GAA', 'GGAA', 'GGAC', 'GGAA', 'GGAA', 'GGAA', 'GGAA', 'GGAC', ' GGAA', 'AGAA', 'GGAA', 'GGAA', 'GGAA', 'GGAT', 'GGAT', 'GGAA', 'AGAA', 'GGAC', 'GGAA', 'GGAA', 'GGAA', 'GGAA', 'GGAA', 'GGAA']
3	5-mer	166	['GAAAG', 'GGTTT', 'GGATT', 'GGATG', 'GAAAT', 'GGAAG', 'GGAAT' TG', 'GGAAG', 'GGATG', 'GGAAG', 'TGAAT', 'GGAGT', 'GGTAG', 'GG 'GGAAG', 'GGAAG', 'GGAGT', 'GGAAG', 'GGATG', 'GGAGG', 'GGAGG', G', 'GGTTG', 'GGAGA', 'TGAAG', 'TGAGG', 'GGAAG', 'GGAGG', 'GGA TGAAG', 'GGAAG', 'GAAAT', 'TGTAT', 'GGATG', 'GATGG', 'GGAAG', ', 'GGAAG', 'GGAGA', 'GGAAG', 'GGAAA', 'GGTGG', 'GGATG', 'GGAA GAGG', 'GGAAA', 'GGAGT', 'GGAGG', 'GGTGG', 'GGAAG', 'GGAAA', ' , 'GGATG', 'GGAGG', 'GGAAG', 'TGAAG', 'GGAAA', 'GGGAG', 'GGAAG AAA', 'GGAAA', 'GGAAG', 'GGATG', 'GGAAA', 'GGAAA']
4	6-mer	288	['TCAGAC', 'GGTTTT', 'CGGATT', 'TGGGTC', 'TGGGTC', 'AAGAAG', ' C', 'TCAGAC', 'TGGATG', 'GGGAAG', 'TGGGAT', 'TGGAAG', 'TGTATT' GAAT', 'TGGAAC', 'GGGGAG', 'AGGGAC', 'GGGAAG', 'TGGAAG', 'TGGG 'TGGATC', 'AGGAAG', 'TGGGAG', 'TGGGTG', 'TGGTAT', 'TGAAAG', 'T ', 'TGGGAG', 'AGGGAG', 'TGGAAG', 'AGGAAC', 'AGGAAC', 'TGGGAG', GAC', 'AGAGAG', 'AGGAAT', 'TGGGTG', 'ACAAAG', 'TGGAAC', 'CGGAA TCTATG', 'AGTGAC', 'TGGGAC', 'ACAATT', 'GAGAAC', 'TGGAAT', 'TG , 'GGAAAC', 'GGGGAG', 'TGTGTC', 'TGAAAT', 'TGAAAG', 'TGGAAG', AG', 'ACAAAG', 'GGGGAG', 'AGGAAG', 'TGGGAT', 'TGAATG', 'TGGGAC GGAAG', 'AGGATG', 'TGGATC', 'GGAAAG']

5	7-mer	412	['AGTTTGAA', 'GGTTTTAT', 'CGGATTGT', 'AGGATGGT', 'AATCAAGC', 'I , 'GGGTTGAT', 'GAGCTGGT', 'GGTAAGAT', 'AAGGTGAA', 'GATGTGGA', ', 'CAGATGGA', 'AGGTTGAT', 'GAGAAGAA', 'CATTTGAA', 'CAGTTGGA', T', 'AGGTTAGA', 'AGTGTGAT', 'AGGCTGAA', 'CAGATGGA', 'TAGCTGGG', TT', 'TGGTTGTT', 'GCTCTGAC', 'AGTGTGAA', 'AGGCTGGT', 'GAGCTGGC GGT', 'AATATTTA', 'CAGCAGGA', 'TGGGAGTT', 'AGGGAGGA', 'GAGCTGG(TGGG', 'TGGCTGGA', 'AGGCAGGA', 'TGTATGGA', 'AAGATGGC', 'AGGAAGI CAAAG', 'GGGCTTGA', 'GGGAAGGA', 'CAGCTTGG', 'AAGCTGGA', 'GGTAT(TGTGGC', 'GAGATGGT', 'GGGCTGGA', 'TGGGTGGA', 'GGGGTGGA', 'TGGC GTCAGGT', 'AGTCTGTA', 'TATTTGAA', 'TGTGTGTA', 'AATTTGGA', 'GGG(
6	8-mer	508	['AGTTTGAA', 'GGTTTTAT', 'CGGATTGT', 'AGGATGGT', 'AATCAAGC', 'I , 'GGGTTGAT', 'GAGCTGGT', 'GGTAAGAT', 'AAGGTGAA', 'GATGTGGA', ', 'CAGATGGA', 'AGGTTGAT', 'GAGAAGAA', 'CATTTGAA', 'CAGTTGGA', T', 'AGGTTAGA', 'AGTGTGAT', 'AGGCTGAA', 'CAGATGGA', 'TAGCTGGG', TT', 'TGGTTGTT', 'GCTCTGAC', 'AGTGTGAA', 'AGGCTGGT', 'GAGCTGGC GGT', 'AATATTTA', 'CAGCAGGA', 'TGGGAGTT', 'AGGGAGGA', 'GAGCTGG(TGGG', 'TGGCTGGA', 'AGGCAGGA', 'TGTATGGA', 'AAGATGGC', 'AGGAAGI CAAAG', 'GGGCTTGA', 'GGGAAGGA', 'CAGCTTGG', 'AAGCTGGA', 'GGTAT(TGTGGC', 'GAGATGGT', 'GGGCTGGA', 'TGGGTGGA', 'GGGGTGGA', 'TGGC'
7	9-mer	603	['AGTTTGAAG', 'GGTTTTATT', 'CGGATTGTC', 'AGGATGGTC', 'AATCAAGC ', 'GGTCATGAT', 'GGGTTGATG', 'GAGCTGGTT', 'GGTAAGATC', 'AAGGTG AT', 'TGGATAAAC', 'CGGCTGAAA', 'CAGATGGAC', 'AGGTTGATA', 'GGGG GGAG', 'CATGTGGAG', 'CGGCTGGAA', 'TGGATGGTG', 'AGGTTAGAG', 'AG CTGAAG', 'TGGGAGGAT', 'GAGATGTAG', 'GAGCTAATT', 'TGTCTTTTG', ' AGGTGGTC', 'CAGCTGGAG', 'CAGATGGAG', 'AAGTTGGAG', 'GGGGAGATG', 'AATCTGTAT', 'TGGATGAGG', 'ACGCATTTC', 'GCTTTGATT', 'AGTTTGGTT , 'GGTGTGGAC', 'TGGATGGAC', 'CGGCTTTTG', 'GGGCAGAGA', 'GGGGTTG C', 'CATTTTGCT', 'CAGCTGGCA', 'TGGATGGAA', 'TAGCAGTAG', 'AGGCT
8	10-mer	697	['GCGAAAGCTA', 'GGTTTTATTT', 'CGGATTGTCA', 'CCGGACTTCT', 'AATC , 'GGTTTCGCTT', 'GGTCATGATA', 'AGGTATGTCA', 'GAGCTGGTTA', 'GAC ', 'AAGGCAAACT', 'GAGCAGTTCA', 'CCTCTGGCCA', 'GGGACAGTCA', 'GC T', 'GGGACAGGTA', 'AAGCAAGCCT', 'ACGCAATCCT', 'GGGATCTGTA', 'C CC', 'GAGATCACCC', 'GATAAGGCCC', 'AGTAAGGTCA', 'AGGAAGGGCT', ' ACC', 'GAGCTGGCCC', 'GGGTTCGCCA', 'AGTGTGGACA', 'GCGGAAGTCA', TTCA', 'AGTTATGTTA', 'AGAGACTCCT', 'GGGAAAACCA', 'GAGGAAGCCC', CAACT', 'AAGCCGGTCA', 'ACACATGCCA', 'ACGAAGACCA', 'AAGATGGCCT'
9	11-mer	806	['GAGTTTGAAGC', 'GGTTTTATTTT', 'CGGATTGTCAA', 'CCGGACTTCTG', 'I GCGGTCGCAAA', 'GGTTTCGCTTT', 'GGTCATGATAC', 'AGGTATGTCAG', 'GA(ACAGGCCTT', 'CGGAACTCCAC', 'GCACATGCCAA', 'GAGCAGTTCAC', 'GGGC(TGTCCTG', 'GGTCTGAACAC', 'GGAGTGTTGAT', 'GGGGTGTTCGT', 'GGGAAGI GACAC', 'GATCTGATCAT', 'GAGAGGGCTAC', 'GAGCTTGAAAT', 'AGGGTGAT(CGC', 'GCGATGGTTGT', 'GGGCACACTGC', 'GGGGGGTTTAT', 'AGGCTGGTCGI C', 'GAGGGGGTCTT', 'GGGGAGATGTC', 'GATCTGGTGTC', 'GCGAGGAACAT' , 'CAAGAGGATAC', 'CGTGACGCATT', 'GAAGTGGAAGT', 'AATGAGATCGC', 'GGAATAACCTA', 'GCAGAGGTGTG', 'GAGGTGGATGG', 'GGAGTGATGGA', 'CI

10	12-mer	922	['AGACTGCGACAG', 'GGTTTTATTTTG', 'CGGATTGTCAAT', 'AGGATGGTCTTG ACGGAAC', 'AGAATACACAAA', 'TGGATATTCAGG', 'GGGAAGCTCAAC', 'AGG' ', 'AGGTTGAGATAA', 'GGACAGGCCTTG', 'ACGCTGGGGCAG', 'CCAAAAGGCA AAAAGCATT', 'TCATTGGGCAAG', 'AGGAATGTCCTG', 'ACACAAAGAATG', 'G TG', 'AGGTTAGAGAAG', 'AGTGTGATCCAG', 'AGGCTGAAGAGG', 'TGAATGGC CTATGATCCTG', 'GGAAATGAACAC', 'AGTCAGACCAGC', 'AGCATGCTGCTG', CAGG', 'AAGATGCGGAAG', 'AGAGAGATCATA', 'AGACTGGAAAAA', 'GGGATG' 'AGGAAGAGACTC', 'GGGAAAACCAGG', 'TCGCTGCTGAAA', 'ACTCTGGAACAG' GGACTT', 'GGAGATGCCTAA', 'GCGCATAGGCAG', 'AGACTGACGAAG', 'CGAC
11	13-mer	1008	['TTTGAAGCGAAAG', 'GGTTTTATTTTGG', 'CGGATTGTCAATA', 'CCGGACTTC A', 'ACGGAGATCTAGA', 'TCGCAAACCTGGT', 'TGGATATTCAGGA', 'TACTTG CTCC', 'GATGAGAATAGAG', 'TTGGAAGCCTAGG', 'TTGTATTACTTGA', 'TCC TCTAAGA', 'GGGGAGAAGAAGA', 'TGGAAAAGCATTT', 'TGGATCAGTTGGA', ' GAGTTCAACT', 'TGGAAGGCTTGGG', 'AAGAATGCCCACT', 'TGGAGGATCATGG', GGGCTAAAGATGG', 'GGTGACTGCAGGA', 'TTGGAGTGCTTGT', 'TGGCAGTGCTG , 'TGGTGGGACTGCT', 'CTGTTGGCCTGAT', 'GGGTTCGCCAAGG', 'AGTGTGGA GA', 'CTGTGGTCCATGG', 'AGGAACATCCAGA', 'AGGAACTTCAGGA', 'GGGGT
12	14-mer	1113	['TTGAAGCGAAAGCT', 'GGTTTTATTTTGGA', 'CGGATTGTCAATAT', 'CCGGACT CTATGCTGGAT', 'ACGGAGATCTAGAA', 'TCGCAAACCTGGTT', 'TGGATATTCAG TTTGGC', 'GGGAAGAGCATCCA', 'ATGATTGTTAATGA', 'TGGAAGCCTAGGAC', T', 'AGGGACAGATGGAC', 'GAGAACTCTAAGAT', 'GGGGACTCTTACAT', 'TGG AAGCAAGCCTGGGA', 'ACGCAATCCTGGAA', 'GGGATCTGTAAAAA', 'TGGAAGGCT CCCAAATGA', 'GTGAAGAGCTTGAA', 'ACCCACTGCAAGCG', 'ATGGAGATAAGGC GCAA', 'GTTCTCATCAATGG', 'TGGCAATCCTGGCT', 'TCTCTCTGAAGGGA', ' , 'TGGTACGTATACGT', 'GACCACAGATGGAG', 'TGGGAGTTATGCAA', 'GCGCTC GAAAACCAGGAG', 'TGTCGCTGCTGAAA', 'TGGAACAGAAATCG', 'ACTATAATCTC GATGGGC', 'GGTCATACTTGATG', 'TGTCACACATGCCA', 'ACGAAGACCATGCA',
13	15-mer	1190	['CTGTTGCTGACTCAG', 'GGTTTTATTTTGGAT', 'CGGATTGTCAATATG', 'CCGG TG', 'CAGATGACGTCGATT', 'ACGGAGATCTAGAAG', 'TCGCAAACCTGGTTG', ' TACCTT', 'AAGGGAGCCTGGTGA', 'CAGAGAATCTGGAGT', 'CAGGACATGAAACTG CGCCTGAAAA', 'CTGAAACACTGCACG', 'GCGGTGGACATGCAA', 'CGTAATCACTG TGGACTTCTCAAAG', 'GTTGAAGCCTGGAGG', 'AAGCAAGCCTGGGAA', 'ACGCAAT , 'AAGAATGACACATGG', 'ATGGAATAGAAGAGA', 'CACAATACCAGAGAG', 'GTG TAG', 'AATGAACACTGGAGG', 'TTCAAAGTCAGACCA', 'TGCGATCTCCGCCTT', CTGCTAA', 'CGTGGTCTCAGGAAA', 'GCGGAAGTCACTGGA', 'AGGTGGTCCTGATG GAGAGAGCGAG', 'CGGTTGCGCTGGATT', 'GTGGGAGAGTGATAG', 'CTGAAGAAGA

The results are obtained for various k-mers. The 15-mer has a score of 1190. The score obtained using the Greedy method may not be the optimal score, as there may exist a motif with a minimum score which can be obtained by finding all possible solutions. But this method provides 15-mer motif in a reasonable time. This score can be improved using other algorithms.

VII. Conclusion and Future Work

The study is conducted to understand the behaviour of the Zika virus. ZIKV shows circadian expression, which also regulates the day-to-day functions in genes. Zika virus infection causes birth defects like neurological disorders in babies, and no proper cure or vaccine is available. This paper attempts to find the probability of every k-mer for a given profile matrix. The Profile most probable k-mer is calculated. This k-mer is most likely to be generated by Profile compared to all k-mers in the text. For example, if the size of k-mer string selected is 12, then the highest probability k-mer is found using the probabilities values is "AGGGGGGAGAGG". Similarly, if the size of k-mer string selected is 15, then the highest probability k-mer is found using the probabilities values is "AGGGGGGAGAGGGAG". Greedy motif Search is done as it is better compared to Brute Force search in real-time. The best motif is set to the first k-mer from each string in DNA. k-mer from each string in DNA is obtained as a collection of strings. Greedy motif search compares whether the motifs score is greater than the best scoring collection of motifs. If it is greater than the best score motif is updated, otherwise ignored. Python programming is used to study the genetic behaviour of the Zika virus genome.

If any value in the profile matrix is zero, then the entire probability of the string becomes zero. If a string is obtained for which the profile matrix value is zero, the string is completely rejected. Such results can be improved using other methods like the Laplace Rule of succession. The score obtained can be further improved using this Laplace Rule of succession.

REFERENCES

[1] Benny, H. David, G. Nick, L. Yaron and T. Massingham, "Genomic DNA k-mer spectra: models and modalities," Genome Biology, 2019.

[2] K. R. P. Ram, R. Jayakumar and A. Sankaridevi, "Apriori-based Frequent Symptomset Association Mining in Medical Databases," International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-7, Issue-5C, 2019.

[3] S. Susmita and P. Vinita, "Whole genome sequencing, variant analysis, phylogenetics, and deep sequencing of Zika virus strains," www.nature.com/scientificreports, 2018.

[4] A. Fabien, D. Stéphanie, M. Caroline, F. Igor, R. Noah, M. Elliott, M. Daria and B. Artem, "Enhanced Zika virus susceptibility of globally invasive Aedes aegypti populations," sciencemag.org, VOL 370 ISSUE 6519, 2020.

[5] K. K. Mary, A. Tomas, F. Veronika, S. S. Ravi and D. Christopher, "Zika: the origin and spread of a mosquito-borne virus," Bull World Health Organ 2016;94:675–686C, 2016.

[6] H. Garg, R. Yeh, W. Douglas M, T. Mehmetoglu-Gurbuz, R. Resendes and B. Parsons, "Enhancement of Zika virus infection by antibodies from West Nile virus seropositive individuals with no history of clinical infection," Garg et al. BMC Immunology (2021) 22:5 https://doi.org/10.1186/s12865-020-00389-2, 2021.

[7] X. Ling, C. Xinru and W. Hui, "Releasing Wolbachia-infected mosquitos to mitigate the transmission of Zika virus," Journal of Mathematical Analysis and Applications,Volume 496, Issue 1,2021, 124804, ISSN 0022-247X, https://doi.org/10.1016/j.jmaa.2020.124804., 2021.

[8] L. Xiao-Feng, J. Zhen, H. Daihai and L. Li, "The impact of contact patterns of sexual networks on Zika virus spread: A case study in Costa Rica,," Applied Mathematics and Computation, Volume 393, 2021, 125765, ISSN 0096-3003, https://doi.org/10.1016/j.amc.2020.125765., 2021.

[9] A. Wang, S. Thurmond, L. Islas, K. Hui and R. Hai, "Zika virus genome biology and molecular pathogenesis. Emerging microbes & infections," https://doi.org/10.1038/emi.2016.141, 2017.

[10] C. Hammack, S. C. Ogden, J. Madden, A. X. C. Medina, P. E, Y. Son, A. Cone, S. Giovinazzi, R. Didier, D. Gilbert, H. Song, M. G, Z. Wen, M. Brinton, A. Gunjan, T. H and M. Heise, "Zika Virus Infection Induces DNA Damage Response in Human Neural Progenitors That Enhances Viral Replication," Journal of Virology, 10.1128/JVI.00638-19, 2019.

[11] L. Jianying, L. Yang, S. Chao, T. D. N. Bruno, R. Yun, L. H. Sherry, H. R. Grace, R. A. Sasha, R. A. Clark, P. Kenneth, V. Nikos, S. Pei-Yong and C. W. Scott, "Role of mutational reversions and fitness restoration in Zika virus spread to the Americas," https://doi.org/10.1038/s41467-020-20747-3, 2021.

[12] C. Rafael, O. S. Igor de, A. B. Heloísa, P. d. S. Lívia, d. N. Raquel, P. Gabriel, S. P. Paulo, P. K. Gary, F. M. Corinne and M. KellyGrace, "Gut microbiota modulation induced by Zika virus infection in immunocompetent mice," https://www.nature.com/srep/, 2021.

[13] G. Rebecca, F. Olivier, T. Benoît, D. Mama, E. Narcisse, M. Nicolas, M. Adeline, H. Jean-Christophe, L. Noémie, C. Elvire, H. Bruno and F. Arnaud, "In utero Zika virus exposure and neurodevelopment at 24 months in toddlers normocephalic at birth: a cohort study," BMC Medical, 2021.

[14] C. Wei-Xin, C. H. L. Regina, K. Parveen, S. L. Tian, Y. Thinesshwary, Y. K. Hao, T. Zi-Yun, K. S. Cyrill, Z. Rong-Rong, L. Xiao-Feng, A. Sylvie, Q. Cheng-Feng and J. H. C. Justin, "A single-dose live attenuated chimeric vaccine candidate against Zika virus," Sealy Institute for Vaccine Sciences, 2021.

[15] J. R. d. S. Severino, J. F. d. M. Jurandy and P. Lindomar, "Simultaneous Circulation of DENV, CHIKV, ZIKV and SARS-CoV-2 in Brazil: an Inconvenient Truth," Science Direct, One health, 2021.

[16] X. Zhuang, A. Magri, M. Hill, A. Lai, A. Kumar, S. B. Rambhatla, C. L. Donald, Lopez-Clavijo, F. R. A, P. K. S, W. H. Chang, P. A. C. Wing, R. Brown, X. Qin, P. Simmonds, T. F. Baumert, D. Ray, A. Loudon, P. Balfe and M. Wakelam, "The circadian clock components BMAL1 and EVERBα regulate flavivirus replication," Nature Communications Volume 10, Issue 1, 1 December 2019, Article number 377, Dec 2019.

[17] S. J. Lim, A. Seyfang, S. Dutra, B. Kane and M. Groer, "Gene expression responses to Zika virus infection in peripheral blood mononuclear cells from pregnant and non-pregnant women," MicrobiologyOpen, Volume 9, Issue 12, 2020.

[18] B. A. Faisal and S. R. S. Md, "MFEA: An evolutionary approach for motif finding in DNA sequences," Informatics in Medicine Unlocked, 2020.

[19] L. Pray, "Discovery of DNA structure and function: Watson and Crick.," Nature Education 1(1):100, 2008.

[20] G. D. Stormo, "DNA binding sites: representation and discovery," Bioinformatics, vol. 16, no. 1, pp. 16-23, 2020.

[21] V. H. Formijn and B. Ben, "Nucleotide composition of the Zika virus RNA genome and its codon usage," Virology Journal, 2016.

[22] X. Peng, C. Xingyu and R. Sanguthevar, "EMS3: An Improved Algorithm for Finding Edit-Distance Based Motifs," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 1, 2021.

[23] X. Peng, C. Xingyu and R. Sanguthevar, "Efficient Algorithms for Finding Edit-Distance Based Motifs," Algorithms for Computational Biology, pp. 212-223, 2019.

[24] K. S. Geir and D. Finn, "A survey of motif discovery methods in an integrated framework," Biology Direct, 2006.

[25] K. D. Modan and D. Ho-Kwok, "A survey of DNA motif finding algorithms," BMC Bioinformatics, vol. 8, 2007.

Enhanced Graphical Representation of Data in Web Application (Case Study: Covid-19 in the UK)

Rockson Adomah[1], Tariq Alwada'n[2], Mohammed Al Masarweh[3]
Artificial Intelligence and Information Research Group, School of Computing
Engineering & Digital Technologies, Teesside University, Middlesbrough, United Kingdom[1, 2]
Management Information System Department, College of Business in Rabigh[3]
King Abdulaziz University, Jeddah, Saudi Arabia[3]

***Abstract*—This paper describes the analysis, design, and implementation of responsive data representation in the web application that can render data asynchronously to users by making an Application Programming Interface (API) request from a webserver. At the same time, provides high-quality downloadable Scalable Vector Graphics (SVG) images for journals, magazines, and other printed media. For this issue, large-scale data that uses open-source Covid-19 data was used to improve the Covid-19 data visualization and the other improvements that can be done for proper representation of such vital data to the general public. During the development process, qualitative research into data representation with responsive charts and/or Scalable Vector Graphics images file has been conducted in contrast of each other to answer questions like what tools and technologies are often used, what are the alternative tools and technology, when, where, and why developers make use of certain approach to data representation.**

***Keywords*—Data representation; data visualization; accessibility standards; scalable vector graphics; Covid-19**

I. Introduction

Visualizing Graphical data through web Technologies has always been one of the attractive research challenges. But the objective of visualizing data in real-time (live) as it updates from a database or API server responsively on both smaller and larger devices is where the real challenge lies [1]. The objective of this paper is to conduct extensive research into data representation, examine possible ways of improvement to existing systems, and developing a robust and accurate system capable of visualizing dynamic data that is accessible through major web browsers as it updates in real-time. Extensive research conducted into the use of SVG graphical images on social medial sites and major news anchor websites during the Covid-19 era gave life to this research.

SARS- CoV-2 also known as Covid-19 is a severe acute respiratory airborne infection that is transmitted from person to person and in most cases without any sign of symptoms [2]. With all the headlines and breaking, news about Covid-19 people still need to know how their communities, cities, and countries are doing with regards to reliable data like infection rate, vaccinated population, death rate, recovery rate, etc. this data will make more sense to the general public depending on how it is being visualized. Therefore, there is the need to research the source of data, the data representation technology, and the design and implementation of such a system to load and visualize credible data using modern web development tools and methodology.

It takes either JSON structured data or XML data from a secured API server and renders the structured data as recommended by the graphical framework providers (Apex Chart) to the graphs on the dashboard in an asynchronous time.

As developers continue to develop robust and dynamic systems to represent complex data from different fields, researchers on the other hand have also been intensively investigating the challenges and variation when it comes to the use of SVG or responsive charts in data graphical data representation systems. We have therefore put together a system that seeks to use both SVG and responsive charts with modern technologies including 'HTML, CSS, and VANILA JAVA SCRIPT' to address both issues. During the development process, research into data representation with charts and SVG image files has been conducted in contrast to each other to come up with when, where, and why developers make use of these two approaches to data representation.

As different methods are being deployed to deal with the recent outbreak of Covid-19 and the need for vaccination against the different Covid-19 variants, government agencies and other health organizations have had the risen need to depend on infection and vaccination data to properly forecast future circumstances to be able to make appropriate plans for the return of normalcy to both business and day to day life [3]. This paper is aimed at investigating the use of web applications to visualize complex Covid-19 data from JSON and XML data structures to readable graphs with both historical and real-time Covid-19 data from the United Kingdom. Informative media outlets have also had to change tactics as a high volume of data from new development needs to be represented in such a way that people from all sorts of life can make sense of. Small businesses and mainstream businesses have had no other option than to shift parts of their business activities online. These and other factors have influenced the need to conduct research into graphical data representation and develop a system that will represent Covid-19 data in a readable graph asynchronously as data sources are being updated for easy access through different electronic devices of different screen sizes and a downloadable SVG file for magazines and newspaper editors to also publish such data to the general public.

This research does not only result in a robust and responsive high volume data representation system but also throws more light on some critical issues in the scope of data visualization [4]. The research indicates a lack of improvement in data visualization from a web development standpoint. Also, this research clearly shows data visualization software are more likely to produce Bitmap images as downloadable graph for print media usage which in turn produces poor quality images as oppose to scalable vector image files. Moreover, this research establishes that the statistical cost of data visualization software which ends up being embedded into the parent web application also increases the total cost of system and its maintenance. The resulted system of this research system is more reliable and effective in visualizing the data presented in an appealing manner with softness and color coding. It also provides downloadable SVG files as oppose to PNG files from similar system from gov.uk and Middlesbrough town council websites. It eliminates the use distorted images as it makes use of SVG images to provide sharp and crisp look and feel to elements in the applications. With comparatively smaller source files size data load time from API or https is relatively faster.

The rest of the paper is organized as follows: the second section introduces a small background about Data Visualization, Scalable Vector Graphics, and Progressive Web Application. Section three proposes the research method while section four introduces the data representation and development tools. Section five presents the research design and implementations and section six introduces the user interface and research results. In the last section, the conclusion and future works are introduced.

II. BACKGROUND

Data visualization or graphical data representation is a subject area that deserves more recognition given the important role it plays in the evolution of data analysis. Historically proven this subject in most cases is overlooked and this can be backed by the lack of extensive research into the subject of graphical data representation for the improvement of methods tools and technologies. The process of making complex raw data simple and clear to understand without specific academic background or expertise should be well appreciated and improved upon. Not enough research papers are targeting the subject area but almost every industry be it a government organization or private organization one way or the other makes use of data visualization as the use of a statistic structure can be extremely beneficial. This paper presents a comparative analysis of data visualization historically through time. This paper compares major technologies and tools vital to the improvement in data representation: Extensible Markup Language (XML), Scalable Vector Graphics (SVG), and Progressive Web Application (PWA).

While analyzing previous research papers, it seems that most researchers focus on what they must visualize rather than how to visualize it properly and accurately. Also, the possibility of finding an easier-to-use graphical data representation software seems to be a major problem as most stakeholders end up using whichever application is available to them. There is also a shortage of standards for measuring the performance of visualization in addition to a shortage of standardized procedures [5]. In producing a design to visualize search results for three major digital libraries, we observed that data visualization and how to measure its effectiveness is difficult and requires the collection of large user experience data. This comes as no surprise because graphical systems that render data from APIs and vector-based drawing programs can be tedious with high volume data to be structured and visualized on graphs for simpler and clear understanding. Bostock and Heer in [6] argued that the representation of data is expected to achieve some level of expressiveness, responsiveness, and flexibility.

For a high-level chart to be created designers are expected to take into consideration certain factors; considering expressiveness “Can I build it?”, efficiency “How long will it take?” and accessibility “Do I know how?” [6]. As a result of this argument, one can argue that we contend that there is still a gap between low-level graphical frameworks and high-level visualization frameworks. Numerous coordinate control graphical systems are simple to memorize but repetitive for complex work, whereas capable visualization frameworks can be threatening to amateurs or novices. Concerning vector-based drawing programs such as Adobe Illustrator approach to data representation and html5 canvas, respectively.

In another article, Romano and others in [7] gave an example with qualitative data analysis (QDA) of films review data from the box office proves to be accurate thereby providing a predictive measure of relative insight.

Bocconcino in [8] has introduces in the form of notes various initial explorations in the area of infographic and graphic visualization as information processing tools for examining, analysing and presenting data as information.

Menges et al in [9] present a technique that recognises fixed elements on Web pages and merges user viewport screenshots relative to fixed elements for an improved representation of the page.

Also, Eye tracking has been used to examine attention in many application domains, such as sports, medical, human-computer interaction studies, and commerce. [10 -13].

A. Scalable Vector Graphics

Unlike Rasta images or bitmaps, SVG is a scalable vector graphic image which means it can be indexed, compressed, scripted, and searched. This is an essential evolution to how graphical data can be represented considering the pictorial factors and scalability nature of the file. The scalability factor also makes SVG is standard developed by the Worldwide Consortium (W3C) [14]. This type of two-dimensional digital graphic is composed of mathematical equations where lines and curves images sharps and text in XML format as displayed in Fig. 1 below. SVGs are SEO-friendly. This means they are easily identified by search engines because the defined in XML text files and keywords are easily recognized.

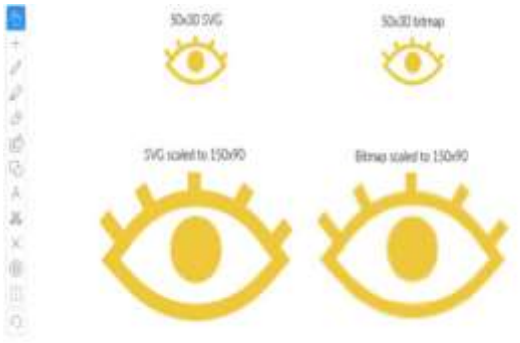


Fig. 1. SVG Scaled vs Bitmap Scaled [15].

Pixel-based images lose their resolution when enlarged as pixels are forced to expand. For vector graphics, the browser will only have to recalculate the math behind it to produce the same quality, but larger graphic as seen in Fig. 1 above. Extensive research conducted into the use of SVG in data visualization projects clearly shows that not enough studies have been done on the use of SVG's nor data visualization and this seconds the claim that data is more likely to be represented by data visualization software like Microsoft Power BI with turns to produce bitmapped images as output files as opposed to SVG file.

B. Progressive Web Application

PWAs as often called stand for the combination of established web technologies and new cutting-edge technologies to create an accessible, engaging, and reliable experience. This is to say PWA can give users the content they sort for regardless of the network conditions. Depending on the level of offline functionalities introduced during development PWAs can enhance user experience with secured HTTPS requests to render engaging data on the web applications. After extensive research, we believe this technology is worth looking into when it comes to data representation mainly because of the resilient offline capabilities [16].

III. Research Method

Observation and review of Existing works within the subject area to provide concise insight into the case study is the first step in every qualitative research analysis. Academic materials such as architecture, design pattern workflow graphics of popular data visualization software, and ideas from the review were analyzed and grouped onto a tabular form for further detailed analysis to be performed during the selection process. In qualitative research, several analysis methods can be used, for example, phenomenology, hermeneutics, grounded theory, ethnography, phonomyography, and content analysis [17]. This was done by the inclusion and exclusion method where materials or ideas relevant to the subject matter that also falls within the scope of the qualitative research approach were therefore selected and examined in contrast with similar alternatives. After the selection process both statistical analyses were performed to address the problem of less research conducted into data visualization. This study supported the assumption of decline in the study within the subject area. Moreover, an extensive content analysis was conducted on the gathered materials and existing data to support or disregard the minimal use of SVG and HTML 5 canvas in data representation works done in recent years. Another source to explore was coded from developers with may seek to represent data on data representation systems with downloadable SVG images as oppose to PNG or JPEG images.

The major instrument used in data collection can be described as observation and review of existing statistical data and or materials for research purposes. Fig. 2 below demonstrates the use of Microsoft data visualization software by Middlesbrough town council covid 19 guidance websites.

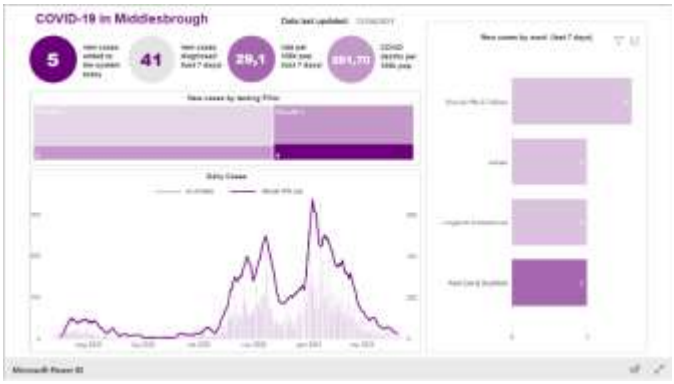


Fig. 2. Microsoft Data Visualization Software by Middlesbrough Town Council Covid 19 Guidance Websites [18].

These materials were collected to compare the technology behind the development of such a system to be well informed in addressing the case study of these projects. Statistical data from many research works done on the use of vector graphic images in data representation was also taken into consideration. There are limitations to these methods of data gathering that cannot be overlooked by any means. These limitations hinder the observation and review process in the content analysis. These include getting access to source codes which developers often regard as a danger to security, another limitation is data privacy adherence which can sometimes be a constraint to the data gathering process of the research.

Other potential issues need to be addressed when taken a qualitative approach towards research that requires a large sum of data. These potentials include relying significantly on the works of others to ascertain or make a recommendation with regards to the research problem especially when the research problem is of a different curriculum or scope. Even though Joachimiak and others in [19] argued that one of the important aspects of successful scientific research is effective data presentation and analysis.

IV. Data Representation and Development Tools

From the research conducted into data representation, design methodologies, as well as a good design practice, are to be followed algorithmically acknowledging all software development standards and ethics to produce a Covid-19 dashboard that will lay more emphasis on tourism in all for nations of the United Kingdom. Research-backed modern technology to design and develop an appealing web application that projects vaccine rollout in the United Kingdom to provide scientific guidance for reopening of the United Kingdom. The

above research suggests the lack of improvement in the design, feel, and function of existing data representation systems. Embedding SVG codes in the system to be able to produce an aesthetically pleasing system capable of downloading graphical images free from pixel enlargement.

A. Data Representation Tools

Data visualizations tools are utilized in industry to bolster decision-making additionally in the scholarly community. Within e-commerce, analytics visualization is most valuable to completely screen all the exercises and to embrace choices in time. In the industry, analytics is very useful to get the company's showcase position. As an example, competitive insights measure numerous opinions, customers, potential advertising, competitors, to reach vital early warnings. In the information technology world, there are cases of data visualization that measure the key performance index of a given operation in an ongoing project.

B. Data Representation Systems

Visualization frameworks are apparatuses planned for the unequivocal reason of information visualization, utilizing deliberations and scientific models suited to this assignment. Such instruments too commonly bolster information administration, layout algorithms, interaction, and animation. Marginally the widely used traditional data visualization systems are spreadsheet applications including MS Excel and Google Spreadsheets MS BI. This research demonstrates that it is time for developers to infuse cutting-edge technology into data visualization systems to enhance functionalities and improve efficiency.

V. Design and Implementations

An iterative and incremental approach to product delivery described in the SCRUM principles an agile project management framework was adopted in developing the system [20]. This decision was based on the nature of the problem at hand and the scope of tasks to be prioritized based on value. The future scope of this project will be based on taking into consideration what could be done to get public data graphically accessible to the public by providing an app store mobile version for apple and a play store version for android users in addition to a web-based system for general mobile, pc and tablet users.

Microsoft visual studio code [21] was selected as the main IDE for this project over the lightweight alternative sublime test because Visual Studio Code has Git integration built-in, which makes it truly simple to right away see the changes you are making in the code. On the cleared out of the sidebar, we will discover the Git symbol were ready to initialize Git as well as perform a few Git commands such as commit, drag, thrust, rebase, distribute, and see into the changes inside the record. VS Code works with any Git store local or remote and offers visual images to resolve clashes before code commits. Also, one of the key highlights of Visual Studio Code is its incredible debugging support. VS Code's built-in debugger makes a difference accelerate your edit, compile and investigate loop. By default, it comes with support for NodeJS and can debug anything that is transpired to JavaScript. Breakpoints can be set, and this easily provides identification of syntax error, see into call stack or factors at run time, and delay or step through code execution. Moreover, Vs code gives us language benefit features such as Peek Definition, Go to Definition, discover all References, and Rename Image.

These characteristics are exceptionally valuable for all software developers. Visual Studio Code makes it possible to arrange JavaScript code as well as code of other languages. All these factors help to reuse code to avoid repetition and save time. For instance, code blocks handling new recovered, new death, and cases were automatically optimized with visual studio codes "IntelliSense" feature to prompt the use of global variables instead of local variables and this could have resulted in bug creation. For reference to source code please refer to appendix A below. The system's objectives provided during the feasibility study are used as the standard from which all the work of system design is initiated. most activities involved at this stage are of a very technical nature which in turn require some level of customization. Also, a system cannot be designed without the active involvement of the user in question.

The user has an important role to play at this stage. All information gathered during the feasibility study will be used systematically during the system design. It should, however, be kept in mind detailed study of the existing system is not necessarily over even after the completion of the feasibility study.

Fig. 3 demonstrates the feasibility study conducted into the appropriate framework. Background research for graphical chart frameworks provided options that were very difficult to choose from. These include the likes of Fusion chart, chart.js, and Apex chart.

C. Selection Process of Software Development Methods

The considered Software methodologies were based on how best each methodology fits the project with regards to key factors including:

- Scalability of the project.
- Complexity of the project.
- Industry.
- Rigidity of structure.
- Stakeholder involvement.

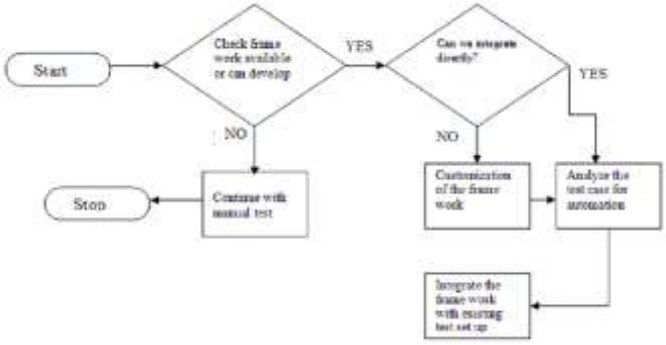


Fig. 3. Feasibility Study Process [22].

D. Information Management

The organization control structure and delivery of information is a life cycle of information management. One of the elements, which impact the growth of a new system, is the cost it will demand. There were measures set in place during the planning stage to minimize cost and these measures include the use of free-to-use open-source software, frameworks, API Data sources, and web hosting domain. The source of information, control, and the delivery mechanism used were inspired by the ISO/IEC/IEEE systems and software engineering life cycle process [23]. This process as shown in Fig. 4 describes the operational needs analysis which in this case is health statistic data that is expected to be fetched from an open-source API to be visually displayed in a graphical component design as the solution to the problem analyzed. It also deals with the implementation and verification process where data from the API has been rendered to the browser but still requires testing and validation to sustain the solution and keep the information updated as the cycle continues in a loop. The cycle involves the application of several technical activities that ensures that information is retained, structured, and securely processed.

Fig. 4. Life Cycle Information Management Process Development Model. [24].

VI. User Interface and Results

General principles of user interface design are the planning goals in making a user interface as represented in Fig. 5, Fig. 6, and Fig. 7. They are elementary to the planning and implementation of all effective interfaces as well as the user interface. These principles explained below are general characteristics of the interface and that they apply to any or all aspects. The user interface principles stand sensible even after a few years of their introduction.

The sole distinction is that they take different forms with core aspects of the principles of design. As described by Toby and others in [25], with low values for a few of the principles investigated. The study has disclosed the existence of those principles in numerous forms and the importance of their existence is for a computer program to be enticing and economical.

The related works and case studies conducted during this research helps to know the individual principles present in varieties of interfaces and conjointly the recognition of these user interfaces. This type of design approach has been termed as a user-centered approach by Stone and others in [26].

Esthetically pleasing a style aesthetic or visually pleasing composition is enticing to the attention. it attracts attention subliminally transfers a message clearly and quickly. visual attractiveness is provided by following the presentation and graphic style principles to be mentioned as well as providing a meaningful distinction between screen components making special groupings orientating screen elements providing three-dimensional illustration and exploiting color and graphics effectively. The good style combines power functionality and eases with a satisfying look.

The clarity the interface should be clear in visual appearance concept and expression. Visual components ought to be comprehensible regarding the user s real-world ideas and functions. Metaphors or analogies ought to be realistic and easy. Interface words and text ought to be straightforward unambiguous and freed from computer jargon. Fig. 6 shows the clarity of the prototype phase in mobile view as emulated by chrome immolator to depict the size of an iPhone 6,7 and 8 Plus display.

Control is feeling that the system is responding to your orders. Feeling that operating the system is not demoralizing and frustrating. The interface should present a tool-like appearance. Control is achieved when the user working at his or her own pace, can determine what to do, how to do it, and gets the task done easily. Directness Tasks should be performed directly. Available alternatives should be visible, reducing the user's mental workload. Tasks are applied by instantly choose an object, then selecting the order to be applied, and then watching the action being accomplished. The graph in Fig. 5 shows the levels of directness felt by the users for different web user interfaces [27].

Fig. 5. The Proposed DashBoard.

Fig. 6. User Interface.

Inefficiency consideration of the eye and hand movements should not be wasted. One's attention should be captured by relevant parts of the screen once required. Serial eye movements between screen parts are design to be foreseeable, obvious, and short sites should be simply scanned. All navigation ways are design to be as short as doable. Avoid frequent transitions between input devices akin to the keyboard and mouse. Familiarity with ideas and visualization that is acquainted with the user, keeping the interface natural, mimicking the user's behavior patterns, victimization real-world metaphors, stop errors from occurring by anticipating wherever mistakes might occur and coming up with to stop them. Also, allow individuals to review, change, and undo actions whenever necessary. In addition, build it tough to perform actions that may have tragic results. Once errors do occur, gift clear directions on a way to correct them and a uniform color pattern to mimic professionalism in the design process are crucial and this can be seen in Fig. 7 as the prototype from the developmental stage with a color pallet from different color code show a clear sign of unpleasantness.

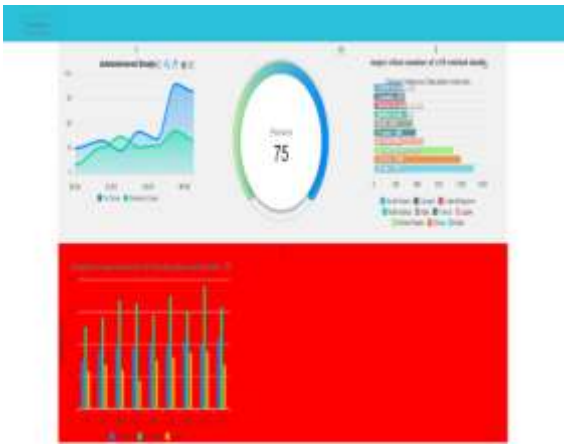


Fig. 7. One of the Resulted Screen.

Recovery must be obvious, automatic, straightforward, and natural to perform. Straightforward recovery from associate degree action greatly facilitates learning by trial and error, and exploration. Developers can build mistakes; a system ought to tolerate people who make common and inescapable mistakes. A forgiving system keeps individuals out of hassle. Individuals prefer to explore and learn by trial and error. For these reasons, the system must be subjected to aesthetical comparison to similar web applications available online. Flexibility is that the system's ability to retort to individual variations. Allowing individual users to decide on the strategy of interaction that is most acceptable to their scenario. is conjointly accomplished through allowing system customization. Responsiveness A user request should be responded to quickly. Feedback is also visual, the amendment within the form of the mouse pointer, or textual, taking the shape of a message. it should even be additive, consisting of a novel sound or tone.

This research does not only result in a robust and responsive high-volume data representation system but also throws more light on some critical issues in the scope of data visualization. The research indicates a lack of improvement in data visualization from a web development standpoint. Also, this research clearly shows data visualization software is more likely to produce Bitmap images as a downloadable graph for print media usage which in turn produces poor quality images as opposed to scalable vector image files. Moreover, this research establishes that the statistical cost of data visualization software which ends up being embedded into the parent web application also increases the total cost of the system and its maintenance.

VII. Conclusion and Future Work

This research has been successfully designed and developed to fulfill objectives that were identified during the requirements analysis. In addition to opening doors for future research into other technologies that may prove to benefit the area of study, it has also proven to practice more effectively with an artifact that is user-friendly, responsive to frame rates, and appealing to users. The system was found to be much reliable and effective in visualizing the data presented appealingly with softness and color-coding. It also provides downloadable SVG files as opposed to PNG files from a similar system from gov.UK and Middlesbrough town council websites. It eliminates the use of distorted images as it makes use of SVG images to provide the sharp and crisp look and feel to elements in the applications. With comparatively smaller source files size data load time from API or HTTPS is relatively faster. The research also highlighted the need for more interactivity in data visualization systems which we believe is worth looking into in the future. Undertaken this research project also brought to our attention the usefulness of a Progressive Web Application (PWA). This technology is new and cons like native APIs not being available on app stores currently prevented us from using this technology but hopefully, there will be more support for this technology in the future to enable data visualization apps to operate basic functionalities even without the internet connections.

REFERENCES

[1] Olshannikova, E., Ometov, A., Koucheryavy, Y., & Olsson, T. (2015). Visualizing Big Data with augmented and virtual reality: challenges and research agenda. Journal of Big Data, 2(1), 1-27.

[2] Ghinai, I., McPherson, T. D., Hunter, J. C., Kirking, H. L., Christiansen, D., Joshi, K., ... & Uyeki, T. M. (2020). First known person-to-person transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the USA. The Lancet, 395(10230), 1137-1144.

[3] Burgos, R. M., Badowski, M. E., Drwiega, E., Ghassemi, S., Griffith, N., Herald, F., ... & Michienzi, S. M. (2021). The race to a COVID-19 vaccine: Opportunities and challenges in development and distribution. Drugs in Context, 10.

[4] Georgiou, K., Mittas, N., Chatzigeorgiou, A., & Angelis, L. (2021). An empirical study of COVID-19 related posts on Stack Overflow: Topics and technologies. Journal of Systems and Software, 111089.

[5] Y. Zhu. "Measuring Effective Data Visualization", In International Symposium on Visual Computing (pp. 652-661), Springer, Berlin, Heidelberg, 2007.

[6] M. Bostock and J. Heer, "Protovis: A Graphical Toolkit For Visualization", In IEEE Transactions On Visualization And Computer Graphics, vol. 15, no. 6, pp. 1121-1128, nov.-dec, 2009.

[7] N. C. Romano Jr, C. Donovan, H. Chen and J. F. Nunamaker Jr, "A Methodology for Analyzing Web-Based Qualitative Data", Journal of Management Information Systems, 19(4), pp.213-246, 2003.

[8] B. Maurizio, "Graphic Representation and Drawing", Proceedings Conference: Immagini? Tra Rappresentazione, Comunicazione, Pedagogia e PsicologiaAt: BRIXEN 27/28.XI, 2017.

[9] R. Menges, H. Tamimi, C. Kumar, T. Walber, C. Schaefer, and S. Staab, "Enhanced representation of web pages for usability analysis with eye tracking", Conference: the 2018 ACM Symposium, 2018.

[10] Andrew T. Duchowski, "A breadth-first survey of eye-tracking applications", Behavior Research Methods, Instruments, & Computers 34, 4 (01 Nov 2002), 455–470. https://doi.org/10.3758/BF03195475.

[11] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Weijer, "Eye tracking: A comprehensive guide to methods and measures", OUP Oxford, 2011.

[12] J. Nielsen, and K. Pernice, "Eyetracking Web Usability", The first edition, New Riders Publishing, Thousand Oaks, CA, USA, 2009.

[13] A. Poole and L. J. Ball, "Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future", In Prospects, Chapter in C. Ghaoui (Ed.): Encyclopedia of Human-Computer Interaction, Pennsylvania: Idea Group, Inc., 2005.

[14] Worldwide Consortium (W3C). "Scalable Vector Graphics (SVG) 1.1", Second Edition. W3C Recommendation 16, 2011.

[15] Bogdan Sandu, "The Benefits Of Using SVG Files On Your Websites. Queness", 2020, Vieweed 02 of July,2021, <https://www.queness.com/post/17855/the-benefits-of-using-svg-files-on-your-websites>.

[16] MDN Web Docs. "Introduction To Progressive Web Apps", Developer.mozilla.org. Viewed 02 of July,2021, <https://developer.mozilla.org/en-US/docs/Web/Progressive_web_apps/Introduction>.

[17] P. Burnard, "Interpreting Text: An Alternative To Some Current Forms Of Textual Analysis In Qualitative Research", Social sciences in health, 1(4), pp.236-245, 1995.

[18] Middlesbrough Council. Data DashBoard. Middlesbrough Council. Viewed 06 of July,2021, <https://www.middlesbrough.gov.uk/community-support-and-safety/coronavirus-covid-19>.

[19] M.P. Joachimiak, J. L. Weisman, and B. May, "Jcolorgrid: Software For The Visualization Of Biological Measurements". Bmc Bioinformatics, 7(1), pp.1-5, 2006.

[20] L. Craig. "Agile and Iterative Development", A Manager's Guide, United Kingdom, Addison-Wesley, 2004.

[21] Visual Studio Code. Visual Studio Code - Code Editing. Redefined. Viewed 06 of July,2021. < https://code.visualstudio.com/>.

[22] A. Ema , and E. Reddy, "Software Test Automation: An Algorithm For Solving System Management Automation Problems", International Conference on Information and Communication Technologies (ICICT 2014), Procedia Computer Science 46, 2015, 949 – 956.

[23] Systems and software engineering — "System life cycle processes". Viewed 06 of July,2021, <https://www.iso.org/standard/63711.html>.

[24] A. Pickard, G. Roedler, and R. Madachy, "Information Management.Sebokwiki", Viewed 06 of July,2021, <https://www.sebokwiki.org/wiki/Information_Management>.

[25] B.H. Toby, "EXPGUI, a graphical user interface for GSAS". Journal of applied crystallography, 34(2), pp.210-213. 2002.

[26] D. Stone, C. Jarrett, M. Woodroffe, and S. Minocha, "User Interface Design and Evaluation. Morgan Kaufmann", The Open University, 2014.

[27] N. Bhaskar and P. P. Naidu, Ch.Kiran Babu and P. Govindarajulu. "General Principles of User Interface Design and Websites", International Journal of Software Engineering (IJSE), Volume (2) : Issue (3) : 2011.

Internet of Things Multi-protocol Interoperability with Syntactic Translation Capability

Nedaa H. Ahmed[1]
Information Systems Department
Faculty of Computers and Information
Fayoum University, Fayoum, Egypt

Haytham Al-Feel[3]
Computer Science Department
Community College
Imam AbdulRahman Bin Faisal University, Saudi Arabia

Ahmed M. Sadek[2]
Computer Science Department
Faculty of Computers and Information
Fayoum University, Fayoum, Egypt

Rania A. AbulSeoud[4]
Electronic and Communication Department
Faculty of Engineering
Fayoum University, Fayoum, Egypt

***Abstract*—Because Internet of Things (IoT) systems contain different devices, infrastructures, and data formats; its success depends on the realization of full interoperability among these systems. Interoperability is a communication challenge that affects all the layers of the system. In this paper, a transparent translator to solve interoperability issues in two layers of an IoT system is proposed. The communication protocol layer is the first layer. In this layer, it is necessary to overcome the difference between the interaction patterns, such as request/response and publish/subscribe. The second layer includes the syntactic layer, which refers to data encoding. This type of interoperability is achieved through the semantic sensor network (SSN) ontology. Tests and evaluations of the proposed translator in comparison with a similar translator were performed using the constrained application protocol (CoAP), message queuing telemetry transport (MQTT) protocol, and hypertext transfer (HTTP) protocol, in addition to different data formats, such as JSON, CSV, and XML. The results reveal the efficiency of the proposed method in terms of application protocol interoperability. In addition, the suggested translator has the added feature that it supports different data encoding standards as compared to the other translator.**

Keywords*—*Internet of things (IoT); interoperability; multiprotocol translation; message payload translation; SSN ontology

I. Introduction

The IoT comprises a collection of different devices connected using different Internet protocols. Examples of these devices include the thermostats, air conditioners, and lightbulbs that can be found in smart homes. In addition, the IoT plays an important role in other domains, such as transportation, healthcare, industrial automation, smart cities, and agriculture. The IoT enables physical objects to perform actions and share data. Therefore, IoT intelligence is bestowed on these objects by using different technologies, such as cloud computing, embedded devices, sensor networks, and Internet protocols. Because of the diversity of IoT systems, many protocols have been developed and applied. Interoperability between the different systems represents an important factor in the success of an IoT; however, it remains a significant challenge. Interoperability problems can be found in different levels, such as at the device, messaging protocol, syntactic, and semantic levels.

Interoperability on the device level refers to the wide range of devices located in an IoT. These may be high- or low-end. Examples of high-end devices include Raspberry Pi and smartphones, which have ample resources and computational capabilities, whereas low-end devices include radio frequency identification tags, sensors, and devices with constrained resources [1]. These devices may support wired or wireless networking protocols, such as Ethernet, ZigBee, Bluetooth, ZWave, 3G/4G cellular technologies, and near-field communication. In addition, these protocols can be standard communication protocols or non-standard proprietary protocols, such as long range (LoRa) and SIGFOX [1]. Sometimes, the devices that need to share information use different network technologies, requiring that that the interoperability among these different devices and network technologies be resolved to enable their integration [1].

Interoperability on the messaging protocol level refers to the multiple application protocols that exist, such as the message queue telemetry transport (MQTT) protocol, constrained application protocol (COAP), and hypertext transfer protocol (HTTP), are used to provide communications. Each protocol has characteristics that support different types of IoT applications [2]. Nevertheless, the various IoT applications should be able to exchange messages independently of messaging protocols to allow a scalable IoT architecture and cross-domain applications. Thus, the success of the interoperability of messaging protocols is manifested in a system's ability to translate between these different messaging protocols.

Syntactic interoperability refers to the fact that the content types of the data sent through the communication protocols can be of different types. Some of the most frequently used data formats are extensible markup language (XML), JavaScript object notation (JSON), and comma-separated values (CSV). The syntactic interoperability problem arises when the sender encodes the message in a specific format and the receiver can

decode received messages only in a different format. Thus, the encoding rules of the sender are incompatible with the decoding rules of the receiver, leading to mismatching of messages [1]. Therefore, this level of interoperability is important to allow a smooth transition of messages among different IoT systems.

The final type of interoperability discussed in this paper is semantic interoperability. The World Wide Web Consortium (W3C) defined it as ``enabling different agents, services, and applications to exchange information, data and knowledge in a meaningful way, on and off the web" [3]. In the traditional IoT scenario, raw sensor data from heterogeneous nodes are provided to the software agent [1]. These data contain no semantic annotations and an extensive effort is required to build intelligent applications. In addition, they may be represented in different units of measurements and have additional information [1], resulting in semantic interoperability problems. These semantic problems between data and information models render IoT applications unable to interoperate both automatically and dynamically because their descriptions and understanding of resources differ [1].

The current middlewares used to achieve application protocol interoperability have limitations, such as adding interoperability problems when working in conjunction with another middleware. Also, some proxies are used to solve the same level of interoperability but have issues like low bandwidth, low processing, and high cost of management. On the other hand, the current works for solving syntactic interoperability problems are very few. The existing solutions can convert between encodings with similar syntax only. Also, there is no current solution that achieves both application protocol and syntactic interoperability together.

In this study, a software architecture was designed to solve the interoperability problems related to both messaging protocols and syntactic levels. The main contributions of this paper are as follows.

1) An IoT translator that can achieve communication protocol and syntactic interoperability is proposed. Semantic interoperability will be addressed in a future paper.

2) The development of a multi-protocol translator that can translate messages among the CoAP, MQTT, and HTTP protocols is described.

3) The use of the semantic sensor network (SSN) ontology to allow conversion among XML, JSON, and CSV data formats to achieve syntactic interoperability is described. This will enable clients to obtain the data they need in any required format, even if they are stored in different formats on the server.

4) A translator based on a hub-and-spoke model, which supports scalability and modularity, is presented. The scalability and modularity will allow the translator to be extended to support more protocols easily, in addition to more data formats.

5) An evaluation is presented that shows the effectiveness of the proposed translator in comparison with the Arrowhead translator.

The remainder of the paper is organized as follows. In Section II, some related studies on solutions for interoperability problems in the IoT are reviewed. Sections III and IV discuss the software architecture and implementation of the proposed solution, respectively. The results and evaluations are presented in Section V. Finally, the conclusions and suggestions for future work are provided in Section VI.

II. Related Work

Different architectures and frameworks are used to solve IoT interoperability on the various levels.

To address syntactic interoperability, Palm et al. [4] presented a theoretical method for translating message payloads among different endpoints. First, this method constructs a syntax tree from the incoming message. Then, it converts the syntax tree into an equivalent syntax tree of the target encoding. The syntactic translation can convert only between an encoding standard and intersecting syntaxes.

To provide messaging protocol interoperability, Derhamy et al. [5] proposed a transparent protocol translator to allow interoperability between communication protocols. This translator depends on a service-oriented architecture (SOA), not on middleware. Thus, it supports low latency and operates transparently. It is also secured through the use of Arrowhead authorization and authentication [6]. Its architecture consists of two spokes and a central hub: the first spoke operates as a service provider spoke and the second as a service consumer spoke. The translator can support any number of protocols, each of which has only two spokes. The authors tested their architecture on the CoAP and HTTP protocols and determined that it was faster than the Californium proxy [7].

Lee et al. [8] proposed an IoT framework based on the software-defined network (SDN) that can intercept all packets from CoAP to MQTT and vice versa. They defined URL rules to specify the resource or the topic and distinguish between homogeneous (e.g., from MQTT client to MQTT client) and heterogeneous (e.g., from MQTT client to CoAP client) traffic. In the homogeneous scenario, the SDN ignores the traffic and these packets are operated as in the original scenario. In the heterogeneous scenario, the SDN switch delivers the packets to the SDN controller and redirects them to the cross proxy for translation. The advantage of this framework is that it causes no delay in a homogeneous scenario. However, the authors provided no evaluation results for their suggested framework.

Ponte [9] is an Eclipse IoT project that provides open APIs to create applications that support the CoAP, HTTP, and MQTT communication protocols. Ponte provided a centralized solution to enable clients using different communication protocols to communicate easily with each other. Data collected from the three different protocols are stored in SQL or NoSQL databases. Therefore, all the clients can access all resources, regardless of the communication protocol they use. In the same direction, Khaled and Helal [10] proposed the Atlas IoT framework to allow communication between clients using MQTT, CoAP, and HTTP. The proposed protocol translator can be deployed on either a cloud infrastructure or the IoT device itself. This framework depends on the IoT device description language [11] and was compared with Ponte

[9]. The results show that energy consumption is reduced in comparison with that of Ponte [9]. A disadvantage of this framework is that not all IoT devices can be part of this interoperable ecosystem: the device should have an Internet connection (e.g., Ethernet, cellular network, or Wi-Fi) and an operating system that supports multithreading to integrate with this ecosystem.

Desai et al. [2] proposed the Semantic Gateway as Service architecture to achieve messaging protocol interoperability among the CoAP, MQTT and XMPP protocols using a multi-protocol proxy. Semantic reasoning using the SSN ontology to solve interoperability was highlighted, but no description of implementation details or evaluation methodology were provided.

To address semantic level interoperability, Gyrard et al. [12] proposed the machine-to-machine measurement (M3) framework to develop semantic-based cross-domain IoT projects and reuse the optimum number of ontologies and rules. In their study, they focused on designing the Linked Open Vocabularies for the Internet of Things dataset to reference and classify semantic-based projects that are relevant to the IoT. In addition, they designed a sensor-based linked open rules dataset of domain rules to infer high-level abstractions from sensor data. David Perez et al. [13] developed an ontology for the smart city scenario, specifically for the SusCity project [14], to facilitate the management of the infrastructure. This ontology consists of several main classes, such as IoT infrastructures, devices, communication interfaces and links, and performance metrics. Evaluations proved the correctness of this ontology. One of its disadvantages is that an automatic ontology update mechanism is required. Gyrard and Serrano [15] presented a methodology called SEG 3.0, a name which comes from segmentation and Web 3.0, which depends on semantic Web technologies. They defined the characteristics and steps of this methodology and subsequently implemented a framework for applying it. The purpose of this framework is to achieve semantic interoperability among IoT projects. In addition, they investigated various use cases to show the correctness of the methodology and that it can be applied to other domains, such as smart cities. Kleine et al. [16] developed a Semantic Web of Things architecture including virtual sensors, smart service proxy, and semantic entities to measure the traffic density of a road. This architecture depends on smartphones located in vehicles moving in the network and represents the sensor data in resource description framework (RDF) form. In the future, the authors plan to enable selective privacy to identify the exact vehicles traveling on certain road sections. Additionally, Kamilaris et al. [17] proposed an eco-system for urban computing, using the concept of the Web of Things together with event processing, mobile computing, semantic Web, and big data analysis techniques to record the real information of smart cities for their residents. They conducted a case study in the city of Aarhus, Denmark. One of the disadvantages of the suggested ecosystem is its lack of privacy or security aspects.

Kamilaris et al. [18] proposed a semantic framework called Agr-IoT for smart farming applications. This framework uses semantic Web techniques to allow reasoning and to facilitate increased information collection and more accurate decision making. In addition, semantic Web techniques help achieve interoperability among different data sources, such as sensors, social media, governmental alerts, connected farms, and regulations. The ontologies used in the framework are the SSN ontology, complex event service ontology, and an agriculture ontology, called AgOnt.

In the healthcare domain, Zgheib et al. [19] presented an IoT system to detect the risk of bedsores, using SSN ontology to achieve interoperability between system components. This system is based on message-oriented middleware to process some constraints, such as the security, scalability, and privacy of medical information.

The above review shows that the methods that achieve messaging protocol and syntactic interoperability remain few in number. Therefore, the proposed architecture is focused on these two levels of interoperability.

III. Proposed Architecture

The proposed architecture is based on that presented in [5]. As shown in Fig. 1, it consists of three parts: the clients, translator, and servers. The clients are service consumers who send different requests to different servers. They can also receive data in any format they need, even if these data are stored in a different format on the server. The proposed translator is used to allow a client using a specific protocol to communicate with a server using a different protocol. It consists of a hub and multiple spokes. As this translator can translate among three different protocols, it has six spokes. It can support additional protocols by adding only the two spokes needed for each of these protocols. At each operation, only two spokes are used according to the request. The hub is a conceptual representation and is represented using an intermediate format and SSN ontology. The intermediate format is used to convert one protocol to another. Each protocol is represented by two spokes. The first spoke is a server spoke, which listens to different requests on a specific port, and the second is the client spoke, which creates a new request according to the information received from the intermediate format in the hub. Inside the hub, the SSN ontology is also located, which is used to describe sensors and their observations, the features of interest, the observed properties, and the actuators [20]. It is used only to achieve syntactic interoperability and is converted into different data formats. The final part of the architecture is the server, which contains the service providers that serve different clients. The proposed translator can be used with different clients and with servers acquired from different vendors.

The following scenarios clarify the architecture.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Request from Representational State Transfer (REST) Client to REST Server

In this scenario, the representational state transfer (REST) points to the HTTP and CoAP protocols. Therefore, this scenario illustrates the events that occur when a CoAP client sends a request to an HTTP server or an HTTP client sends a

request to a CoAP server. As shown in Fig. 2, the server spoke, a CoAP server in this case, listens to the requests from a CoAP service consumer. A CoAP request is converted to the appropriate HTTP request using the intermediate format. Subsequently, the HTTP request is forwarded to the HTTP service provider.

If the syntactic interoperability service is required, the SSN ontology is used. In this case, if the request is PUT or POST, an INSERT or UPDATE statement modifies the SSN ontology. However, if the request is GET, there exist two situations. The first is where the service consumer requires the payload of a specific resource in the same format as that of the service provider. In this case, the service provider replies directly with the payload to the HTTP client spoke without needing to use the SSN ontology. When the HTTP client spoke receives the response, a CoAP response is generated by the CoAP server spoke using the intermediate format and is forwarded to the CoAP consumer. In the second situation, the service consumer requires the payload of a specific resource in a format different from that which exists in the service provider. In this case, the service provider replies with a "NOT ACCEPTABLE" code to the HTTP client spoke. If the CoAP server spoke receives a "NOT ACCEPTABLE" message, it generates a SELECT query to obtain the specified resource data in a standard format, converts this format to the required format, and finally responds to the CoAP client consumer with these data. The same steps are taken when the HTTP client requires a service from the CoAP server. The only difference is that the service consumer in this case is an HTTP client and the service provider is a CoAP server. An additional difference is that, inside the translator, the two spokes that should be used in this case are the HTTP server and the CoAP client spokes. The mapping between the HTTP and CoAP protocols is shown in Table I. However, it is important to note that the CoAP protocol contains an OBSERVE request that does not exist in the HTTP protocol. The CoAP OBSERVE request was used only with the MQTT protocol.

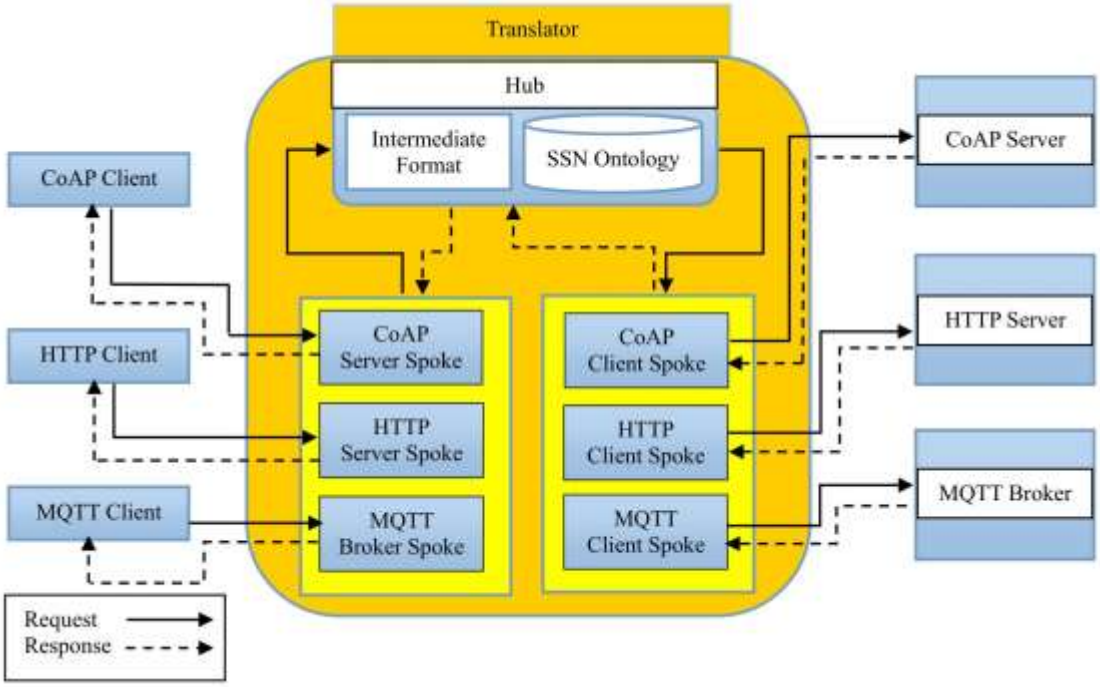


Fig. 1. Proposed Architecture.

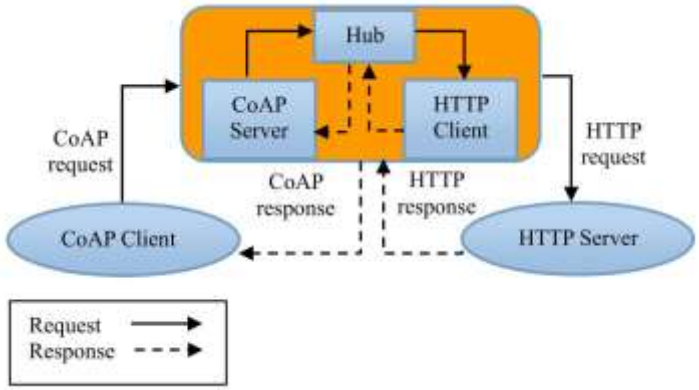


Fig. 2. Request from Constrained Application Protocol Client to Hypertext Transfer Protocol Server.

B. Publishing a Resource in Representational State Transfer Server (Constrained Application Protocol or Hypertext Transfer Protocol)

As shown in Fig. 3, in this scenario the service consumer is an MQTT client and the service provider is a REST server. Thus, the MQTT client must publish a message on any resource in the REST server.

The MQTT broker spoke listens to the PUBLISH messages from the MQTT service consumer. The MQTT message is converted to the PUT REST request using the intermediate format in the hub. Subsequently, the PUT request is forwarded to the REST service provider. In addition, the SSN ontology is required only if the syntactic interoperability is activated. In that case, the MQTT server spoke updates the value of the resource in the SSN ontology. Note that UPDATE query exists in the SPARQL query language, and therefore, DELETE and INSERT queries must be used together to update the data in the SSN ontology. As the MQTT protocol does not contain a content format field in its header, the content format is checked manually. According to the content format, it is possible to determine the manner in which the message should be processed to extract the resource or topic name with its payload and store them in the SSN ontology. The pseudocode for this is shown in Algorithm 1. When the REST client spoke receives a response, an MQTT message response is generated by the MQTT server spoke and forwarded to the MQTT publisher. The response of the PUBLISH message can be either a None or PUBACK packet or a PUBREC packet according to the quality of service (QoS) level [21].

C. Subscribing a Resource in Constrained Application Protocol Server

The subscription of a message to the CoAP server differs slightly from that to the HTTP server, because the HTTP protocol does not have an OBSERVE request as does the CoAP server. As shown in Fig. 4, the main steps in this scenario are the same as described previously. The difference is that the SUBSCRIBE message in the MQTT protocol corresponds to the OBSERVE request in the CoAP protocol. As the MQTT protocol does not contain the content format field in its header, it is not necessary to use the SSN ontology, although syntactic interoperability is activated.

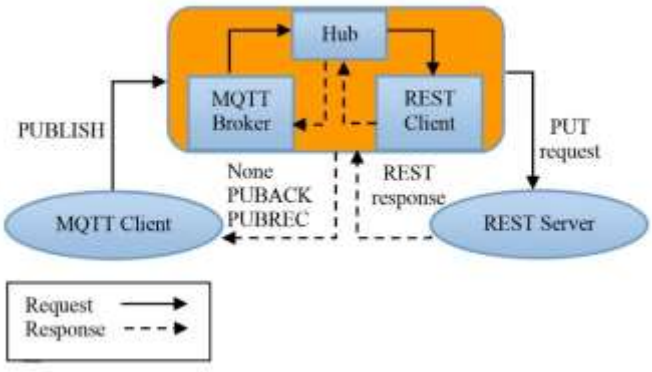


Fig. 3. PUBLISH Message from Message Queue Telemetry Transport Client to Representational State Transfer Server.

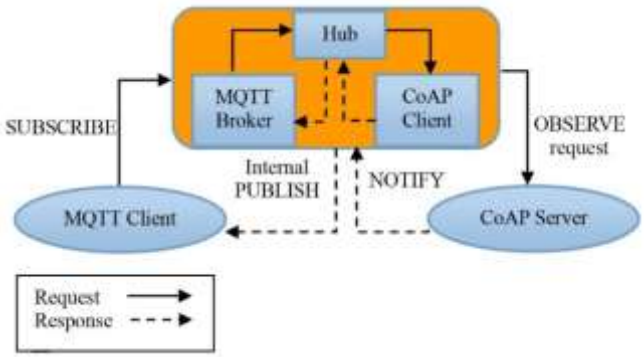


Fig. 4. SUBSCRIBE Message from Message Queue Telemetry Client to Constrained Application Protocol Constrained Application Protocol Server.

D. Subscribing a Resource in the Hypertext Transfer Protocol Server

As mentioned above, the HTTP protocol does not have an OBSERVE request. Therefore, the GET request, which is implemented periodically, is used. As shown in Fig. 5, when the MQTT client subscribes to a specific resource in the HTTP server, an HTTP client makes a GET request with a pre-configured periodic time. If the returned payload is different from the last payload, the MQTT broker publishes internally.

Algorithm 1: Checking the content format for the MQTT protocol

```
Input: A string variable message which is MQTT payload
Output: An integer variable type which is the format of MQTT payload
initialization;
if message begins with "{" and ends with "}" then
 type ←0; // zero means it is JSON format
end
else if message begins with "<" and ends with ">" then
 type ←1; // one means it is XML format
end
else if the number of commas is equal in each line then
 type ←2; // two means it is CSV format
end
else
 type ←3; // three means it is not supported format
end
return type
```

with the new payload to the MQTT client. The periodic GET request would be canceled if the MQTT client unsubscribed on this topic.

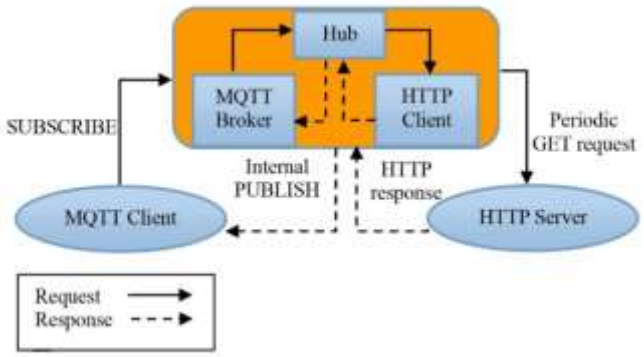


Fig. 5. SUBSCRIBE Message from Message Queue Telemetry Transport Client to Hypertext Transfer Protocol Server.

IV. Implementation

The architecture was implemented using the JAVA language, and the spokes were implemented using the Java libraries. For example, the CoAP spokes, for both client and server, were implemented using the Californium CoAP library [7]. The CoAP server spoke contained only the CoAP server with RootResource [7]. The CoAP client spoke was used to perform the CoAP request and return the response to the calling spoke. For the MQTT protocol, the MQTT server spoke was implemented using the Mosquitto broker [22], and the MQTT client spoke was implemented using the Eclipse.

Paho client developed by Eclipse Foundation; however, the Jersey open-source libraries created by Eclipse Foundation and Oracle Corporation for the HTTP protocol were used to implement the HTTP server and HTTP client spokes.

A. Mapping among different Protocols

The mapping among the three different protocol spokes was implemented as shown in Table I.

B. Intermediate Format

The intermediate format was used in the hub to enable interchanges between the protocol spokes. It holds the basic header fields of request and response. The structure of the intermediate format is presented in Table II.

C. Semantic Sensor Network Ontology

The SSN ontology [20] is an ontology developed by W3C to provide standard modeling for sensor devices, actuators, sensor platforms, their observations, and so on. It was used in the RDF format [23]. The purpose of using the SSN ontology is to achieve syntactic and semantic interoperability. Here, the syntactic interoperability was achieved to solve the problem of encoding and decoding messages between the sender and receiver in different formats. By doing so, if, for example, the payload is stored in the plaintext format at the REST server, the REST client can obtain the payload in different formats, such as JSON or XML. This study converted three different formats: CSV, JSON, and XML. In Fig. 6, a subset of this ontology is shown. It represents the classes and properties that used in the proposed architecture.

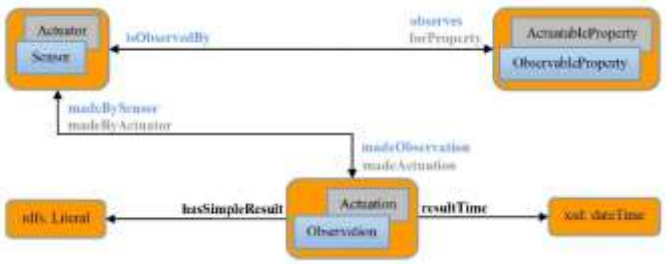


Fig. 6. Subset of Semantic Sensor Network Ontology.

TABLE I. Mapping among Three Protocols

	HTTP	CoAP	MQTT
Code Request	GET request in case of mapping with GET request of CoAP and with the MQTT SUBSCRIBE.	GET request in case of mapping with GET request of HTTP. OBSERVE request in case of mapping with MQTT SUBSCRIBE	SUBSCRIBE
	PUT request	PUT request	PUBLISH retained message
	POST request	POST request	PUBLISH retained message
	DELETE request	DELETE request	PUBLISH retained message with zero-byte payload
Code Response	“200” OK	“2.00” OK	0x00 - Success - Maximum QoS 0 0x01 - Success - Maximum QoS 1 0x02 - Success - Maximum QoS 2
	“404” NOT FOUND	“4.04” NOT FOUND	Not supported
	“406” NOT ACCEPTABLE	“4.06” NOT ACCEPTABLE	Not supported
	“204” NO CONTENT	“2.04” CHANGED	Not supported
Code Error	“400” BAD REQUEST	“4.00” BAD REQUEST	0x02 Connection Refused, identifier rejected
	“401” UNAUTHORIZED	“4.01” UNAUTHORIZED	0x05 Connection Refused, not authorized
	“500” INTERNAL SERVER ERROR	5.00” INTERNAL SERVER ERROR	0x03 Connection Refused, Server unavailable
Object	Resource name	Resource name	Topic name

TABLE II. Definition of Intermediate Format

Variable Name	Type	Discussion
uniqueKey	int	is the message Id
Code	String	is the CRUD operation in case of request and response code or error code in case of response
Object	String	is the resource or topic to be operated on
Payload	String	is the body of the message
payloadFormat	String	is the format of the payload

V. Testing and Evaluation

In this simulation, a CoAP server, HTTP server and MQTT broker were used in a weather information service transmitting information from different geographic locations. This service measured the temperature, humidity, pressure, wind direction, and wind speed. The HTTP and CoAP protocols could support different formats of the payload, such as plaintext, CSV, JSON, and XML formats. However, in the MQTT protocol, the content format was application-specific, as this protocol does not contain the content format or accept fields in its header. If the service used one sensor and provided only the value, the plaintext format was used; otherwise, any one of the three formats was used. The XML, CSV, and JSON payload structures are shown in Fig. 7(a), Fig. 7(b), and Fig. 7(c), respectively. The approximate lengths of each format are listed in Table III.

For converting the payload from one format to another, the intermediary ontology was used, as it represents the data in a structured form.

The delay caused by the translator for protocol translation and format conversion was measured and evaluated. All the tested scenarios are shown in Fig. 8. In test 8-a), a CoAP request was generated from a CoAP client to the translator, which then generated the corresponding HTTP request to the HTTP server.

Test 8-b) followed the form of test 8-a), except that the translator generated the corresponding MQTT message to the MQTT broker. However, test 8-c) involved an HTTP request generated from an HTTP client to the translator, which then generated the corresponding CoAP request to the CoAP server. Test 8-d) was similar to test 8-c), except that the translator generated the corresponding MQTT message to the MQTT broker. Test 8-e) involved an MQTT message generated from an MQTT client to the translator, which then generated the corresponding HTTP request to the HTTP server. Finally, test 8-f) followed the form of test 8-e), except that the translator generated the corresponding CoAP request to the CoAP server.

The translator was run on a laptop with an Intel Core i5-2520M processor running Windows 8 at 2.50 GHz and 4.00 GB RAM. The delay introduced by the translator and the delay caused by using the SSN ontology were measured. As shown in Fig. 9, six Java milliseconds timers (t_1 - t_6) were used to compute these delays. The descriptions of these timers are listed in Table IV.

The following equations were used in the delay calculations, where (1) represents the time the packet takes within the translator and (2) the time required to perform processing on the different formats of data plus the execution time of the SPARQL query.

$$D_{translate} = (t_2 - t_1) + (t_4 - t_3) \quad (1)$$

$$D_{ssnOnt} = t_6 - t_5 \quad (2)$$

The tests were performed 1000 times per scenario. The average time required for protocol translation is summarized in Table V.

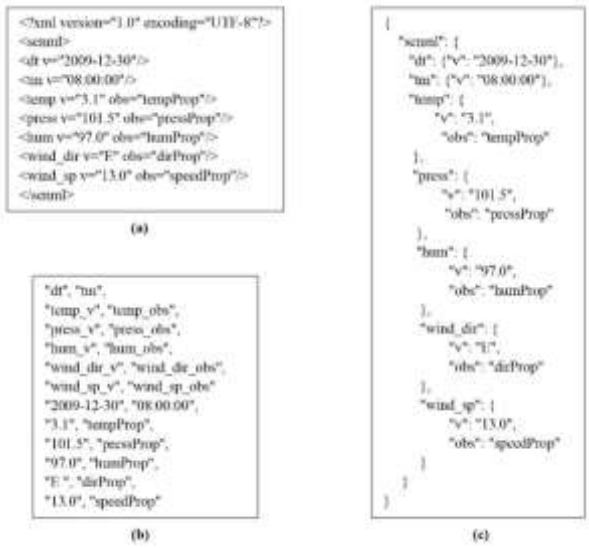


Fig. 7. Different Payload Structure Formats.

TABLE III. Length of Payloads in Bytes

The Payload Structure	Length (bytes)
XML	236
JSON	246
CSV	234

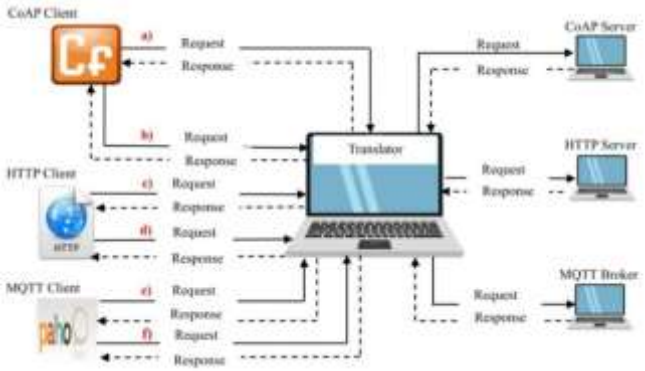


Fig. 8. Test Scenarios.

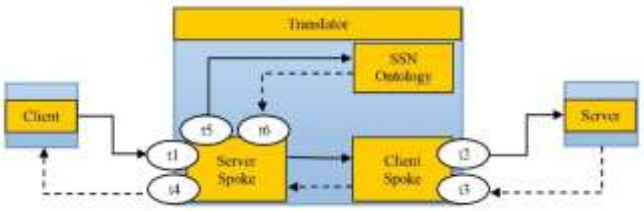


Fig. 9. Time Stamp for Delay Measurements.

As shown in Table V, the time required to convert the protocols is very short. Scenarios (e) and (f) have the minimum delay because the server spoke in this case is the MQTT broker. The MQTT broker operates as a pipeline and does not perform any processing, as do HTTP and CoAP servers. Therefore, the MQTT broker is very simple as compared to the CoAP and HTTP servers.

TABLE IV. TIMING INSTRUMENTATION

t_1	=	Request arrives at the application server spoke of the translator.
t_2	=	Request leaves the application client spoke of the translator.
t_3	=	Response arrives at the application client spoke of the translator.
t_4	=	Response leaves the application server spoke of the translator.
t_5	=	Query leaves the server to the SSN ontology plus preprocessing the data if needed.
t_6	=	Response from the SSN ontology plus processing the data if needed.

TABLE V. AVERAGE DELAY OF THE PROTOCOL TRANSLATION

Scenario	Client	Server	Method	Delay (ms)
a)	CoAP Client	HTTP Server	GET	3.96
			PUT	4.03
b)	CoAP Client	MQTT Broker	GET	3.10
			PUT	3.05
c)	HTTP Client	CoAP Server	GET	3.21
			PUT	3.38
d)	HTTP Client	MQTT Broker	GET	3.81
			PUT	3.09
e)	MQTT Client	HTTP Server	Subscribe	3.10
			Publish	2.64
f)	MQTT Client	CoAP Server	Subscribe	2.45
			Publish	2.21

TABLE VI. AVERAGE PROCESSING TIME

Scenario	From	To	Delay (ms)
1)	JSON	SSN	23.69
2)	XML		23.57
3)	CSV		30.19
4)	SSN	JSON	11.78
5)		XML	11.86
6)		CSV	11.71

In Table VI, the average computed delay in achieving syntactic interoperability is shown. Scenarios 1, 2, and 3 represent the delay in processing and updating the data in the SSN ontology. These three scenarios are used when performing a POST, PUT, or PUBLISH request. Scenarios 4, 5, and 6 represent the average delays caused by selecting the data from the SSN ontology and converting it to the required format. These scenarios were used when performing GET requests. As can be seen in this table, updating the ontology (Scenarios 1–3) consumes more time than selecting the data from it (Scenarios 4–6). However, this delay is acceptable in IoT applications and does not affect the performance of the architecture. Using the SSN ontology, different clients can receive the payload in any format they require, even if this payload exists in the server in a different content format.

VI. DISCUSSION

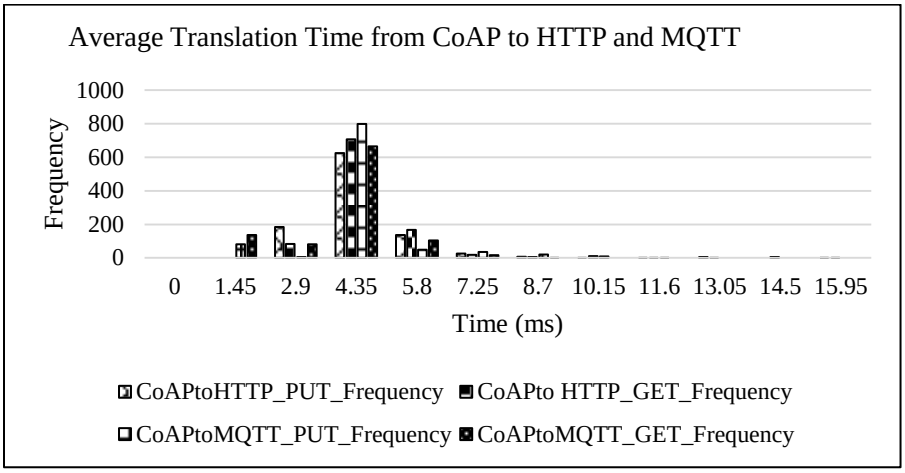
To validate the efficiency of the suggested translator, the delay caused by the proposed translator was compared with that caused by the Arrowhead translator [5]. The comparison was applied only to scenario (c) in Fig. 8, as the authors implemented only this scenario. The Arrowhead translator was run on hardware that was different from that used here on the proposed translator. For this reason, it was unavailable to compare the proposed translator with the Arrowhead translator directly. In the case of the Arrowhead translator, the authors measured and evaluated the delay introduced in comparison with the Californium proxy [22]. They ran the Californium proxy on the same hardware they used for running their proposed translator. The same procedure was followed to evaluate the proposed translator in comparison with the Arrowhead translator.

As can be seen in Table VII, the Arrowhead translator's delay is about 43.5\% of that of the Californium proxy when implemented on the same platform, whereas the suggested translator`s delay is only about 23.6\% of that of the Californium proxy on the same platform. In addition, the proposed translator can map between different data encoding standards, a capability that is not available in the Arrowhead translator. In Fig. 10 and 11, the results of 1000 requests for each scenario are shown.

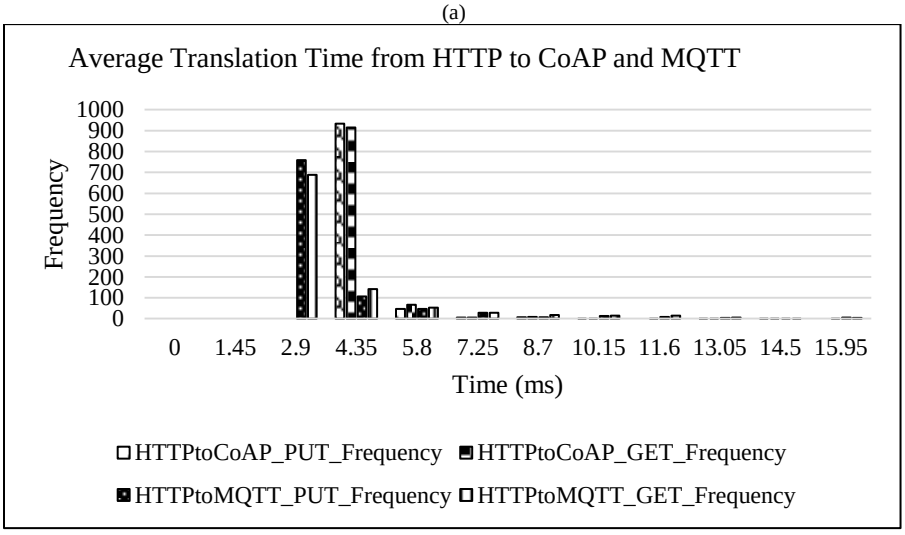
The histogram charts show that there exists some anomaly. However, this anomaly is due to the Java libraries that were used, and therefore, it was beyond control.

TABLE VII. COMPARISON OF THE PROPOSED AND THE ARROWHEAD TRANSLATOR

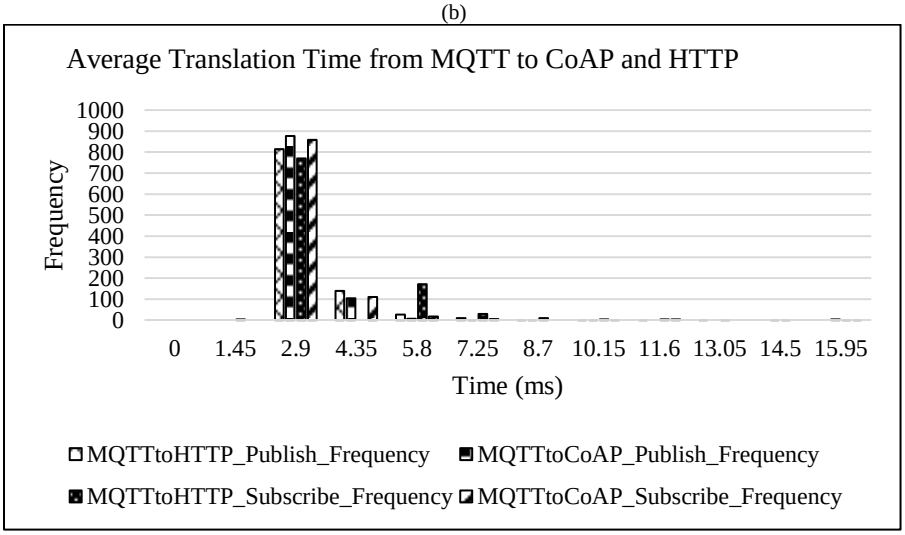
The suggested Translator		The Arrowhead Translator	
Delay of it (ms)	Delay of CP on the same hardware (ms)		Delay of it (ms)
3.21	13.59	177	77



(a)

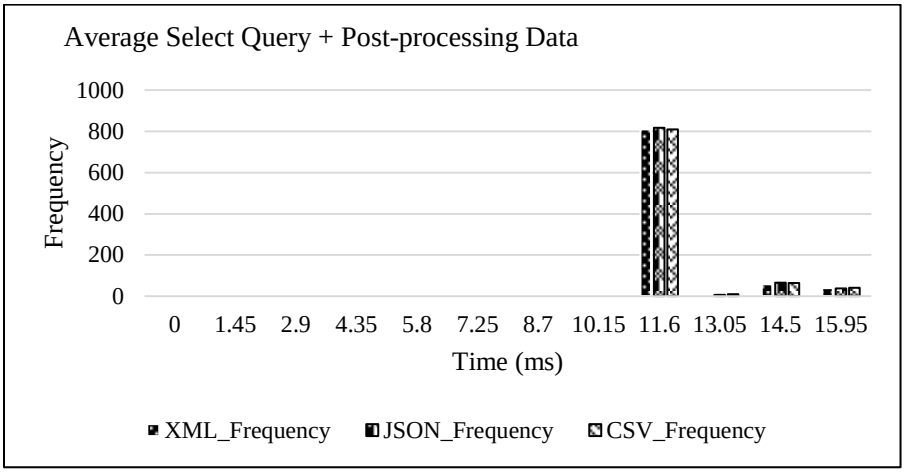


(b)

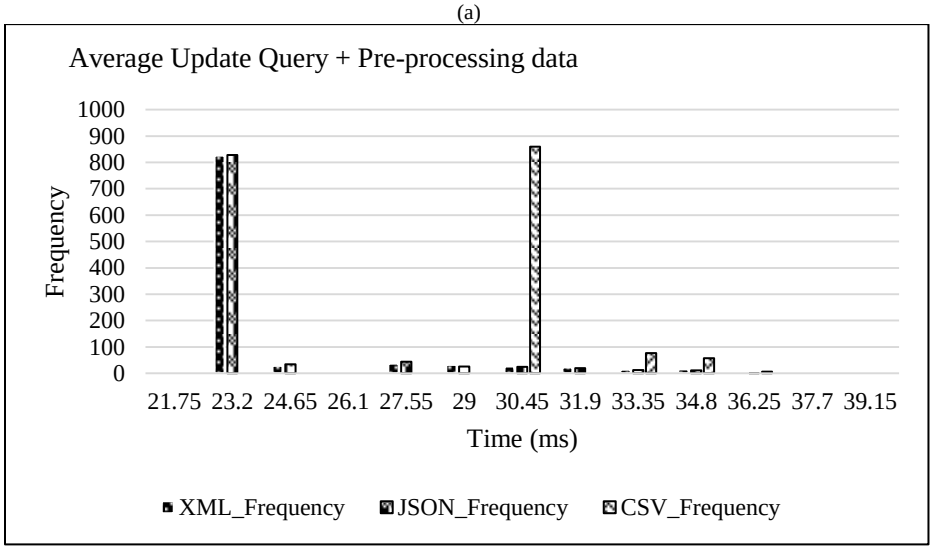


(c)

Fig. 10. Average Translation Time among different Protocols.



(a)



(b)

Fig. 11. Average Delay of Query on Semantic Sensor Network Ontology and Processing Data.

VII. Conclusion

One of the greatest challenges in the IoT is to achieve interoperability. A transparent and reliable architecture based on SOAs was suggested in this paper. The goal was to achieve protocol and syntactic interoperability using the SSN ontology. The proposed architecture can integrate different IoT application protocols, such as MQTT, HTTP, and CoAP. In addition, it can convert between different payload formats, such as JSON, XML, and CSV.

The evaluations showed that the proposed translator's performance is better than that of the Arrowhead translator and introduced a shorter delay. The proposed translator`s delay is approximately 19.9\% of that of the Arrowhead translator. This difference in delay is due to the simplicity of the implementation. In addition, the suggested translator has an advantage over the Arrowhead translator in that it can achieve syntactic interoperability, unlike the Arrowhead translator using SSN ontology.

Future work will attempt to achieve semantic interoperability using the current architecture by reasoning the data stored in the SSN ontology. For this reason, it was preferred to use semantic Web technology to achieve syntactic interoperability rather than any other database. This allowed us to achieve two types of interoperability using one technology without introducing a long delay.

References

[1] M. Noura, M. Atiquzzaman, and M. Gaedke, "Interoperability in internet of things: taxonomies and open challenges," Mob. Networks Appl., vol. 24, no. 3, pp. 796–809, 2019.

[2] P. Desai, A. Sheth, P. Anantharam, "Semantic gateway as a service architecture for IoT interoperability," Proc. - 2015 IEEE 3rd Int. Conf. Mob. Serv. MS 2015, pp. 313–319, Aug. 2015.

[3] A. D. P. Venceslau, R. M. C. Andrade, V. M. P. Vidal, T. P. Nogueira, and V. M. Pequeno, "IoT semantic interoperability: A systematic mapping study," ICEIS 2019 - Proc. 21st Int. Conf. Enterp. Inf. Syst., vol. 1, no. Iceis, pp. 523–532, 2019.

[4] E. Palm, C. Paniagua, U. Bodin, O. Schel´en, "Syntactic translation of message payloads between at least partially equivalent encodings," Proc. IEEE Int. Conf. Ind. Technol., vol. 2019-Febru, pp. 812–817, 2019.

[5] H. Derhamy, J. Eliasson, J. Delsing, "IoT interoperability—on-demand and low latency transparent multiprotocol translator," IEEE Internet Things J., vol. 4, no. 5, pp. 1754–1763, 2017.

[6] P. Varga et al., "Making system of systems interoperable – The core components of the arrowhead framework," J. Netw. Comput. Appl., vol. 81, pp. 85–95, Mar. 2017.

[7] M. Kovatsch, M. Lanter, Z. Shelby, "Californium: scalable cloud services for the internet of things with coap," 2014 Int. Conf. Internet Things, IOT 2014, pp. 1–6, 2014.

[8] C.-H. Lee, Y.-W. Chang, C.-C. Chuang, Y. H. Lai, "Interoperability enhancement for internet of things protocols based on software-defined network," 2016 IEEE 5th Glob. Conf. Consum. Electron. GCCE 2016, pp. 4–5, 2016.

[9] M. Dave, M. Patel, J. Doshi, and H. Arolkar, "Ponte Message Broker Bridge Configuration Using MQTT and CoAP Protocol for Interoperability of IoT," Commun. Comput. Inf. Sci., vol. 1235 CCIS, pp. 184–195, Mar. 2020.

[10] A. E. Khaled, S. Helal, "Interoperable communication framework for bridging restful and topic-based communication in IoT," Futur. Gener. Comput. Syst., vol. 92, pp. 628–643, 2019.

[11] A. E. Khaled, A. Helal, W. Lindquist, C. Lee, "Iot-ddl–device description language for the "t" in IoT," IEEE Access, vol. 6, pp. 24048–24063, 2018.

[12] A. Gyrard, C. Bonnet, K. Boudaoud, M. Serrano, "LOV4IoT:a second life for ontology-based domain knowledge to build semantic web of things applications," Proc. - 2016 IEEE 4th Int. Conf. Futur. Internet Things Cloud, FiCloud 2016, pp. 254–261, 2016.

[13] D. P. Abreu, K. Velasquez, A. M. Pinto, M. Curado, E. Mon-teiro, "Describing the internet of things with an ontology: The suscity project case study," Proc. 2017 20th Conf. Innov. Clouds, Internet Networks, ICIN 2017, pp. 294–299, 2017.

[14] J. Fernandes et al., "Building a smart city IoT platform - the suscity approach", 48nd Spanish Congr. Acoust. Iber. Encount. Acoust., pp. 557–566, 2017.

[15] A. Gyrard, M. Serrano, "Connected smart cities: interoperability with seg 3.0 for the internet of things," Proc. - IEEE 30th Int. Conf. Adv. Inf. Netw. Appl. Work. WAINA 2016, no. 2, pp. 796–802, 2016.

[16] O. Kleine, S. Ebers, M. Leggieri, "Monitoring urban traffic using semantic web services on smartphones - a case study," 2015 12th Annu. IEEE Int. Conf. Sensing, Commun. Netw. - Work. SECON Work. 2015, pp. 1–6, 2015.

[17] A. Kamilaris, A. Pitsillides, F. X. Prenafeta-Bold, M. I. Ali, "A web of things based eco-system for urban computing-towards smarter cities," Proc. 24th Int. Conf. Telecommun. Intell. Every Form, ICT 2017, 2017.

[18] A. Kamilaris, F. Gao, F. X. Prenafeta-Boldu, M. I. Ali, "Agri-IoT: a semantic framework for internet of things-enabled smart farming applications," 2016 IEEE 3rd World Forum Internet Things, WF-IoT 2016, pp. 442–447, 2017.

[19] R. Zgheib, R. Bastide, E. Conchon, "A semantic web-of-things architecture for monitoring the risk of bedsores," Proc. - 2015 Int. Conf. Comput. Sci. Comput. Intell. CSCI 2015, pp. 318–323, 2016.

[20] M. Compton et al., "The SSN ontology of the W3C semantic sensor network incubator group," J. Web Semant., vol. 17, pp. 25–32, Dec. 2012.

[21] J. Toldinas, B. Lozinskis, E. Baranauskas, and A. Dobrovolskis, "MQTT Quality of Service versus Energy Consumption," Proc. 23rd Int. Conf. Electron. 2019, Electron. 2019, Jun. 2019.

[22] R. A. Light, "Mosquitto : server and client implementation of the MQTT protocol," J. Open Source Softw., vol. 2, pp. 1–2, 2017.

[23] J. Z. Pan, "Resource Description Framework," in Handbook on Ontologies, Springer, Berlin, Heidelberg, 2009, pp. 71–90.

Comparing SMOTE Family Techniques in Predicting Insurance Premium Defaulting using Machine Learning Models

Mohamed Hanafy Kotb[1], Ruixing Ming[2]
School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, China[1]
Department of Statistics, Mathematics, and Insurance, Faculty of Commerce, Assuit University, Asyut 71515, Egypt[1]
School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, China[2]

***Abstract*—Default in premium payments impacts significantly on the profitability of the insurance company. Therefore, predicting defaults in advance is very important for insurance companies. Predicting in the insurance sector is one of the most beneficial and important study areas in today's world, thanks to technological advancements. But because of the imbalanced datasets in this industry, predicting insurance premium defaulting becomes a difficult task. Moreover, there is no study that applies and compares different SMOTE family approaches to address the issue of imbalanced data. So, this study aims to compare different SMOTE family approaches. Such as Synthetic Minority Oversampling Technique (MOTE), Safe-level SMOTE (SLS), Relocating Safe-level SMOTE (RSLS), Density-based SMOTE (DBSMOTE), Borderline-SMOTE(BLSMOTE), Adaptive Synthetic Sampling (ADSYN), and Adaptive Neighbor Synthetic (ASN), SMOTE-Tomek, and SMOTE-ENN, to solve the problem of unbalanced data. This study applied a variety of machine learning (ML)classifiers to assess the performance of the SMOTE family in addressing the imbalanced problem. These classifiers including Logistic Regression (LR), CART, C4.5, C5.0, Support Vector Machine (SVM), Random Forest (RF), Bagged CART(BC), AdaBoost (ADA), Stochastic Gradient Boosting, (SGB), XGBOOST(XGB), NAÏVE BAYES, (NB), k-Nearest Neighbors (K-NN), and Neural Networks (NN). Additionally, model validation strategies include Random hold-out. The findings obtained using various assessment measures show that ML algorithms do not perform well with imbalanced data, indicating that the problem of imbalanced data must be addressed. On the other hand, using balanced datasets created by SMOTE family techniques improves the performance of classifiers. Moreover, the Friedman test, a statistical significance test, further confirms that the hybrid SMOTE family methods are better than others, especially the SMOTE -TOMEK, which performs better than other resampling approaches. Moreover, among ML algorithms, the SVM model has produced the best results with the SMOTE- TOMEK.**

***Keywords*—Machine learning; classification; insurance; imbalanced data; SMOTE family; statistical analysis**

I. Introduction

In the era of the industrial revolution, all businesses seek digital transformation. One of the key elements of digital transformation is your ability to manage data. Data Science and business analytics is the tool that is being employed on the holy grail of data to extract hidden insights. Since the amount of data is exponentially increasing, therefore the systematic process of data science is gaining popularity in recent times. Like any other industry, 'THE INSURANCE' industry is no exception, and in fact, it is one of the key areas where data science is being practiced at a large scale. Many insurance companies are now employing ML techniques that provide a more systematic way of obtaining a more accurate and representative outcome than the traditional statistic approach.

One of the main challenges with ML approaches in classification is that they are influenced by the data set's unequal class distribution. In other words, when the data is uneven, many ML algorithms may simply disregard the tiny class and assign the majority of the cases to the common class, resulting in high overall model accuracy. Still, the prediction models' efficiency for the tiny class will be drastically diminished. Thus, this study aims to apply a variety of SMOTE family techniques to deal with the imbalanced data problem to improve the performance of ML models in predicting the small class efficiently. In our study, we will develop 117 ML models for predicting insurance premium defaulting {(9 of SMOTE family methods) × (13 of ML models) = 117 model}.

The following is the structure of this paper: Section II presents the previous studies. Section III explains the methodology included data collection, Data Preparation, and imbalanced data problem. Section IV explains model training and parameter optimization. Section V presents the evaluation methods. Section VI shows the results. Section VII shows the results of the statistical tests. Section VIII and IX represent the conclusion and the future work, respectively.

II. Related Work

In the study of [1], they employed several data level methodologies to try to address the unbalanced data issue to predict the occurrence of claims in insurance. The AdaBoost model with oversampling and the hybrid technique produced the highest accurate results. And [2]; they used big insurance data to build eight ML algorithms to predict the occurrence of claims, and they handled the highly imbalanced data using the over-sampler technique. The random forest classifier outperformed the other algorithms. Furthermore, [3] constructed a model for forecasting insurance claims; they generated four classifiers to predict the claims, with the XGBoost model outperforming the others. And [4] predicted

the frequency of vehicle insurance claims using two competing approaches, logistic regression and XGBoost. According to this study, the XGBoost model outperforms logistic regression. Further, the [5] study is to investigate data mining approaches for developing a predictive classifier for vehicle insurance claim prediction. Their studies revealed that neural networks were the best predictor. And [6], this study intends to provide an accurate way for insurance companies to forecast whether or not the customer relationship with the insurance company will be renewed or not. In this paper, random forests were shown to be the top-performing algorithm. And [7], this study starts with data enrichment and works its way up to model development to predict customer churn. And they applied class weights to the prediction model due to the imbalance of the samples. And in [8] the aim of this paper is to compare and contrast the results of different machine-learning techniques for churn prediction; according to the results of this study, the Random Forest and ADA improve outperform all other methods. The study of [9] shows that after using resampling techniques to solve the imbalanced data problem, the efficiency of all ML classifiers in predicting auto insurance fraud is enhanced. Besides, the Stochastic gradient boosting classifier obtained the best result after using the SMOTE-ENN resampling technique among all the other models. And [10] created a new approach for improving the accuracy of fraud prediction. And to solve the unbalanced data problem, they re-balance the data through the method "Resample" of Weka before applying testing and learning. According to this study, Random Forest outperforms all other algorithms in terms of fraud prediction. And [11] predicts fraudulent claims and estimates insurance premium amounts for a range of customers depending on their personal and financial data. The results showed that the Random Forest outperforms the other two algorithms on the Insurance claim dataset. And to deal with the unbalanced data distribution, the research of [12] provides a novel insurance fraud detection technique. The paper is based on constructing insurance fraud detection models based on data partitions derived from under-sampling. The results show that DT outperforms other algorithms.

To accentuate the importance of our study and the gap that we will fill in this study, we summarized a list of recent research that works on classification in the insurance industry by applying the ML models is presented in Table I.

TABLE I. REVIEW OF RESEARCH WORKS IN THE FIELD OF CLASSIFICATION IN THE INSURANCE INDUSTRY

The study	ML models											SMOTE FAMILY									Statistical analysis
	LR	Decision tree (CART or C4.5, or C5.0)	SVM	RF	Bagged CART	ADA	SGB	XGBOOST	NB	k-NN	NN	ADASYN	ANS	BLSMOTE	DBSMOTE	RSLS	SLS	SMOTE	SMOTE-ENN	SMOTE-Tomek	
[1]		CART, C5.0, C4.5		√	√	√	√	√										√			
[2]	√	CART, C5.0, C4.5		√				√	√	√											
[3]		C4.5						√	√		√										
[4]	√							√													
[5]	√	C4.5									√										
[6]	√		√	√		√				√	√										
[7]				√																	
[8]	√	CART	√	√		√	√		√	√	√										
[9]	√	CART, C5.0, C4.5	√	√		√	√	√	√	√	√							√	√		
[10]	√	C4.5	√	√		√			√		√										
[11]		C4.5		√					√												
[12]		√	√								√										
Present study	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√

Table I demonstrates that there is an absence of application and detailed comparison of the common SMOTE family approaches for handling unbalanced data in the insurance industry. This research aims to look into the impact of SMOTE family techniques on boosting the performance of machine learning models in the insurance industry. So, in this study, we applied numerous SMOTE family approaches for solving the imbalanced data problem to fill in the gaps in the previous studies. As compared to earlier studies, the following are our study's original advances and key procedures:

- Using feature scaling to standardize different data features.
- Implementing and comparing different SMOTE family techniques, including nine different methods.
- Hold-Out is applied as a prominent cross-validation algorithm to perform the validation process.
- Comparison of the efficiency of SMOTE family techniques using different ML algorithms, including 13 different models.
- Using various evaluation approaches, such as Accuracy, sensitivity, specificity, and AUC, to assess the performance of the developed models.
- Showing how the various SMOTE family strategies affect the performance of classifiers.
- Using the Friedman test to analyze the differences among several SMOTE family approaches and indicating the best method among the others.

III. Methodology

This study compares various SMOTE family approaches to handle the imbalanced data problem to discover the optimal methodology and classifier for forecasting insurance premium defaulting. The following are the methodology steps used to attain the objectives of this paper:

- Data Gathering.
- Data Preparation.
- Implementing SMOTE family techniques to solve the issue of the Imbalanced data.
- applying ML classification algorithms.
- Analyzing the outcomes.

Fig. 1 shows the Flow chart of the proposed work in our study.

A. Data Collection

This research has used datasets from an insurance company of Egypt. between 2014 and 2020 years. This data collection has a number of variables that can influence insurance premium defaulting. This dataset includes information on the 93520 clients with ten various features. There are four categorical variables (area type, Accommodation, Marital status, Default or Not), and six continuous &discrete variables (Age Income, Number of Vehicles owned, number of Late payments, number of premiums that paid, Premium amount, the number of dependents for the insured) with no missing values and columns.

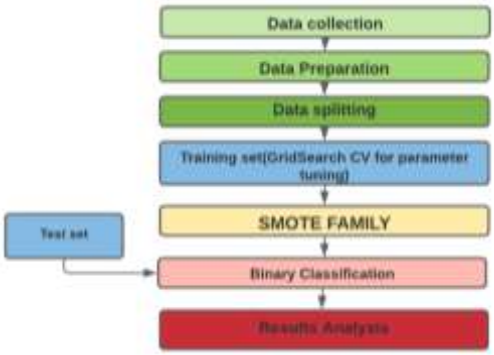


Fig. 1. Working Diagram of Proposed Model.

B. Data Preparation

One of the most crucial stages in ML is data preparation. This procedure turns raw data into an understandable format. This phase will eliminate the errors, which may exist in the dataset, making datasets easier to manage [2]. And the data preprocessing can be summarized into the following two steps.

1) Feature scaling: Feature scaling is a method of normalizing the range of independent variables in a dataset. Most ML algorithms employ the Euclidean distance between two data points, hence without Feature Scaling, the ML algorithms may not perform properly [13]. And in this study, the values range in our dataset is not similar for most variables, so we will apply the Standardization technique as a feature scaling method to rescale the data variables. As a consequence, all of the variables become to have a mean of zero and a standard deviation of one, which is typical of a normal distribution.

The data were scaled using the following algorithm:

$$\boldsymbol{Z} = \frac{\boldsymbol{X} - \mu}{\boldsymbol{\sigma}}$$

Where μ is the mean and $\boldsymbol{\sigma}$ is the standard deviation.

2) One-hot encoding for categorical features: In machine learning, one-hot encoding is the process of converting categorical data into a format that can be fed into ML algorithms. Because most of the ML models works only with the numerical inputs.

C. Imbalanced Data Problem

It's worth noting that most ML algorithms in classification operate best when each class's number of instances is roughly equal. Because the unbalanced data lead to the majority class dominates the minority class. Consequently, algorithms are biased toward the majority class, and their performance become is unreliable [1,14,15]. Our datasets are severely uneven, and the two categories of insurance premium defaulting are not equivalent; in reality, the dataset contains more samples from non-defaulted (90% of the observations)

and defaulted classes (only 10 % of observations). Several techniques have been proposed to address the issue of imbalanced data, SMOTE family is one of the highly effective strategies for resolving the issue of imbalanced data.

SMOTE family: Is a collection of numerous oversampling techniques evolved from SMOTE.

1) Synthetic Minority Oversampling Technique (SMOTE): SMOTE is a statistical strategy that generates new instances to increase the number of minority samples in the dataset. This approach takes feature space samples for each target class and its nearest neighbours, then generates new samples that blend the features from the target case with the features from its neighbours. The new cases are not exact replicas of extant minority cases [16].

2) Adaptive Synthetic Sampling (ADASYN): ADASYN's core concept is to apply a weighted distribution for different minority class instances according to the possibility of learning them. With more artificial instances generated for the minority class instances that are harder to learn than minority class instances that are simpler to learn. Consequently, this technique enhances data distribution learning by eliminating or decreasing the bias brought on by data imbalanced and adaptively pushing the classification decision boundary toward difficult instances [17].

3) Borderline-SMOTE (BLSMOTE): BLSMOTE is a new minority over-sampling technique founded on the SMOTE method that over-samples only the minority examples at the borderline, where the number of majority neighbours of each minority instance is used to split minority instances into three groups: SAFE/DANGER/NOISE. Only the DANGER is employed to generate synthetic instances [18].

4) Density-based SMOTE (DBSMOTE): DBSMOTE, a new over-sampling approach. This method is based on a density-based clustering concept and is intended to oversample a randomly shaped cluster obtained by DBSCAN. DBSMOTE creates synthetic instances by finding the shortest path between each positive instance and a minority-class cluster's pseudo centroid. As a result, the synthetic dataset that results are dense around the core of a group of original positive cases [19].

5) Adaptive Neighbor Synthetic (ANS): The requirement of the number of nearest neighbours as a critical parameter to synthesize instances is one of SMOTE's drawbacks. And The Adaptive Neighbor Synthetic Minority Oversampling Technique (ANS) is a new adaptive technique that tries to avoid this drawback by dynamically adapts the number of neighbours required for oversampling around different minority regions [20].

6) Safe-level SMOTE (SLS): SMOTE synthesizes minority instances at random along a line connecting a minority instance, and it's chosen nearest neighbours while disregarding surrounding majority instances. SLS is a technique that meticulously samples minority instances along the same line with varied weight degrees, which is referred to as the safe level. The safe level is calculated using the minority instances of the nearest neighbours [21].

7) Relocating Safe-level SMOTE (RSLS): SLS creates synthetic minority instances in the vicinity of original instances while avoiding majority instances nearby. This may cause some classifiers to become confused. Furthermore, SLS generates synthetic instances without employing minority outcast instances; thus, some valuable information of the minority class may be lost in the dataset. And by merging two methods, the RSLS tries to address these two flaws in SLS. The first is to check and move these synthetic instances away from any potentially nearby majority instances. The second is using the 1-nearest neighbour strategy to deal with minority outcasts [22].

8) HYBRID techniques: smote family that are considered as over-sampling methods have their own set of benefits and drawbacks. Combining the Over-sampling methods with the under-sampling can help reap the benefits of both.

a) SMOTE-ENN: The SMOTE-ENN technique is one of the most well-known techniques for improving outcomes by combining the SMOTE that represent an over-sampling technique with the Edited Nearest Neighbors (ENN) that represent an under-sampling technique [23].

b) SMOTE-Tomek: The SMOTE-Tomek technique combines the SMOTE that represents an over-sampling technique with the Tomek that represents an under-sampling technique to improve outcomes [23].

IV. Model Training with Parameter Optimization

A. Model Validation

By using the cross-validation technique, the data were divided into training and testing subsets. Cross-validation of input data is used to prevent machine learning models from overfitting and underfitting. This study used the Random holdout as a popular cross-validation procedure.

You can see a scheme of holdout CV in Fig. 2

Fig. 2. Holdout CV.

- The data is randomly split into a training and test set.
- A model is trained using only the training set.
- Predictions are made on the test set.
- The predictions are compared to the true values.

B. Overfitting and Underfitting

Machine model's training and validation scores will be recorded at lower levels in the case of Underfitting. In comparison, overfitting is defined as a pattern of high training scores combined with low validation results. Model parameters must be optimized to avoid overfitting and underfitting circumstances. The grid search technique, which is a popular tuning tool, was used to optimize the parameters of the models. Table II shows the best values for model parameters.

TABLE II. MACHINE LEARNING MODELS WITH THEIR SPECIFIC PARAMETER'S SETTINGS

K-NN	K=23	SVM	C =0.8
CART	cp = 0.006329114.	RF	mtry = 2
C4.5	C = 0.01. M = 5.	NN	size = 1. decay = 0.1
LR	no tuning parameters.	ADA	nIter = 150. method = Real adaboost.
NB	laplace = 0. usekernel = TRUE. adjust = 0.4.	C5.0	trials = 90. model = tree. winnow = FALSE
XGB	nrounds = 50. max_depth = 2. eta = 0.3. gamma = 0. colsample_bytree = 0.8. min_child_weight = 1. subsample = 1.	SGB	n. trees = 50. interaction. depth = 1. shrinkage = 0.1. n. minobsinnode = 10.
BC	no tuning parameters.		

V. EVALUATION METHODS

Methods of evaluation are critical in comparing and selecting the best model [1].

TABLE III. EVALUATION METHODS

Accuracy	Referred to the overall correctly prediction	$\frac{(TP + TN)}{(TP + FP + TN + FN)}$
Sensitivity	Referred to the correct rate of predicting the default class.	$\frac{TP}{(TP + FN)}$
Specificity	Referred to the correct rate of predicting the non-default class	$\frac{TN}{(FP + TN).}$

The evaluation methods employed in this study are shown in Table III. Where TP is the number of true positives, FP represents the number of false positives, TN represents the number of true negatives, and FN represents the number of false negatives.

Where:

1) TP: is the aggregate number of clients who accurately attributed to default class.

2) FP: is the aggregate number of clients who inaccurately attributed to the default class.

3) TN: is the aggregate number of clients who accurately attributed to non-default class.

4) FN: is the aggregate number of clients who inaccurately attributed to the non-default class.

Besides the evaluation methods in Table III, we also used the AUC, AUC is a universal quality metric for models. AUC of 1 indicates a perfect model, whereas an AUC of 0.5 indicates a random model.

Analyzing and comparing the performance of the classifiers is an important procedure. Although evaluation measures are straightforward to employ, the results obtaining from the evaluation measures may be misleading. As a result, determining the optimal model or technique according to their abilities is a difficult task. This problem will be solved using statistical significance tests [24]. A common statistical test method for determining the differences between two or more related sample means is called the ANOVA test. The ANOVA's null hypothesis is that all resampling procedures are equivalent, and the stated discrepancies are just coincidental [25]. There are three assumptions that must take into account before we applied the ANOVA test.

1) All samples must follow the normal distribution.

2) The sample cases should be independent of one another.

3) There should be roughly equal variance among the methods (SMOTE family methods).

The Anderson–Darling normality test [25] is used in this study to determine whether data is normal or not. The null hypothesis of this Anderson–Darling normality test is that the data follow a normal distribution. And we will accept this null hypothesis if the p-value of the test is more than 0.05; otherwise, we will reject the null hypothesis if the p-value $\leq$ 0.05.

If one of the ANOVA's assumptions be broken, the Friedman test [26] will be used instead of the ANOVA test to investigate differences among the methods. The Friedman test's null hypothesis is that all SMOTE family methods perform the same. And we will accept the null hypothesis if the p-value of the test is more than 0.05; otherwise, we will reject the null hypothesis if the p-value $\leq$ 0.05. And rejecting the null hypothesis means that at least one of the SMOTE family strategies perform differently from others. For each SMOTE family approach, the accuracy, sensitivity, specificity, and AUC values are used to compare the ability of the different resampling techniques to tackle the problem of unbalanced data.

The Freidman test ranks each classifier's data for each SMOTE family technique, then examines the ranks values [27].

As a result, for each SMOTE family technique, the Friedman test generates a sum of ranks, which aids in determining which SMOTE family method is the most effective among the others.

VI. RESULTS

The performance of the various ML classifiers on the unbalanced dataset and also on the balanced data that was generated by the SMOTE family methods is shown in Table IV. Various assessment measure methods, including accuracy, sensitivity, specificity, and AUC, are utilized to gain a better knowledge of the models' performance.

Table IV shows the accuracy, sensitivity, specificity, and AUC of each ML strategy on balanced and imbalanced datasets created by the SMOTE family. The most important outcomes are from Table IV; there is a substantial discrepancy between specificity and sensitivity with the unbalanced data.

TABLE IV. PERFORMANCE OF THE CLASSIFIERS

ML	Evaluation	unbalanced	ADASYN	ANS	BLSMOTE	DBSMOTE	RSLS	SLS	SMOTE	SOMTE – TOMEK	SMOTE - ENN
K-NN	Accuracy	0.9105	0.8023	0.8092	0.8583	0.8363	0.8737	0.8725	0.8035	0.6654	0.7283
	sensitivity	0.11911	0.5198	0.5128	0.4013	0.4118	0.3247	0.3142	0.5024	0.7408	0.6607
	specificity	0.96489	0.8243	0.832	0.8914	0.8674	0.9127	0.912	0.8267	0.6625	0.7345
	AUC	0.542	0.67205	0.6724	0.64635	0.6396	0.6187	0.6131	0.66455	0.70165	0.6976
LR	Accuracy	0.9062	0.7791	0.781	0.8074	0.7849	0.7849	0.8787	0.7849	0.7072	0.7767
	sensitivity	0.1727	0.7568	0.769	0.7201	0.7201	0.7201	0.4754	0.7201	0.8349	0.7687
	specificity	0.9638	0.7852	0.7863	0.8182	0.7941	0.7941	0.9124	0.7941	0.701	0.7792
	AUC	0.56825	0.771	0.77765	0.76915	0.7571	0.7571	0.6939	0.7571	0.76795	0.77395
SVM	Accuracy	0.9025	0.776	0.7676	0.7981	0.7733	0.8913	0.8899	0.7684	0.7082	0.7707
	sensitivity	0.032	0.7659	0.7751	0.7262	0.7323	0.4968	0.4418	0.7843	0.8384	0.7722
	specificity	0.97	0.7813	0.7715	0.8078	0.7808	0.9288	0.9268	0.7018	0.7717	0.7725
	AUC	0.501	0.7736	0.7733	0.767	0.75655	0.7128	0.6843	0.74305	0.80505	0.77235
NB	Accuracy	0.9018	0.8539	0.8591	0.856	0.8213	0.8659	0.8667	0.8566	0.8741	0.8826
	sensitivity	0.032	0.5366	0.5305	0.5611	0.5886	0.4357	0.4968	0.5488	0.5213	0.4969
	specificity	0.9693	0.8814	0.8874	0.8818	0.8426	0.8971	0.898	0.8834	0.8985	0.9091
	AUC	0.50065	0.709	0.70895	0.72145	0.7156	0.6664	0.6974	0.7161	0.7099	0.703
C5.0	Accuracy	0.9028	0.898	0.8974	0.8986	0.8955	0.8994	0.8967	0.9003	0.7254	0.8074
	sensitivity	0.2031	0.2031	0.2237	0.2237	0.1913	0.2326	0.2355	0.2178	0.7478	0.6711
	specificity	0.9606	0.9554	0.9532	0.9545	0.9537	0.9548	0.9516	0.9568	0.7258	0.818
	AUC	0.58185	0.57925	0.58845	0.5891	0.5725	0.5937	0.59355	0.5873	0.7368	0.74455
C4.5	Accuracy	0.8988	0.8926	0.8919	0.8907	0.9165	0.9233	0.9247	0.9168	0.7437	0.8074
	sensitivity	0.1264	0.2267	0.2208	0.2031	0.3653	0.5589	0.5425	0.387	0.7269	0.7025
	specificity	0.9622	0.9478	0.9476	0.9476	0.9583	0.9524	0.955	0.9572	0.7467	0.816
	AUC	0.5443	0.58725	0.5842	0.57535	0.6618	0.75565	0.74875	0.6721	0.7368	0.75925
CART	Accuracy	0.898	0.8782	0.8834	0.8811	0.8851	0.8824	0.8795	0.8786	0.7688	0.8271
	sensitivity	0.09395	0.3978	0.3919	0.4243	0.386	0.4007	0.4361	0.4361	0.7617	0.6572
	specificity	0.96373	0.9194	0.9254	0.9205	0.9277	0.9236	0.9178	0.9169	0.7712	0.8398
	AUC	0.52884	0.6586	0.65865	0.6724	0.65685	0.66215	0.67695	0.6765	0.76645	0.7485
BC	Accuracy	0.9057	0.9036	0.9057	0.9061	0.9048	0.9025	0.903	0.904	0.7429	0.7911
	sensitivity	0.1783	0.1992	0.2445	0.2236	0.2062	0.2167	0.2202	0.2202	0.7269	0.6676
	specificity	0.956	0.9524	0.9518	0.9536	0.9533	0.9502	0.9504	0.9516	0.7458	0.8009
	AUC	0.56715	0.5758	0.59815	0.5886	0.57975	0.58345	0.5853	0.5859	0.73635	0.73425
XGB	Accuracy	0.9103	0.9088	0.9107	0.9076	0.9092	0.8925	0.8956	0.9101	0.7527	0.8158
	sensitivity	0.1574	0.147	0.1783	0.1679	0.1749	0.2898	0.2898	0.1714	0.7164	0.6363
	specificity	0.9622	0.9613	0.9613	0.9587	0.96	0.9349	0.9382	0.9611	0.7569	0.8291
	AUC	0.5598	0.55415	0.5698	0.5633	0.56745	0.61235	0.614	0.56625	0.73665	0.7327
ADA	Accuracy	0.9105	0.9089	0.90975	0.9063	0.9091	0.89	0.8909	0.90935	0.7458	0.80745
	sensitivity	0.04245	0.17485	0.19575	0.2045	0.19755	0.34035	0.3351	0.1923	0.7443	0.6694
	specificity	0.96978	0.95965	0.9592	0.955	0.95845	0.929	0.93035	0.959	0.7478	0.81815
	AUC	0.506115	0.56725	0.577475	0.57975	0.578	0.634675	0.632725	0.57565	0.74605	0.7437
SGB	Accuracy	0.9117	0.909	0.9088	0.905	0.909	0.8875	0.8862	0.9086	0.7389	0.7991
	sensitivity	0.1505	0.2027	0.2132	0.2411	0.2202	0.3909	0.3804	0.2132	0.7722	0.7025
	specificity	0.9642	0.958	0.9571	0.9513	0.9569	0.9231	0.9225	0.9569	0.7387	0.8072
	AUC	0.55735	0.58035	0.58515	0.5962	0.58855	0.657	0.65145	0.58505	0.75545	0.75485
RF	Accuracy	0.9107	0.9071	0.9078	0.9096	0.9101	0.9096	0.9071	0.9082	0.751	0.8216
	sensitivity	0.0773	0.1992	0.2341	0.1435	0.11911	0.1505	0.1365	0.2132	0.7478	0.6328
	specificity	0.9678	0.9562	0.9547	0.9624	0.96445	0.962	0.9602	0.9564	0.7532	0.8356
	AUC	0.52255	0.5777	0.5944	0.55295	0.54178	0.55625	0.54835	0.5848	0.7505	0.7342
NN	Accuracy	0.9109	0.6809	0.7676	0.7653	0.7488	0.847	0.8029	0.7728	0.7153	0.7404
	sensitivity	0.1714	0.8334	0.7045	0.6871	0.6278	0.5268	0.6278	0.6836	0.8349	0.7966
	specificity	0.962	0.6752	0.7756	0.7743	0.7605	0.8714	0.818	0.7825	0.7096	0.7387
	AUC	0.5667	0.7543	0.74005	0.7307	0.69415	0.6991	0.7229	0.73305	0.77225	0.76765

From Table IV we can see that in the column of imbalanced Dataset, all of the accuracy results are greater than 90%, all sensitivity values are less than 18 % and all specificity results are greater than 96 %, indicating that all classifiers are biased toward the majority class. So, the problem must be addressed because it led to inaccurate results. And, after using various SMOTE family techniques to solve the unbalanced problem, we can see a significant improvement in the ML systems' ability to forecast the minority class. For example, while utilizing imbalanced data, the SVM got a sensitivity of 3.2 %, but the result increased to 83.84% with the SOMTE -TOMEK technique.

VII. Results of Statistical Tests

The ML algorithms perform differently with the different balanced data created by various SMOTE family techniques. As a result, finding the appropriate SMOTE family approach to get the greatest results from ML algorithms is quite difficult. Thus, we will use a Statistical significance test that will help us in this difficult task of deciding on the optimum SMOTE family technique. And before doing the ANOVA test, it's important to check the normality assumption.

TABLE V. The Results of the Anderson-Darling Normality

Accuracy	A = 6.013,	p-value = 7.099e-15
Sensitivity	A = 3.977,	p-value = 5.834e-10
Specificity	A = 6.3676,	p-value = 1.005e-15
AUC	A = 6.013,	p-value = 7.099e-15

Table V shows the normality test results according to the Anderson-Darling normality test on the accuracy, sensitivity, specificity, and AUC. The p-value is less than 0.05; thus, the null hypothesis is rejected, and the ANOVA test cannot be employed.

Because one of ANOVA's assumptions related to the normal distribution is broken, we will use the Friedman test to compare the resampling strategies in both datasets instead of the ANOVA test. The Friedman test results are shown in Table VI.

TABLE VI. The Friedman Test Results

Accuracy	chi-squared= 40.345	df=8	p-value = 2.763e-06
sensitivity	chi-squared = 35.235	df=8	p-value = 2.423e-05
specificity	chi-squared = 43.959	df=8	p-value = 5.793e-07
AUC	chi-squared = 42.201	df= 8	p-value = 1.242e-06

Table VI shows that the p-value of the Friedman test for Accuracy, Sensitivity, Specificity, and AUC is lower than the (0.05). As a result, we will reject the null hypothesis, and the following conclusion can be drawn at least one of the SMOTE family techniques performs differently from the other methods.

Table VII shows the rank, sum of ranks, and median determined from the Friedman test for Accuracy, Sensitivity, Specificity, and AUC. And Table VII confirm the following results:

1) For the accuracy, the RSLS technique could be more effective than the other techniques

2) For the sensitivity, the DBSMOTE technique could be more effective than the other techniques

3) For the Specificity and the AUC, the SMOTE_TOMEK technique could be more effective than the other techniques.

TABLE VII. Additional Information from Friedman Test Results

	RANK	SMOTE FAMILY	SUM OF RANKS	MEDIAN
Accuracy	1	RSLS	82.5	0.89
	2	DBSMOTE	78.5	0.8955
	3	SMOTE	78	0.9003
	4	BLSMOTE	77.5	0.8907
	5	ANS	77	0.8919
	6	SLS	76.5	0.8899
	7	ADASYN	58	0.8926
	8	SMOTE_ENN	36	0.8074
	9	SMOTE_TOMEK	21	0.7429
		Overall		0.8786
Sensitivity	1	SMOTE_TOMEK	112	0.7478
	2	SMOTE_ENN	95	0.6694
	3	ANS	65	0.2445
	4	SMOTE	60	0.387
	5.5	BLSMOTE	54	0.2411
	5.5	RSLS	54	0.3909
	7	SLS	52	0.3804
	8	ADASYN	49	0.2267
	9	DBSMOTE	44	0.3653
		Overall		0.4361
Specificity	1	DBSMOTE	80.5	0.9533
	2	RSLS	79	0.9288
	3	BLSMOTE	77.5	0.9476
	4	SLS	75	0.9268
	5	SMOTE	72.5	0.9516
	6	ANS	71	0.9476
	7	ADASYN	70.5	0.9478
	8	SMOTE_ENN	36	0.816
	9	SMOTE_TOMEK	23	0.7467
		Overall		0.9127
AUC	1	SMOTE_TOMEK	106	0.74605
	2	SMOTE_ENN	99	0.74455
	3	ANS	67	0.59815
	4	BLSMOTE	60	0.5962
	5.5	RSLS	56	0.657
	5.5	SMOTE	56	0.66455
	7	SLS	53	0.65145
	8	ADASYN	49	0.58725
	9	DBSMOTE	39	0.6396
		Overall		0.6724

Fig. 3, 4, 6 and 5 shows the SMOTE family methods' boxplot based on the accuracy, specificity, sensitivity, and AUC, respectively, where data refers to the original data.

To summarize, in this study, we aim to solve the imbalanced problem with SMOTE family methods; the assessment measures as to the accuracy, sensitivity, specificity, and AUC are utilized to compare models more compactly. Accuracy can be a useful measure if data has the same number of samples per class. However, with an imbalanced set of samples, accuracy is not helpful at all because the model predicts the value of the majority classes for all predictions. So, when it comes to selecting the best models, AUC will take precedence. From Fig. 3, 4, 5 and 6, we can note that ML models achieve the highest accuracy and the highest specificity with the original data. On the other hand, ML models achieve the lowest results for the sensitivity and AUC measures; this refers to ML algorithms do not give accurate results using imbalanced datasets, and they cannot predict all the target classes. Therefore, solving the imbalanced data problem is notably necessary. And by using the balanced dataset after applied SMOTE family, the sensitivity and accuracy-test results are not significantly improved. And it is logical because, on the balanced data, most ML classifiers will consider all classes, which will lead to lower sensitivity and accuracy results than the imbalanced data that considers only one class and ignore the other class. Moreover, the specificity and AUC results using the balanced dataset are significantly improved, especially with the hybrid SMOTE methods.

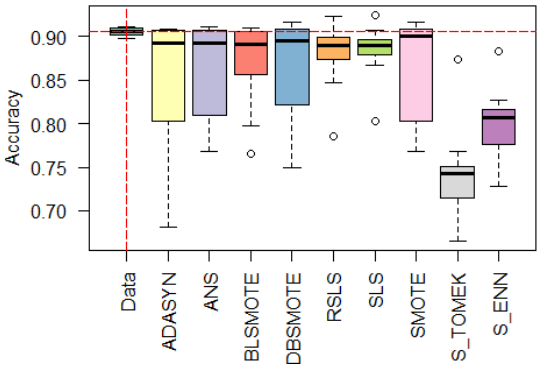


Fig. 3. The Boxplot of the Original Data and SMOTFAMILY based on the Accuracy.

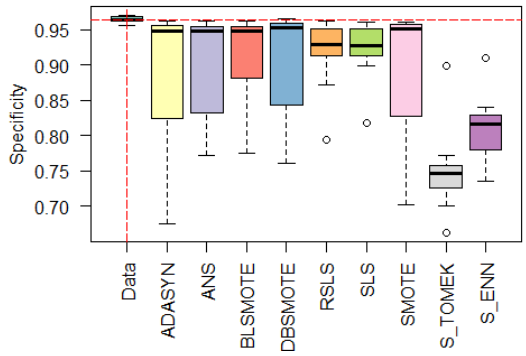


Fig. 4. The Boxplot of the Original Data and SMOTE Family Methods based on the Specificity.

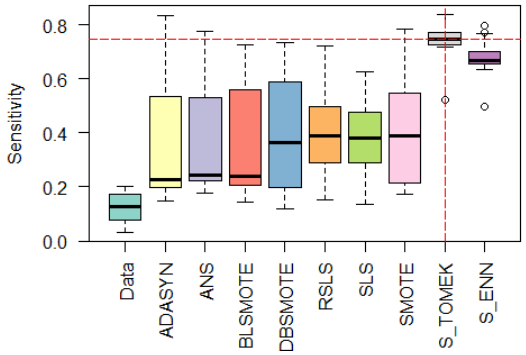


Fig. 5. The Boxplot of the Original Data and SMOTE Family Methods based on the Sensitivity.

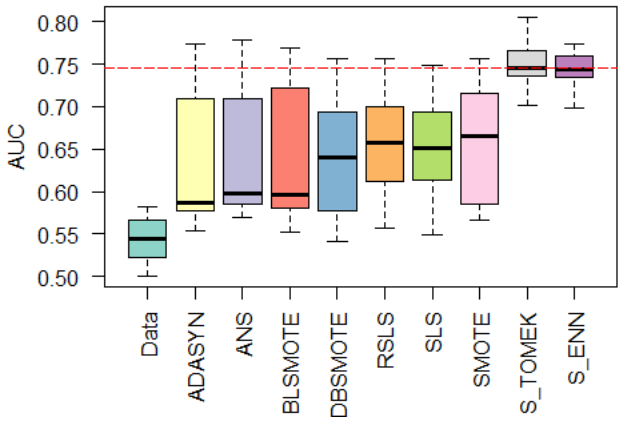


Fig. 6. The Boxplot of the Original Data and SMOTE Family Methods based on the AUC.

Finally, based on the AUC comparison of ML models, the performance of the SVM classifier with the SMOTE-TOMEK method was 80.5%, which was the highest compared with all models.

VIII. Conclusion

The findings show that, algorithms are unable to make accurate predictions with unbalanced data. In contrast, the results demonstrate that algorithms performance has improved when using the various balanced data obtained by different SMOTE family techniques. The findings of the validation approach show that classifiers perform differently on the different balanced data, making it difficult to choose the appropriate resampling technique. The Friedman test was used to determine the optimal resampling approach. According to the AUC, the results of this test show that the hybrid resampling methods are better than others, and especially the SMOTE-TOMEK performs better than alternative resampling approaches. Moreover, among ML algorithms, the SVM model has produced the best results with the SMOTE - TOMEK. According to the results of this paper, we recommend using hybrid resampling strategies to solve the unbalanced data problem as both SMOTE- TOMEK and SMOTE-ENN provided the best performance.

IX. Future Work

The study can be broadened to incorporate hybrid and deep learning algorithms. Other performance indicators might be used to assess performance. The algorithm's timing measures could also be a useful indicator of algorithms performance. Algorithms could also be evaluated with different datasets from various sectors that suffer from the problem of unbalanced data to prove the efficiency of the hybrid resampling strategies to solve the imbalanced data problem.

References

[1] Mohamed Hanafy and Ruixing Ming, "Improving Imbalanced Data Classification in Auto Insurance by the Data Level Approaches" International Journal of Advanced Computer Science and Applications (IJACSA), 12(6), 2021. http://dx.doi.org/10.14569/IJACSA.2021. 0120656.

[2] Hanafy, Mohamed, and Ruixing Ming. "Machine learning approaches for auto insurance big data." *Risks* 9.2 (2021): 42.

[3] Abdelhadi, Shady, Khaled Elbahnasy, and Mohamed Abdelsalam. "A proposed model to predict auto insurance claims using machine learning techniques." *Journal of Theoretical and Applied Information Technology* 98.22 (2020).

[4] Pesantez-Narvaez, Jessica, Montserrat Guillen, and Manuela Alcañiz. "Predicting motor insurance claims using telematics data—XGBoost versus logistic regression." *Risks* 7.2 (2019): 70.

[5] Weerasinghe, K. P. M. L. P., and M. C. Wijegunasekara. "A comparative study of data mining algorithms in the prediction of auto insurance claims." *European International Journal of Science and Technology* 5.1 (2016): 47-54.

[6] Stucki, Oskar. "Predicting the customer churn with machine learning methods: case: private insurance customer data." (2019). Master's dissertation, LUT University, Lappeenranta, Finland.

[7] Mau, Stefan, Irena Pletikosa, and Joël Wagner. "Forecasting the next likely purchase events of insurance customers: A case study on the value of data-rich multichannelenvironments." *International Journal of Bank Marketing* (2018).

[8] Sabbeh, Sahar F. "Machine-learning techniques for customer retention: A comparative study." *International Journal of Advanced Computer Science and Applications* 9.2 (2018).

[9] Hanafy, Mohamed, and Ruixing Ming. "Using Machine Learning Models to Compare Various Resampling Methods in Predicting Insurance Fraud." *Journal of Theoretical and Applied Information Technology* 99.12 (2021).

[10] Itri, Bouzgarne, et al. "Performance comparative study of machine learning algorithms for automobile insurance fraud detection." *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*. IEEE, 2019.

[11] Kowshalya, G., and M. Nandhini. "Predicting fraudulent claims in automobile insurance." *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, 2018.

[12] Hassan, Amira Kamil Ibrahim, and Ajith Abraham. "Modeling insurance fraud detection using imbalanced data classification." *Advances in nature and biologically inspired computing*. Springer, Cham, 2016. 117-127.

[13] S. Aksoy and R. M. Haralick, ''Feature normalization and likelihood-based similarity measures for image retrieval,'' Pattern Recognit. Lett., vol. 22, no. 5, pp. 563–582, Apr. 2001.

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

[15] H. Byeon, Predicting the depression of the South Korean elderly using SMOTE and an imbalanced binary dataset. International Journal of Advanced Computer Science and Applications, vol. 12, no. 1, pp. 74-79, 2021.

[16] Siriseriwan, Wacharasak. "SMOTEfamily: a collection of oversampling techniques for class imbalance problem based on SMOTE (2018)." *URL http://cran. r-project. org/package= SMOTEfamily*.

[17] H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning", *Proc. IEEE Int. Joint Conf. Neural Netw. IEEE World Congr. Comput. Intell.*, pp. 1322-1328, Mar. 2008.

[18] Han, Hui, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." *International conference on intelligent computing*. Springer, Berlin, Heidelberg, 2005.

[19] Bunkhumpornpat, Chumphol, Krung Sinapiromsaran, and Chidchanok Lursinsap. "DBSMOTE: density-based synthetic minority over-sampling technique." *Applied Intelligence* 36.3 (2012): 664-684.

[20] Siriseriwan, Wacharasak, and Krung Sinapiromsaran. "Adaptive neighbor synthetic minority oversampling technique under 1NN outcast handling." *Songklanakarin J. Sci. Technol* 39.5 (2017): 565-576.

[21] Bunkhumpornpat, Chumphol, Krung Sinapiromsaran, and Chidchanok Lursinsap. "Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem." *Pacific-Asia conference on knowledge discovery and data mining*. Springer, Berlin, Heidelberg, 2009.

[22] Siriseriwan, Wacharasak, and Krung Sinapiromsaran. "The effective redistribution for imbalance dataset: relocating safe-level SMOTE with minority outcast handling." *Chiang Mai Journal of Science* 43.1 (2016): 234-246.

[23] Batista, Gustavo EAPA, Ronaldo C. Prati, and Maria Carolina Monard. "A study of the behavior of several methods for balancing machine learning training data." *ACM SIGKDD explorations newsletter* 6.1 (2004): 20-29.

[24] Demšar, Janez. "Statistical comparisons of classifiers over multiple data sets." *The Journal of Machine Learning Research* 7 (2006): 1-30.

[25] R. A. Fisher, Statistical Methods and Scientific Inference. Oxford, U.K.: Hafner Publishing Co, 1956.

[26] Friedman, Milton. "The use of ranks to avoid the assumption of normality implicit in the analysis of variance." *Journal of the american statistical association* 32.200 (1937): 675-701.

[27] Friedman, Milton. "A comparison of alternative tests of significance for the problem of m rankings." *The Annals of Mathematical Statistics* 11.1 (1940): 86-92.

A Gray Box-based Approach to Automatic Requirements Specification for a Robot Patrol System

Soojin Park
Graduate School of Management of Technology
Sogang University, Seoul, Korea

***Abstract*—The black box-based requirements specification models representative by the use case model focus on specifying system behaviors exposed outside. While these models are sufficiently effective in specifying requirements for business applications behavior, they are limited in specifying requirements for embedded systems with relatively very short interaction sequences with users. To solve this problem, we have proposed a gray box-based requirements specification method to specify the inner logic of an embedded system, including a tool for automatic generation of requirements specification from some analysis models in our previous work. This study proves the benefits of the proposed software requirements specification method by applying it to a robot patrol system and showing the possibility of general use of the proposed method in the embedded system domain. Compared with our previous work, we enhance the tool for automatic generation of requirements specification, called *SpecGen*, and prove the benefit of the proposed method from multiple aspects. The application result on the robot patrol system case is quantitatively demonstrating that our proposed requirements specification method improves development productivity and enhances overall software product quality, including code quality.**

Keywords*—Embedded system; automatic requirement specifications generation; mobile robots; use case specification

I. INTRODUCTION

Embedded software refers to software embedded in various electronic products, from small appliances, including mobile phones, digital cameras, and MP3s to robotics systems [1]. Scenario-based specification techniques widely used in business applications are often used in the requirements specification for embedded software. The primary purpose of the scenario-based specification technique represented by the use case model [2] is to describe the interaction between the system and the environment in which the system is used. Business applications are realizing real-world business as services supported by software systems. Hence, most required behaviors of a business application can be captured from the statements for specifying interactions between the user and the system. Although the service provided by the embedded system results from each event generated by the user, it cannot be observable from the outside of the system which internal action the system performs until the service result is derived. In other words, requirements extractable from visible interactions between an embedded system and environmental factors in which it is used are relatively limited [3, 4].

Thus, when the requirements for an embedded system are specified using a use-case model, the amount of information identified in the requirements specification is insufficient as a specification for developers [5]. As is shown in Fig. 1, for an example of a generic flow of events for "Power On" use case of an embedded system is specified as: (1) A user pushes the power-on button to start the system; (2) The system is invoked and waits for the user's other request. Such use case specification is insufficient to be used as a requirements specification to guide the development team designs the embedded system. To overcome this lack of information when the use-case model is applied to the embedded system requirements specification, most existing studies [6-9] pre-populate various design diagrams such as state, sequence, class, or data flow diagrams .etc.

Even if we select a suitable design diagram to specify the inner mechanism of exposed system behavior, we should decide the depth of each design diagram. The deeper the depth of the diagram, the more sophisticated the system's behavior can be included in the requirements specification. We usually face a "what versus how" dilemma [10] in specifying requirements, which means a "how" in the preceding step means again the "what" in a subsequent step. Over-elaborated design diagrams in the requirement stage could violate the definition of a requirement in that it specifies solutions, not problems [11]. Furthermore, it can cause a raising initial system development cost. The requirements model, which includes the requirements set for embedded software developers, should provide the interaction requirements between the system's internal components. As the developers could refine the interaction between components by digging the depth, guidelines for the appropriate elaboration depth are needed to obtain the necessary information in the requirements stage.

Our previous work [12] proposed an extended requirements specification model from use case specification to satisfy these needs. The use case specification guides us to maintain a view of the target system as a single black box [13] when we specify requirements. In contrast, we named the proposed model a "gray-box" based requirements specification in the sense that it is a model based on a perspective that partially looks at the interaction between top-level components among interactions inside the embedded system. It suggests the trade-off between the elaboration depth and the effort to design interactions among internal components of an embedded system when utilized as a requirements specification.

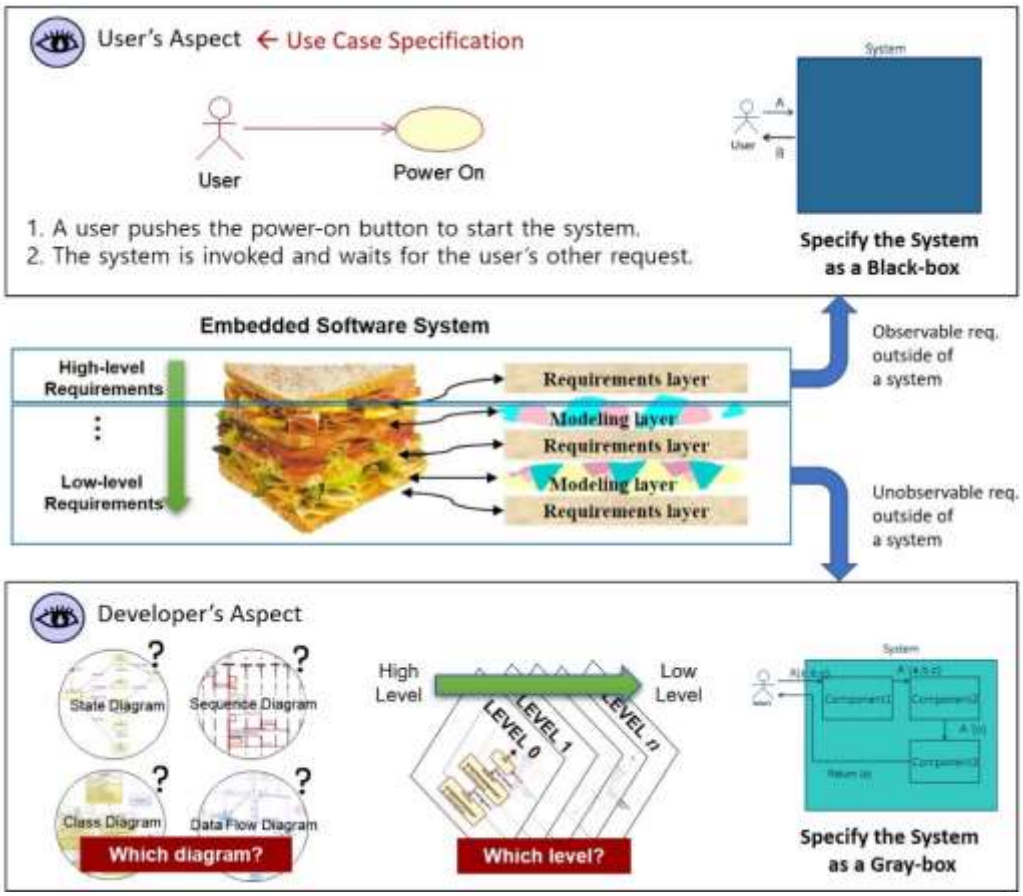


Fig. 1. Different Aspects of a Black Box-based Requirements Specification and a Gray Box-based Requirements Specifications.

The work reported here extends our previous work [12] in the following aspect:

- To show the extensibility of the proposed model through the application case of the more complex embedded system. While the scope of the case study was limited to a module of a mobile phone in [12], in this study, we extend the applicable domain area of the proposed model to the whole of a robot patrol system that is a different domain from the previous work. To prove the extensibility of the applicable area of the model is to prove that the proposed gray box-based requirements specification model can be a general method to specify requirements of embedded systems, not for only a specific system or domain.
- To prove the benefit of the proposed model using more various aspects. The author in [12] explains the benefit by showing that each elapsed time for the software development phases following the requirements phase is decreased when the proposed model is provided to the developers. Besides the enhancement of the productivity of developers, in this study, we show the software product quality and the code quality are also enhanced from using the proposed model through the more sophisticatedly designed experiment.
- To update the automatic generation of the gray box-based requirements specifications, which is renamed SpecGen. We re-developed the tool for utilizing a more prevalently used and better supported UML authoring tool when developers make an analysis model that is the source of the gray box-based requirements specifications. This work could be valuable in helping more developers use the proposed model and the supporting tool.

The rest of the paper is organized as follows: Section 2 investigates the trend of existing studies. Section 3 gives an overview of the gray box-based requirements specification method. Section 4 explains an automatic transition from a design diagram in UML to Microsoft Word typed tabular specifications implemented in *SpecGen*. Section 5 shows each step of requirements specification for a robot patrol system and some fragments of the state diagram and automatically generated specifications by *SpecGen*. Section 6 explains how we designed an experiment for showing the effectiveness using the proposed requirements specification model for the robot patrol system and discusses the results of the accomplished experiment. The conclusion of the paper is discussed in Section 7.

II. Related Work

A. Model-driven Development based Approaches

In embedded software development, model-driven development (MDD) based approaches are now widely used. MDD has the merit that developers can find the software's

essential features, thanks to information on the complicated system structure as an abstracted model [14]. The most typical MDD methods are the COMET method by H. Gamaa [15], which integrates object-oriented and concurrent processing concepts. The OCTOPUS method [16] models the system using a structural, functional, and dynamic model.

The research that addresses the requirements specification problems based on the model created by applying an MDD based approach can be found in [17-19]. Lattemann and Lehmann [17] define controller, actuator, and sensor as three main components that comprise the embedded system and suggest that the controller that controls the entire system should be intensively specified among the three roles. Lavi and Kudish [18] classify the model to be analyzed into the E-level representing the external structure and behavior of the system and the S-level representing the conceptual model of the system inside. They suggest an automatic documentation method for requirements specifications based on activity diagrams and state diagrams for specification and analysis of E-Level processes. Glinz [19] utilizes hierarchical activity diagrams after the relation between system state and objects that comprise the system with a source for the specification of requirements in an embedded system is identified.

Existing works only refer to the necessity that the entire system should be divided into lower systems. Each modeling phase should be recursively applied for the requirement specification of the embedded system. But there is no guideline for stopping the recursion for the elaboration depth of the model to be built. To solve this problem, we have started this study from the work that defines the elaboration depth of the analysis model for requirements specification, which was ignored in the previous studies while preserving their advantages.

B. Requirements Pattern-based Approaches

Another notable approach for requirements specification for embedded systems is requirements pattern-based one. Denger et al. [20] propose a natural language pattern to specify requirements in the embedded systems, including 1) meta models for the description of requirements and 2) meta-models for events and responses that we use to verify the completeness of the pattern language. The proposed patterns seem slightly less common compared to commercial phrase requirements. Matsuo et al. [21] use natural language controlled for requirements, limiting how they can combine simple sentences into more complex sentences. They proposed three different types of frames: noun frame, case frame, and feature frame, and they use the frames to parse requirement specifications, and organize them according to different perspectives, and verify requirement completeness. However, there exists a limit that the frame-based approaches seem to be more difficult for non-specialists to understand and apply. Konrad and Cheng [22] define formal specification pattern systems for embedded systems. These patterns are used to describe system properties mapped to linear time logic. Patterns are classified into qualitative (occurrence or order) and real-time (period, periodic, or real-time) patterns. There is a limit that we should specify the supporting model in a UML 12 variant. Postet al. [23] provide the successful application case of this system to automotive requirements. However, the application coverage is not complete.

A pattern is a set of solutions that are commonly applicable to recurring problems. Therefore, the pattern-based approach has an inherent limitation in the scope of its application. This study is different from the pattern-based approach in that it aims to develop a specification method generally applicable to the embedded systems.

III. Overview of Gray box-based Software Requirements Specification Model

The proposed gray box-based software requirements specification for embedded systems leverages partially cultivated analysis artifacts. However, designing all aspects of an embedded system is not proper in the requirements specification phase, considering that software requirements should focus on what services should be provided in the future. Thus, we have limited the design area to the following two diagrams to which the collaboration behavior between the inner components is to be extracted:

A. Top-level State Diagram of a Controller

A state diagram that shows state changes in the system corresponding to events occurring inside and outside the system is a typical diagram used to design dynamic views of the embedded system. Therefore, we selected it as the diagram to specify the internal behavior of the system corresponding to the event specified in the use case specification. After choosing to use state diagrams as the source of the requirements specification, another remaining issue was identifying which component could represent the state transition of a whole embedded system. A state of an entire system is a specific situation where specific values are assigned to all attributes of components comprising the system. Therefore, the question, which component should have the ownership of the state of a whole system is a controversial issue. Referencing Broy and Stauner [24], we have classified the roles of the main components in an embedded system into controller, actuator, or sensor. Among the three stereotypes of components, we defined the controller coordinating behaviors of other actuators and sensors as the component possessing the states of a whole system.

B. Sequence Diagrams Specifying Interactions of all Top Leveled Components - Controllers, Sensors, and Actuators

The proposed model specifies events exchanged among external subjects in a time sequence. As a flow of events in a use case is reflected in a sequence diagram, each internal interaction invoked by external stimuli from an actor is also designed using a sequence diagram to keep the same context. The owners of the events on the sequence diagram created in this step are actors, controllers, and sensors/actuators (that execute the controller's commands). The states in the top-leveled state diagram of the controller are added as annotations on the lifeline of the controller object in the sequence diagram, as depicted in Fig. 2. In the next step, the sequence diagram added state transitions of a whole embedded system is the source of the automatically generated requirements specification.

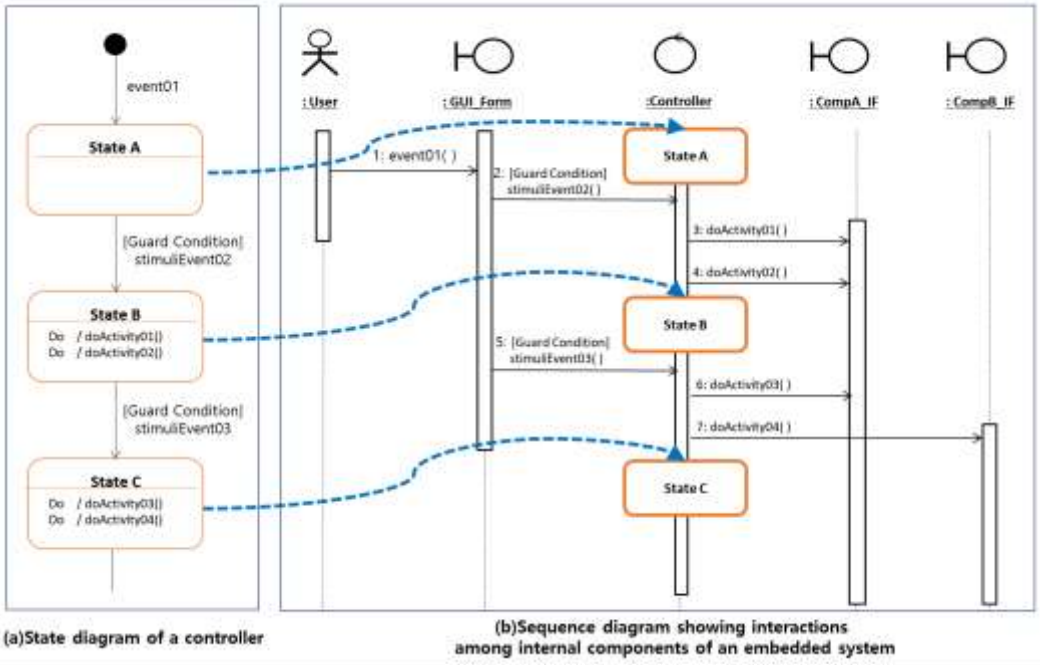


Fig. 2. The Relationship of Information Specified in (a) A State Diagram of a Controller and (b) A Sequence Diagram for showing Interactions.

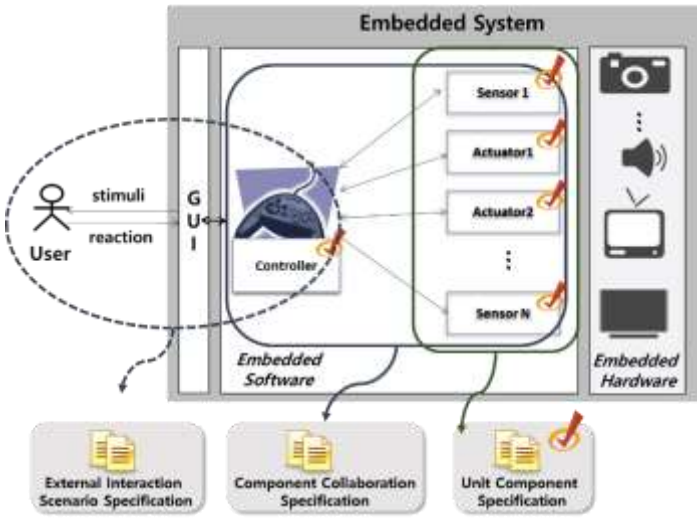


Fig. 3. The Coverage of Three different Requirements Specifications of the Proposed Gray Box-based Requirements Specification Model.

As shown in Fig. 3, once the two kinds of design diagrams are completed by developers, the following three different software requirements specifications can be automatically generated.

- External Interaction Scenario Specification: specifies the interaction between a system and an actor corresponding to the system's external environment. The information included in this specification is equivalent to the information contained in the use case diagram.
- Component Collaboration Specification: specifies the state changes of a controller due to inter-component interaction. The information included in this specification contains state-related information included in the state diagram for the component of the controller that controls the actuator and sensor of the embedded system. In addition, the information recorded in the sequence diagram, which is the result of designing the sequence of commands that the controller receives external stimuli and sends commands to other actuators and sensors, is extracted as this specification.
- Unit Component Specification: specifies the behaviors to be implemented by a specific component. This specification is written by classifying all operation calls in the previously extracted component collaboration specification by corresponding to the receiver and binding them. These unit component specifications are APIs for each class or component in the development stage, in other words.

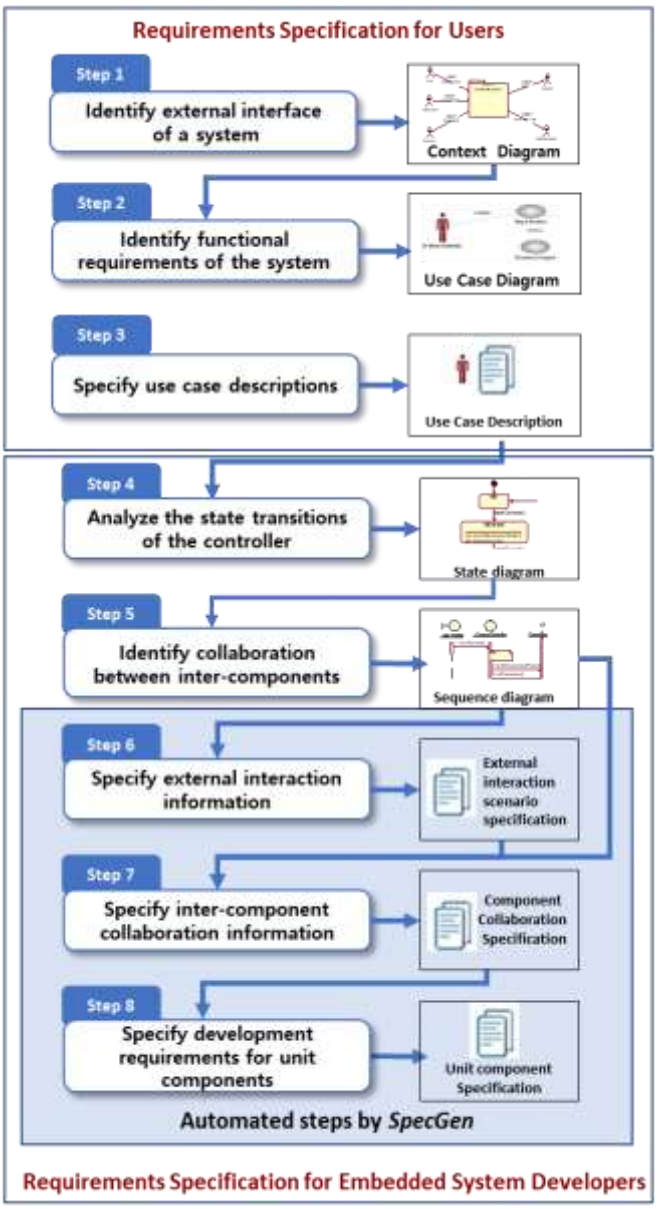


Fig. 4. The Process to Build the Proposed Gray Box-based Software Requirements Model.

Our proposed method does not exclude the steps of the existing black-box requirement specification but includes them. Fig. 4 illustrates each step to construct those specifications. A context diagram and a use case model are specified due to steps 1~3 in Fig. 4. A context diagram represented by UML (Unified Modeling Language) shows which users or interfacing systems are engaged to provide a service in the target system. Specifying the context diagram is not included in the general guidelines of the use case model. However, embedded systems are literally "embedded" in a hardware system. Thus, clear identification of external objects that an embedded system should interface with is critical. We also marked interface systems outside as actors in the context diagram for consistency with the use case model. A use case diagram that specifies the system's services is the same as a typical use case diagram. Steps 4~5 are for developing a design artifact, including a top-level state diagram of a controller and sequence diagrams for specifying collaboration between the controller and other sensors/actuators to respond to each stimulus outside of an embedded system. The following steps 6~8 are to automatically extract the three specifications explained above from the designed diagrams through steps 4~5. We also developed an automatic tool, *SpecGen*, to support these steps. The requirements specifications generated by *SpecGen* define the internal behaviors of an embedded system, which will be utilized as a guideline set for embedded system developers in the following development phases.

IV. SpecGen: A Tool for Automatic Generation of Requirements Specification from Design Diagrams

One of the originalities of our work is to provide a tool for top-leveled design artifacts to be transformed to requirements specifications automatically. This feature is important from three perspectives:

- In writing the requirements specification as a development guideline for developers, the information included in the design diagram created by the developer or designer is linked without loss.
- Since most developers refer to automatically generated requirements specifications and development proceeds, as a result, it does not matter if very few members with design ability use various UML diagrams in the development team.
- And, the support of automated tools can minimize the effort required to write requirements specifications in hand.

Our previous study [12] utilized ArgoUML [25] as the authoring tool for designing diagrams. In this study, we changed the authoring tool to StarUML 4.0 [26] as ArgoUML has not been versioned up. Fig. 5 shows a fragment of the transformation from a UML diagram in StarUML to a requirements specification as a Microsoft Word file by SpecGen. For using SpecGen, the first step is to extract diagrams authored using StarUML to an XMI file. To extract needed information from the XMI file, we should understand the structure of each object in the XMI file representing the UML model extracted from StarUML. Although we can catch the owner object of the first lifeline is "User" intuitively from the given sequence diagram, we can find it out after tracing several lines in the exported XMIs, as depicted in Fig. 5. It depicts the example with the shortest trace selected due to space constraints, but in some cases, the desired information is extracted through a traverse of more than ten lines of XML. Similarly, we analyzed all relevant XMI structures and compared the attributes in each requirements specification we defined. We implemented the transformation rules identified as such with SpecGen.

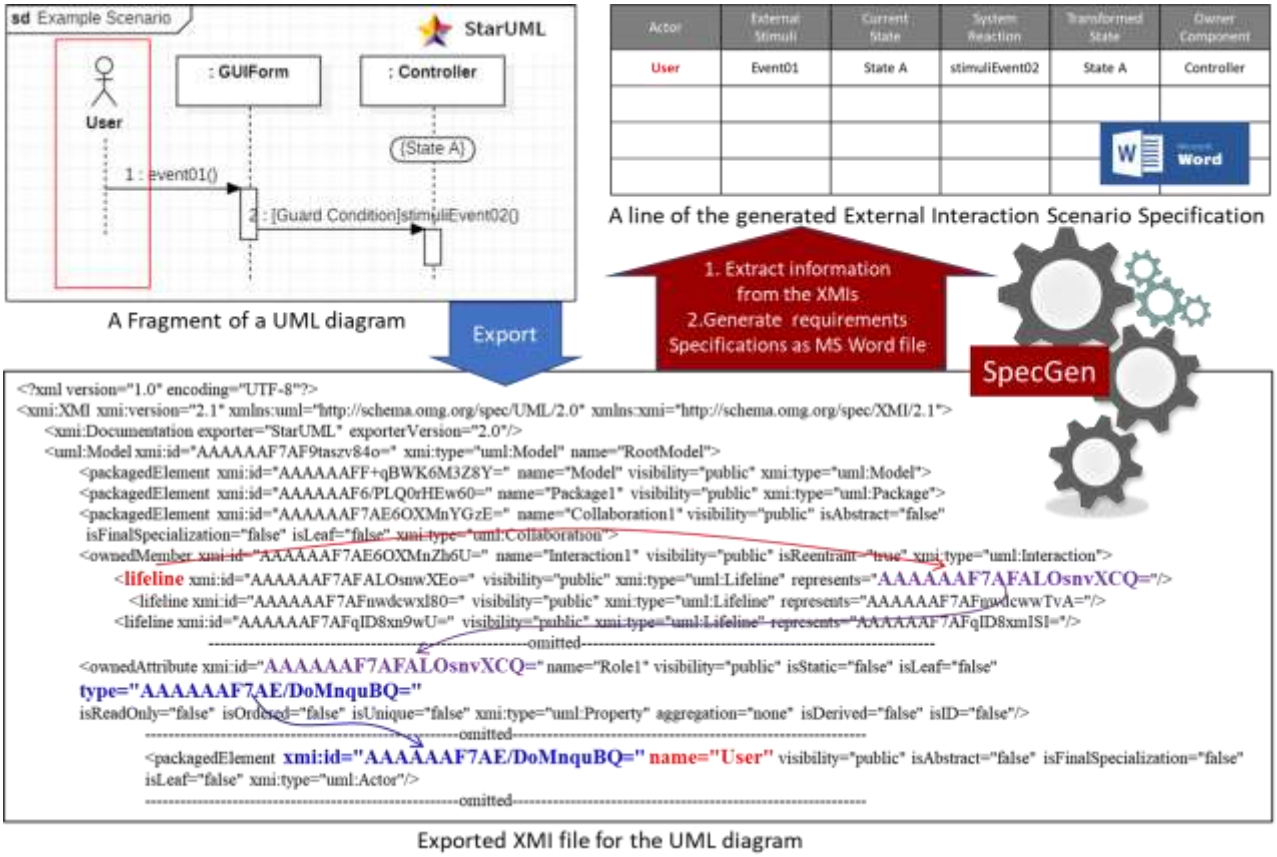


Fig. 5. A Fragment of the Transformation from a UML Diagram to a Requirements Specification by SpecGen.

V. Case Study: Automatic Requirements Specification for a Robot Patrol System

We selected a robot patrol system (RPS) as a target system to which the proposed requirement specification method is applied. RPS is a robot system that provides a service by sending out an alarm when an intruder is detected as it patrols the designated section. The reason to choose a robot system as a target system is that it is one of the typical system domains requiring the three components - controller, sensor, and actuator - in an embedded system as defined by Broy and Stauner [24]. Whereas an operation given for a robot patrol system is simple as "Keep patrolling here," many inner-sided interactions invisible to users are required to patrol within an area. These features are consistent with the feature of the target domain area to which we apply the proposed method.

The followings are the results and explanations of each step in Fig. 4 of the proposed model applied to RPS.

Step1: Identify the external interface of a system

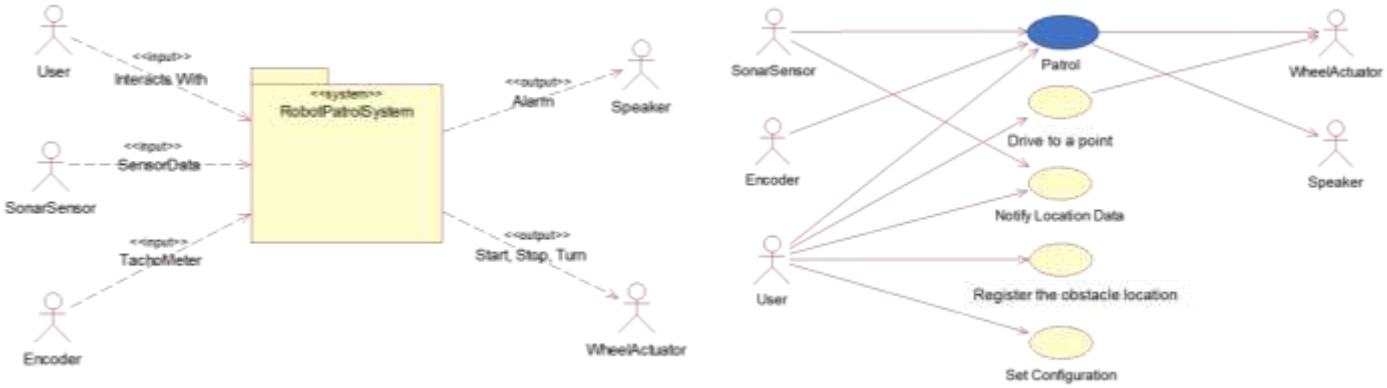
Fig. 6(a) is the context diagram (level 0 data flow diagram) to show the external interface of RPS. To keep the consistency with the following use case diagram, we specify all external entities as actors. The context diagram defines which entities are the sources of data and which entities are the data destinations. In RPS, whereas, *User*, *SonarSensor*, and *Encoder* are the data sources, *Speaker* and *WheelActuator* are the data destinations.

Step2: Identify functional requirements of the system

Fig. 6(b) is the use case diagram, which specifies functional services be provided by the target system. The use cases of RPS are: Patrol, Drive to a point, Notify location data, Register the obstacle location, Set configuration. And, the active actors that invoke a use case are the data source of the context diagram. So, the active actors of RPS are User, SonarSensor, and Encoder. The data sources of the context diagram come to be passive actors being the systems to be interfaced in the use case diagram. In RPS, Speaker and WheelActuator are the passive actors.

Step 3: Specify use case descriptions

Table I shows the use case description of the "Patrol" service. We select a tabular style, and most compartments of the use case specification are specified, including pre-conditions and post-conditions. There are one basic flow and two alternative flows in the "Patrol" use case.



(a) (b)

Fig. 6. Artifacts of Black Box-based Requirements Specification for Robot Patrol System: (a) Context Diagram and (b) Use Case Diagram.

TABLE I. ARTIFACTS OF BLACK BOX-BASED REQUIREMENTS SPECIFICATION FOR ROBOT PATROL SYSTEM: USE CASE SPECIFICATION FOR "PATROL"

Use Case Name	Patrol
Actor	User
Brief Description	The robot patrols the area based on the range of user input.
Basic Flow of Events	This use case begins when the user enters the destination range for patrol by GUI and gives the order to start. [1] The robot patrolling system(RPS) saves the start and destination positions and switches direction to destination positions. And the RPS gives the start command to Wheel Actuator. [2] The RPS reads the sensor values and identifies the intruder. If the intruder is detected, the flow goes to [A1]. [3] The RPS reads the current location and checks whether the RPS arrives at the destination. [3.1] If the robot arrives at the destination, give the stop command to Wheel Actuator, where the use case ends. [3.2] If the robot does not arrive at the destination, the flow goes to [2].
Alternative Flow 1	**[A1] Intruder detection** [1] If the intruder is detected, the RPS gives a stop command to the Wheel Actuator. [2] The RPS causes alarm bells through the speaker, at which point the use case ends.
Alternative Flow 2	**[A2] User's stop command** If the user gives the order to stop at any time, the RPS gives a stop command to the Wheel Actuator, at which point the use case ends.
Exception Paths	N/A
Extension Points	N/A
Pre-conditions	The RPS was initialized state. The robot's starting position is (0, 0). And the direction of the robot is assumed to be 90 degrees.
Post-conditions	The RPS is the stationary state according to user instructions.

With the artifacts depicted in Fig. 6 and 7, we can see what should be developed for the RPS. However, it does not provide sufficient information to guide what should be implemented because it defines only the interactions between actors and the system. If only these artifacts are given to developers, comparatively many decisions should be made by individual developer's capability to realize the specified requirements. If only these artifacts are provided to the developer, the individual developers must make a relatively large number of decisions, which could be a significant burden. The burden to the developers comes from the lack of details in requirements specification will be discussed in Section 6 with experimental results.

Step 4: Analyze the state transitions of the controller

In an embedded system, various sensors and actuators are equipped. However, only the controller has a meaningful state in an embedded system during the system's execution as other sensors or actuators are passive objects that receive commanders from the controller. For this reason, the state diagram of a controller should be created as a diagram explaining the behaviors of a whole embedded system. Fig. 7 shows the top-level state diagram of PatrolSystemController, which controls all other components in RPS. There are five meaningful states while the PatrolSystemController runs: Idle, Initialized, Patrolling, StoppedAtTheDestination, StoppedByIntrution.

Step 5: Identify collaboration between inter-components

After analyzing the state transitions of the controller, the next step is to identify collaboration between components. According to the use case specification in Table I, there are a basic flow and two alternative flows in the "Patrol" use case. Fig. 8 is the sequence diagram for the scenario combining the basic flow and alternative 2 (intruder detection). The one different point comparing with typical sequence diagrams is that the states are additionally annotated on the lifeline of the

controller. We can confirm that the four states- Idle, Initialized, Patrolling, StoppedByIntrution – which are related to the scenario, are annotated on the lifeline of the PatrollSystemController in Fig. 8. The collaboration in Fig. 9 shows that PatrollSystemController controls the sequence of messages to WheelActuatorIF, SonarSensorIF, and SpeakerIF, identified as actors as external modules to interface in the use case diagram. DirectionCalculator and IntrusionDetectionIF newly identified in designing the sequence diagram are also identified as the objects collaborating to provide the "Patrol" service.

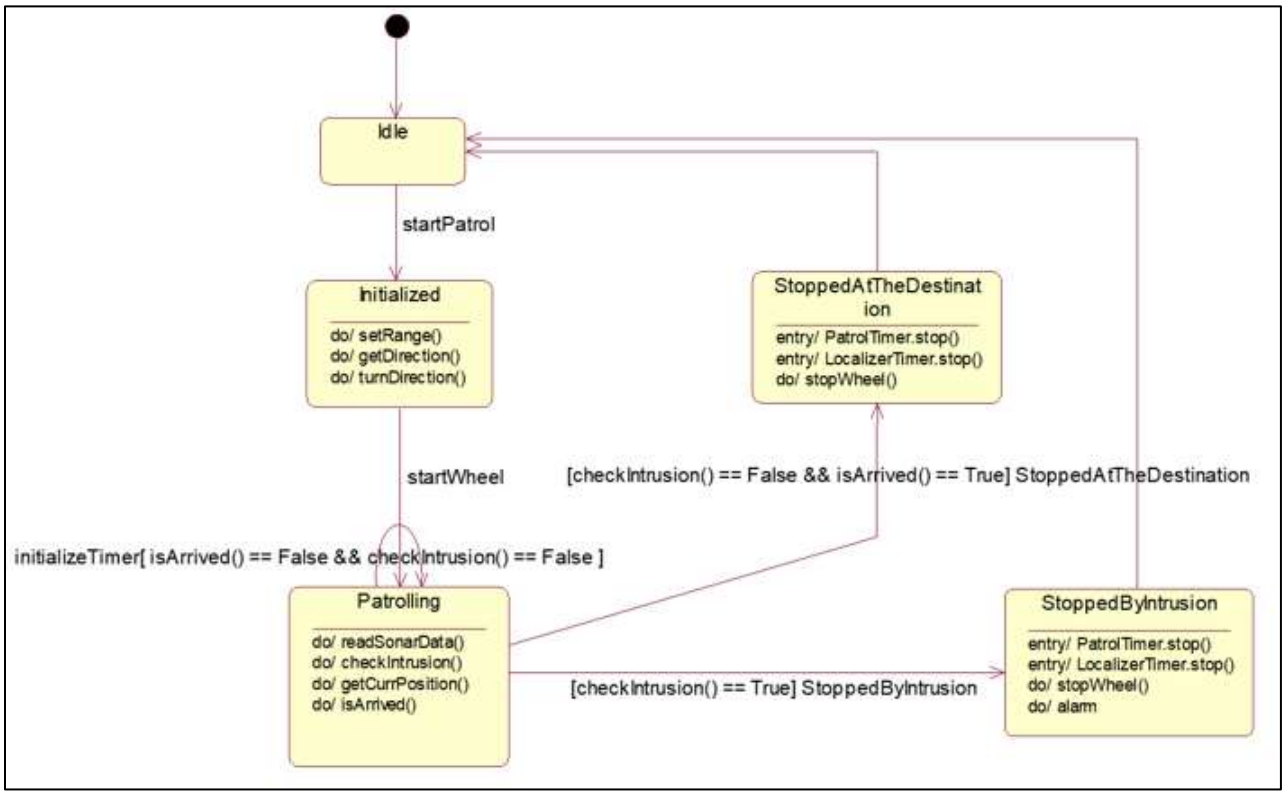


Fig. 7. The State Diagram for Patrol System Controller.

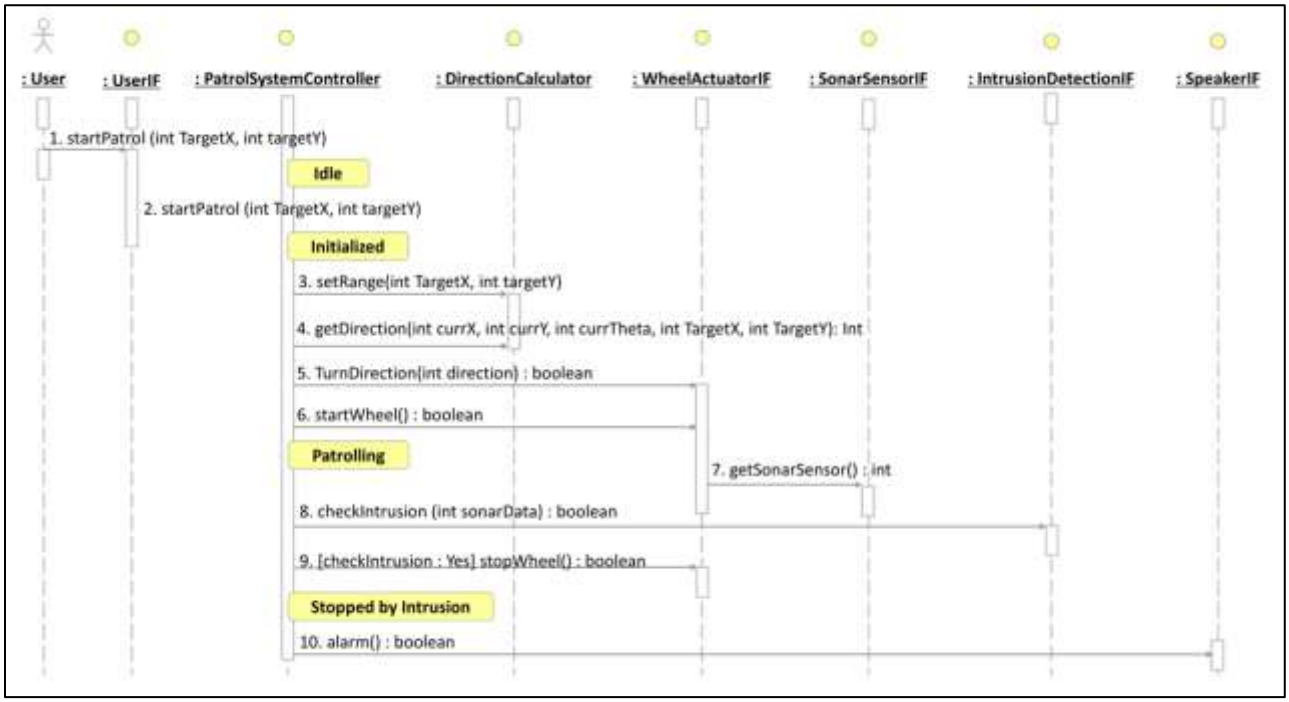


Fig. 8. The Sequence Diagram for the Scenario of the Composition of the Basic Flow and the Alternative Flow 2 in the "Patrol" use Case.

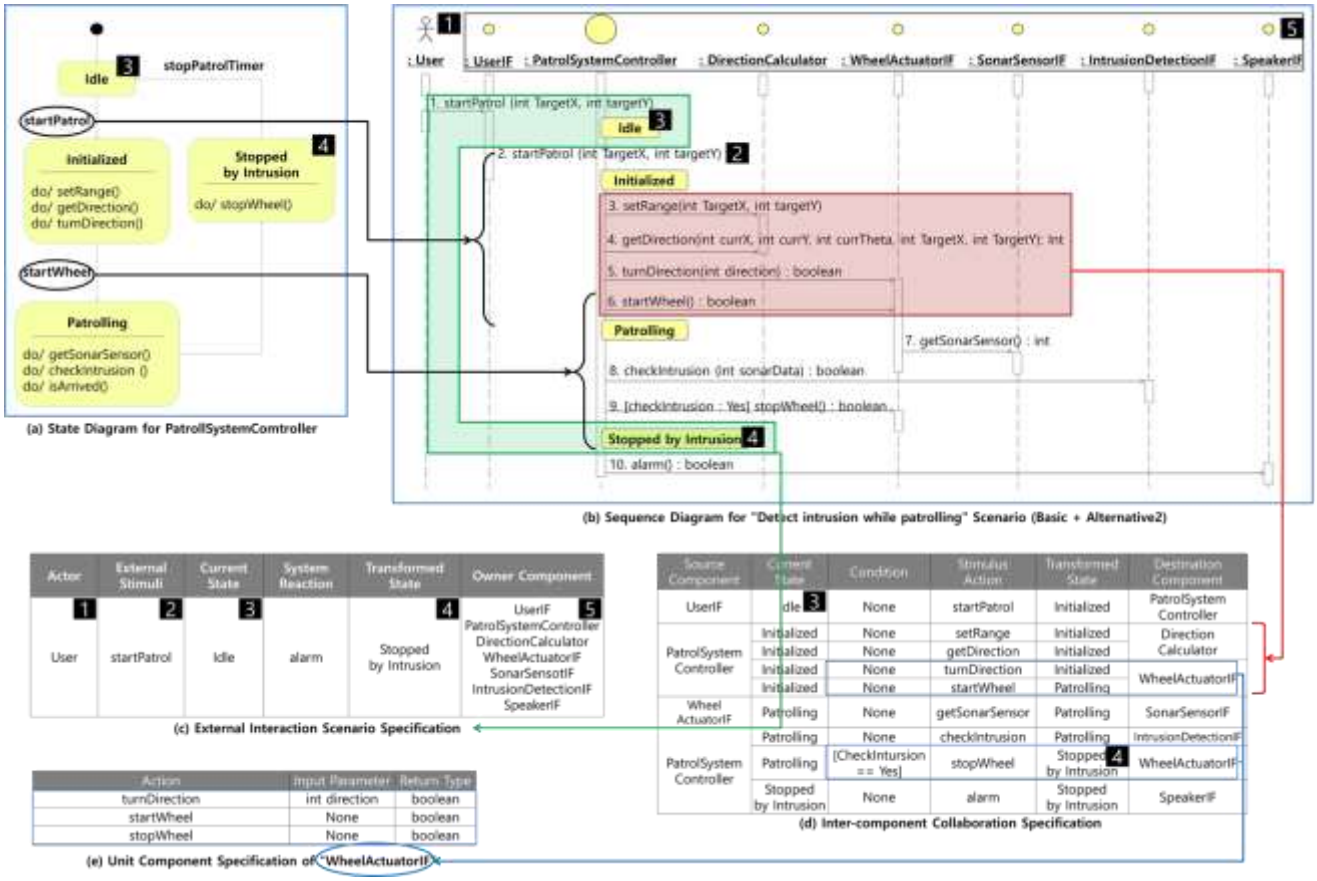


Fig. 9. Information Transformation from the Sequence Diagram to Requirements Specification for "Detect Intrusion While Patrolling" Scenario in the Robot Patrol System.

Unlike the general sequence diagram, one more thing to note is that the operation connected to each message in the sequencediagram must match the operation in the state diagram designed earlier. The operations that appeared in the state diagram and the state diagram are identical. As a result, the sum of the sequence diagrams created as many times as necessary contains all the information identified in the state diagram. Thus, the source of automatic generation of specifications through the following steps is the set of sequence diagrams as the result of step 5.

When step 5 is completed, sequence diagrams are created for each scenario combined based on the flow of events of the use case. As described above, the sequence diagram guided by the proposed model additionally specifies the controller's state transition information in the timeline. Using SpecGen, three additional requirement specifications are created through steps 6-8 based on the controller's state transition and the message sequence information that the controller controls to perform the scenario.

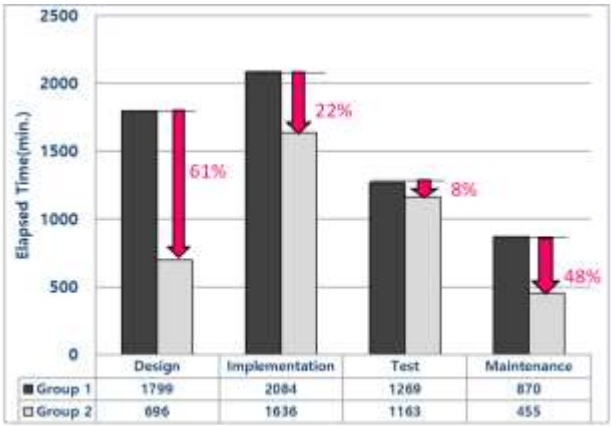
Step 6: Specify external interaction information

First, the contents of the external interaction scenario specification described in Fig. 9(c) are the same as the previous use case specification information. The external interaction scenario specification table specifies the stimuli and reactions between external actors and the whole system. Fig. 10 shows the relationship between the diagram and the automatically extracted and generated fragment of each specification. For understanding, the matching information is denoted by the same black-boxed number. The generated row in Fig. 9(c) specifies that User invokes startPatrol (External Stumilus) when the system is idle (Current State). Then, UserIF, PatrolSystemController, DirectionCalculator, WheelActuator-IF, SonarSensotIF, IntrusionDetectionIF, SpeakerIF (Owner Component) collaborate each other. The system's last reaction is to alarm (System Reaction), and the final state is stopped by intrusion.

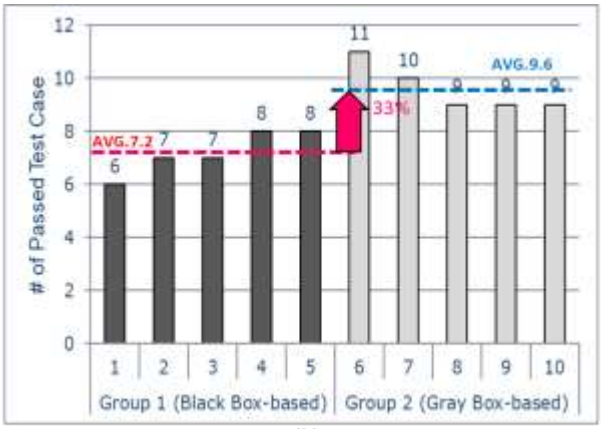
The specified content covers just a part of the "Patrol" use case. Fig. 9(c) specifies the external interaction scenario specification, including another scenario for the regular patrolling without any intrusion.

Step 7: Specify inter-component collaboration information

The second specification is generated from the message passing information in the sequence diagram. As annotated in the sequence diagram, the state transitions of the controller are also reflected in the inter-component collaboration specification. It is extracted one-to-one from each message on a sequence diagram. The first action to invoke each collaboration starts at the message from GUI (Graphical User Interface) object to the controller, not the message from an actor already specified in the external interaction scenario specification.

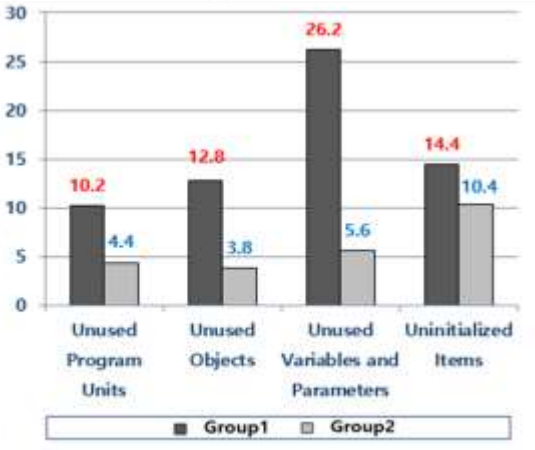


(a)



(b)

Group	Team	Unused Program units	Unused Objects	Unused Variables & Parameters	Uninitialized Items	Total number of Bad Symptoms
1 (Black -box)	1	0	3	22	2	27
	2	18	22	43	47	130
	3	10	10	22	13	55
	4	0	2	3	0	5
	5	23	27	41	10	101
Average		**10.2**	**12.8**	**26.2**	**14.4**	**63.6**
2 (Gray -box)	6	5	3	3	2	13
	7	3	2	6	10	21
	8	6	5	4	9	24
	9	4	7	12	19	42
	10	4	2	3	12	21
Average		**4.4**	**3.8**	**5.6**	**10.4**	**24.2**



(c)

Fig. 10. Comparison of the Experiment Results from Groups 1 and 2: (a) Elapsed Development Time, (b) The Number of Passed Test Cases, (c) The Number of Harmful Symptoms in the Code from Static Analysis.

Fig. 9(d) shows the inter-component collaboration specification for the "detect intrusion while patrolling" sequence diagram. A user invokes the startPatrol event. And then, the event makes UserIF trigger startPatrol message when the controller's state is Idle. Fig. 9(d) captures the message passing after the triggering according to the information in Fig. 9(b). The state transition of the PatrolSystemController in executing the "detect intrusion while patrolling" scenario is depicted in the Transformed State column in the specification. The state is transit from the Idle state to the states of Initialized → Patrolling → Stopped by Intrusion in the sequence.

Fig. 9(d) is the inter-component collaboration specification for the "Patrol" use case that includes the messages extracted from another sequence diagram for the regular patrolling scenario, denoted by shading. The content of Fig. 9(d) could be the key development specification for the developer in charge of the "Patrol" use case.

Step 8: Specify development requirements for unit components

The third specification automatically generated by SpecGen is the unit component specification. SpecGen classifies each message captured in the inter-component collaboration specification according to the destination component. For example, Fig. 9(e) is a table of the group of messages - turnDirection, startWheel, stopWheel- of which destination component is the same, WheelActuatorIF. The table becomes the unit component specification of WheelActuatorIF later if all of the other incoming messages to WheelActuatorIF appeared in other sequence diagrams are added. Fig. 9(e) is the completely generated unit component specification for the WheelActuatorIF of RPS. In the case of WheelActuatorIF, the extracted actions to implement the scenario in Fig.10 are equivalent to the whole set of actions in RPS. The three actions required to implement WheelActuatorIF are the same as the APIs for WheelActuator, which means that the developer in charge of the WheelActuator should implement each action in the unit component specification. And the other developers can reference the specification when they need to call any actions of WheelActuator.

VI. Evaluation of the Automatically Designed Requirements Specification for a Robot Patrol System

A. Experimental Design

To evaluate the effectiveness of the proposed requirements specification method, we designed an experiment. The scope of the experimental development is a basic flow and an alternative flow of the "*Patrol*" use case, which is described in

Table I. The participants of this experiment were 48 third or fourth-year university students from a computer engineering program, and they took the 8-week UML education course before they participated in the experiment. So, they can be classified as novice-level developers with more or fewer experiences in software development. 4 to 5 of them created a team.

Consequently, ten teams participated in this experiment. We divided the ten teams into two subgroups, one of which (group1) was given only the existing use case specification we named as the black box-based specification. The other group (group2) was given the artifact set of the provided requirements specification method as the gray box-based specification. The followings are the lists of the provided requirements specification artifacts for the two groups:

- For group1: black box-based requirements specification
    - Use case diagram
    - Use case specifications
- For group2: gray box-based requirements specification (black box-based requirements specification + additional requirements specifications generated by SpecGen)
    - Use case diagram
    - Use case specifications
    - External interaction scenario specification
    - Inter-component collaboration specification
    - Unit component specification

The results we wanted to confirm through this experiment and the corresponding measures we used were as follows:

- Enhancement of software development productivity: We compared the development time by asking each team to record each development phase's required time on the PSP sheet (Personal Software Process) [27]. The objective of the comparison is to confirm that the automatically generated requirements specifications from SpecGen contribute to decreasing embedded software development time.
- Enhancement of software product quality: We have defined 12 test cases for the given "Patrol" use case and tested the result from ten teams according to the test cases. And then, we compared the number of passed test cases related to functional aspects by each group, groups 1 and 2. We wanted to check if the provided requirements specifications from SpecGen can help the number of passed test cases increase.
- Enhancement of code quality: We expected that the requirements specifications generated from SpecGen contribute to enhancing the implemented code quality. If our expectation is correct, the number of bad symptoms from group 2 will be less than group 1. To confirm the assumption, we used the static analysis tool, Understand [28], to measure the number of bad symptoms inherent in the implementation codes from the two groups.

B. Experimental Results

- Enhancement of software development productivity: According to the PSP record documented during the two-week experimental development, the elapsed time of group1 in each development phase was shorter than group2. The teams' average total elapsed development time in group1 and group2 to develop the same use case, "Patrol," were 6,022 minutes and 3,950 minutes, respectively. The development time of group1 is the same as 66% of the elapsed time of group2.

As shown in Fig. 10(a), it is confirmed that group2 performed all steps, which are successive to the requirements specification, in less time than group1. In particular, the time taken for group2 to perform the design activity was less by 61% compared to group1. This decrease in development time can be interpreted as the benefit of the additionally provided requirements specifications generated by *SpecGen*, which include the analysis model in the early stage.

On the other hand, in the test phase, the execution time of group2 was less by only 8% compared to group1. In this experiment, the teams accomplished only integration tests since only a use case, "Patrol," was the development scope.

Since the use case specification is the exact requirements artifact provided for both groups, there is little difference in the information used for testing, so we can understand that there is no significant difference in the time taken for testing. To compare the time required for maintenance, we made the same request to change the requirements for each team belonging to both groups. The time to reflect the change request to the implementation code and test the changed code was measured as the maintenance time. As the teams in group2 can utilize the component collaboration specification generated from *SpecGen*, they took almost half the time of group1.

To ascertain that the elapsed time declines do not come from individual student's programming capabilities, we performed a simple regression analysis to analyze the correlation between the development time of each group and individual grades in programming-related courses. As a result, the R-Square value is 0.01, which explains no influence between students' development time and individual capabilities. Thus, the development time decline could be interpreted as the benefit of the proposed software requirements specifications.

- Enhancement of software product quality: We checked the number of passed test cases without any detected errors. The counted test cases are related to the functional requirements of the RPS, and the total test cases were 12. Fig. 10(b) shows that 7.2 test cases, averagely, are passed through the test in the product of group1. On the other hand, the average number of the passed test case in group2 was 9.6. This result means that compared with the development result of group1, the development result of group2 satisfies the given requirements by 33% more completely.

- Enhancement of code quality: We used Understand, a static analysis tool, to evaluate each group's quality of source codes. As a result, the number of detected bad symptoms of the source code implemented by group1 was 2.6 times larger than group2. The types of detected errors were unused program units, unused variables and parameters, unused objects, and uninitialized items, as shown in Fig. 10(c). These errors can be risks in software maintenance or reduce the efficiency of memory utilization in the future. Moreover, there is a wide variation in the number of detected errors extracted from five individual teams in group1, from 5 to 130. These figures produce evidence that there was no design guideline for developers (students), which can cause the quality of source code to depend wholly on individual developers' capability. On the other hand, we found a comparatively slight variation, from 13 to 42, in the numbers of detected total errors from five teams in group2. It shows that the proposed requirements specifications helped developers in group2 to construct uniformly qualified codes.

C. Comparison with Related Work

As proved by the experimental result, the proposed method help enhance the productivity of embedded software development and the quality of the product itself and implementation code. We analyze that the enhancement comes from providing (1) guidelines for the degree of detail for each analysis diagram, (2) support of an automating tool for the creation of specifications from the analysis diagrams, and (3) the specification methods for each development phase. Table II shows the results of comparing several related works and this study, based on the satisfaction of the features as mentioned above. Compared with that other related work [13, 18, 29, 30] limits providing guidelines for the degree of detail for each diagram and supporting an automatic tool for the proposed specification methods, Table II shows that this study acquires originality by providing the critical features mentioned above.

TABLE II. COMPARISON REQUIREMENT SPECIFICATION METHODS

	[13]	[18]	[29]	[30]	This Study
Guidelines for the degree of detail for each diagram	X	X	X	X	O
Automatic creation of specifications	X	X	X	X	O
Specification method for each deployment phase	O	O	O	X	O

VII. CONCLUSION

This study presents a gray box-based software requirements specification method for embedded system domain and guidelines for constructing an analysis model in the requirements phase, which can be a source of requirements extraction. The case study on a robot patrol system development demonstrates how the proposed guidelines are realized during the analysis model development and which information is documented as a requirements specification from the analysis model. An experiment to show the quantitative benefits of applying the proposed specification method and the revised supporting tool is conducted. The result of comparing this study and several related works based on the critical success features that brought about the enhancement demonstrated by the experimental results is also discussed.

Compared with our previous study, the originalities in this work could be captured in that:

- It proves the extensibility of the proposed gray box-based approach to automatic requirements specification by showing the result from applying it to the whole system of a robot patrol system different from the case study in the previous work. It shows that the proposed model is not a solution dedicated to a specific domain.
- It shows the evaluation results of the proposed approach with more various aspects. In addition to the decrease of the elapsed time for the software development phases after requirements, this study shows that the number of passed test cases of the target system can be increased by using the requirements specification automatically generated by the *SpecGen*, an automating tool for supporting the proposed model. Furthermore, the evaluation result shows that the source code's detected bad symptoms are decreased by a meaningful amount in the development group using the proposed approach compared with the other group not using it. All of the findings were measured quantitatively on an actual robot patrol system development, not a contrived system only for an experiment, which can be one of the originalities of our work.
- It provides more accessibility for embedded software developers by utilizing a more popular open-source UML authoring tool. In the previous work, the automating tool runs with ArgoUML. But, ArgoUML is not a widely used tool, and the upgrading is stopped. In this work, we re-build the automating tool, *SpecGen*, integrating with StarUML, one of the most popular UML authoring tools. Thus, more developers who already experienced StarUML can easily adopt *SpecGen* in their development.

REFERENCES

[1] E. A. Lee, "What's ahead for embedded software?," Computer, vol. 33, no. 9, pp. 18-26, 2000.

[2] G. Booch, J. Rumbaugh, and I. Jacobson, The unified modeling language user guide, Pearson Education India, 2005.

[3] T. Pereira, F. Alencar, and J. Castro, "Requirements Engineering for Embedded Systems: The REPES Process," in *Proceedings of the 21st Workshop on Requirements Engineering*, 2018.

[4] S. Ernst, B. Tenbergen and K. Pohl, "Requirements engineering for embedded systems: An investigation of industry needs," in *Proceedings of the Int. Working Conf. on Requirements Engineering: Foundation for Software Quality*, pp. 151-165, 2011.

[5] S. Ferg, "What's wrong with Use Cases?" Available at: http://jacksonworkbench.co.uk/stevefergspages/papers/ferg--whats_wrong_with_use_ cases.html (accessed 25/08/2021, 2021).

[6] P. Zave, "An Operational Approach to Requirements Specification for Embedded Systems," IEEE Transactions on Software Engineering, vol. SE-8, no. 3, pp. 250-269, 1982.

[7] J. M. Thompson, M. P. E. Heimdahl, and S. P. Miller, "Specification-Based Prototyping for Embedded Systems," in *Proceedings of ACM SIGSOFT Symposium on Foundations of Software Engineering 1999*, pp. 163-179, 1999.

[8] J. Lavi, and J. Kudish, "Systems modeling & requirements specification using ECSAM: an analysis method for embedded & computer-based systems," Innovations in Systems and Software Engineering, vol.1, pp.100-115, 2005.

[9] M. R. Sena Marques, E. Siegert, and L. Brisolara, "Integrating UML, MARTE and SysML to improve requirements specification and traceability in the embedded domain," in *Proceedings of the 12th IEEE International Conference on Industrial Informatics (INDIN)*, pp. 176-181, 2014.

[10] L. Dean, and W. Don, Managing software requirements: A use case approach, Addison-Wesley Professional, 2003.

[11] J. M. Nicholas, Project management for business and engineering: Principles and practice, Elsevier, pp.121. 2004.

[12] S. Park, "Software requirement specification based on a gray box for embedded systems: a case study of a mobile phone camera sensor controller," Computers, vol. 8, no. 20, pp. 1-11, 2019.

[13] N. G. Leveson, M. P. E. Heimdahl, H. Hildreth and J. D. Reese, "Requirements specification for process-control systems," IEEE Transactions on Software Engineering, vol. 20, no. 9, pp. 684-707, 1994.

[14] B. P. Douglass, Real-time UML : developing efficient objects for embedded systems, Addison-Wesley Longman Ltd., 2000.

[15] H. Gomma, "Designing concurrent, distributed, and real-time applications with UML," in *Proceedings of the 28th International Conference on Software Engineering*, pp. 1059-1060, 2006.

[16] J. Marin, T. Blanco, and J. J. Marin, "Octopus: A Design Methodology for Motion Capture Wearables," Sensors, vol. 17, no. 8, pp.1875, 2017.

[17] F. Lattemann, and E. Lehmann, "Methodological approach to the requirement specification of embedded systems," in *Proceedings of the International Conference on Formal Engineering Methods(ICFEM)*, pp. 183-191, 1997.

[18] J. Z. Lavi, and J. Kudish, "Systems modeling & requirements specification using ECSAM: a method for embedded computer-based systems analysis," in *Proceedings of the 11th IEEE International Conference and Workshop on the Engineering of Computer-Based Systems*, pp. 2-11, 2004.

[19] M. Glinz, "Statecharts for requirements specification-as simple as possible, as rich as needed," in *Proceedings of the ICSE2002 workshop on scenarios and state machines: models, algorithms, and tools*, 2002.

[20] C. Denger, D. M. Berry, and E. Kamsties, "Higher Quality Requirements Specifications through Natural Language Patterns," in *Proceedings of the 2003 IEEE International Conference on Software - Science, Technology and Engineering*, pp. 80-90, 2003.

[21] Y. Matsuo, K. Ogasawara, and A. Ohnishi, "Automatic Transformation of Organization of Software Requirements Specifications," *in Proceedings of the 4th International Conference on Research Challenges in Information Science*, pp. 269-278, 2010.

[22] S. Konrad and B. H. C. Cheng, "Facilitating the Construction of Specification Pattern-based Properties," in *Proceedings of the 13th International Conference on Requirements Engineering*, pp. 329-338, 2005.

[23] A. Post, I. Menzel, and A. Podelski, "Applying Restricted English Grammar on Automotive Requirements - Does it Work? A Case Study," in *Proceedings of the Requirements Engineering: Foundation for Software Quality*, pp. 166-180, 2011.

[24] M. Broy, and T. Stauner, "Requirements engineering for embedded systems," Informationstechnik und Technische Informatik, vol. 41, pp. 7-11, 1999.

[25] ArgoUML. Available at: https://sourceforge.net/projects/argouml/ (accessed 25/08/2021, 2021).

[26] StarUML. Available at: https://staruml.io/ (accessed 25/08/2021, 2021).

[27] W. S. Humphrey, The Personal Software Process (sm), vol. 11, Carnegie Mellon University, Software Engineering Institute, 2000.

[28] Understand. Available at: https://www.scitools.com/ (accessed 25/08/2021, 2021).

[29] F. Lattemann, and E. Lehmann, "A methodological approach to the requirement specification of embedded systems, " in *Proceedings of the 1st IEEE International Conference on Formal Engineering*, pp.83-191, 1997.

[30] M. Glinz, "Statecharts for requirements specification-as simple as possible, as rich as needed," in *Proceedings of the 24th International Conference on Software Engineering (ICSE 2002)Workshop: Scenarios and state machines: models, algorithms, and tool*, pp.1-6, 2002.

A Comparison of BAT and Firefly Algorithm in Neighborhood based Collaborative Filtering

Hartatik[1], Bayu Permana Sejati[2]
Faculty of Computer Science
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia

Hamdani Hamdani[3]*
Department of Informatics
Universitas Mulawarman
Samarinda, Indonesia

Andri Syafrianto[4]
Department of Informatics
STMIK El Rahma
Yogyakarta, Indonesia

***Abstract*—The recommender system is a knowledge-based filtering system that predicts the users' rating and preference for what they might desire. Simultaneously, the neighborhood method is a promising approach to perform predictions, resulting in a high accuracy based on the common items. This method, furthermore, could affect the resulting accuracy value because when each user provides limited data and sparsity, the accuracy of value might be narrow down as a consequence. In this research, we use the Swarm Intelligent (SI) technique in the recommender system to overcome this problem, whereby SI will train each feature to optimal weight. This technique's main objective is to form better groups of similar users and improve recommendations' accuracy. The intelligent swarm technique used to compare its accuracy to help provide recommendations is the Firefly and Bat Algorithm. The results show that the Firefly Algorithm has slightly better performance than the Bat Algorithm, with a difference in the mean absolute error of 0.02013333. The significance test using the independent t-test method states that no statistically significant difference between Bat and Firefly algorithm.**

***Keywords*—Bat algorithm; firefly algorithm; collaborative filtering; recommender system; swarm intelligent**

I. Introduction

The recommender system is research that is most popular in the business world. The recommender system is developed on several algorithms to find the best pattern from data and provide a recommendation by filtering a preference-based on the user's interests or needs [1], weighted update [2]. For example, one of the recommender system's prominent implementations is the amazon online store to offer similar books according to customer search history when visiting the online store.

Plenty of techniques can develop recommendation systems such as Content-Based Filtering, Collaborative Filtering, and Knowledge-Based Filtering. Content-Based Filtering, moreover, is a user modelling process in which users' interest is inferred from the items the user interacts with [3][4][5]. The items refer to usually textual, for example, email or web pages. In Content-based Filtering, the most defining features are used to model items and users. Meanwhile, the most discriminatory parts are identified and stored as vectors containing the components and their weights. The user model usually consists of user-item features. User models and recommendation candidates are compared to generate recommendations, for example, using the vector space model and the cosine similarity coefficient.

The second technique is Collaborative Filtering (CF). This technique will inform the user based on the feedback from other users who have relatively similar attributes [6][7]. The semblance of two users' tastes is calculated based on the history presented rating [8]. There are two CF approaches, namely the neighborhood-based method and the latent factor model - matrix factorization [8].

Neighborhood-based methods provide a study about the relationships between items or between users. The item approach is based on the user's preference for an object based on the same users' ranking of similar items [9]. Another case with the latent factor model - matrix factorization, which converts items and users to the same latent factor space [9]. The latent space is then used to explain the ranking by characterizing the product and user in terms of automatically generated factors from user feedback.

The neighborhood-based method is popular enough because of its simplicity, efficiency, and ability to produce accurate and personalized recommendations [8]. Neighborhood-based methods make predictions based on common items that help the accuracy value when each user provides limited data, and the spread, the resulting accuracy value will be small-scale [9]. Some researchers have tried adding techniques such as clustering [10] and intelligent swarm [11]. The addition of the SI uses the recommender technique to learn the optimal weight of each feature. Thus, it can form better groups of similar users and improve recommendations.

This research tried to combine IS techniques (BA and FA) with the neighborhood-based method to solve the sparsity problem and improve the accuracy of the recommendation system. First, IS methods are trained to get each feature's optimal weight, which is then used by the neighborhood-based method to get the rating prediction. At the end of the experiment, MAE and RMSE values will be obtained, which show the comparison of the accuracy of the BA and FA in providing recommendations.

II. Related Work

In the last few decades, metaheuristic techniques have experienced rapid development due to their increased search efficiency. Researchers have widely used the SI metaheuristic technique to find the most optimal solution search space

*Corresponding Author: hamdani@unmul.ac.id

phenomenon [11]. In System recommendations, the intelligent swarm is used to learn each feature's optimal weight to get a group of similar users who can be called active users [12].

The research uses weight updates for group decision-making that have similar parameters to be used by decision-makers (DMs). This model involves stakeholders who may have the same or different parameters in choosing parameters so that they can accommodate the interests of all DMs to obtain alternative decisions [2]. Furthermore, S. Ujjin and P. J. Bentley, [13] used the Particle Swarm Optimization (PSO) technique to generate a set of weights for user features and used a modified Euclidean function to generate recommendations. Their approach shows a considerable improvement over the Pearson algorithm compared to the Genetic algorithm.

Meanwhile, R. Katarya and O. P. Verma, [14] used K-Means to provide initial parameters for the PSO algorithm. The PSO algorithm itself is used to optimize Fuzzy C Means Clustering. Experiments conducted on the MovieLens Dataset show that there are several improvements from the existing method.

Research conducted by J. Sobecki, [15] compared several SI algorithms, namely Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Intelligent Weed Optimization (IWO), Bee Colony Optimization (BCO), and Bat Algorithm (BA) [16], in recommendations of student courses. Prediction accuracy shows a difference of only 0.1 between the five algorithms.

Tried to improve the CF-based recommendation system's quality using two Swarm Intelligence methods, namely BA and ABC [17]. This research aims to improve the quality of the traditional recommendation system. The authors use PCC to find similarities between users in the utility matrix and Swarm Intelligent methods to find the best weight of items to get good neighbors called active users. The data used in this study is the Jester dataset by Ken Goldberg, from AUTO Lab, UC Berkeley. This study indicates that BA has a 6.9% better quality than ABC, obtained a lower root mean squared error (RMSE) score than ABC, and higher precision, recall, and F1 score. In another research, [17]. Used BA to improve the recommendation system's quality [9]. The difference with the first research was in using the trust-aware matrix technique to create a utility matrix, which is then optimized - using BA. The data used in this study is a dataset from MovieLens, Epinions, CiaoDVD, and Filmtrust. To measure the performance of BA, the writer compared it with Particle Swarm Optimization (PSO), where BA obtained better results, both in terms of measuring Mean Squared Error (MAE) as much as 3.84% better and BA reaching 85.54% compared to PSO of 85.54% compared to PSO of 81.85%.

Another research combined Fuzzy C Means (FCM) and Bat Optimization to solve CF's sparsity and scalability problem [18]. BA serves to find the most optimal number of clusters from FCM. This study uses Movie Lens film rating data to be compared with the number of different neighbors and other clustering methods such as K-Means and SOM Clusters. This study indicates that the MAE score obtained by FCM and BA is smaller than other methods in the sense that BA can improve the quality of the clustering-based recommendation system.

Improving CF's performance using swarm intelligence can be done with several other swarm intelligence methods such as PSO. M. Wasid and V. Kant, introduced a new strategy in the recommendation system by combining it with Fuzzy Features or FPSO-CF [19]. This study aims to improve CF's accuracy, where PSO is used to find user weights and represent user features so that the authors use fuzzy sets more efficiently. Experiments were carried out using a movie lens dataset, with 60 films and 497 users. This study indicates that FPSO-CF has higher accuracy and lower error than Pearson CF, Fuzzy CF, and Fuzzy Genetic CF.

Research conducted by [20] using the Firefly Algorithm improves the quality of a collaborative filtering-based recommendation system. The MovieLens Dataset is used and collected by the GroupLens research project at the University of Minnesota [20]. The study results indicate that the proposed method significantly improves the recommendations' accuracy and improves the recommendations' prediction quality and performance.

The above studies prove that the combination of SI and CF techniques helps achieve better personalized recommendations for users. Therefore, continuing this work further, in this paper, we have tried to compare the Bat (BA) [21] algorithm and firefly [16], [22] in the calculation of feature weights and the measurement of the Pearson Correlation Coefficient to find the neighbors of active users. Finally, top-N recommendations were made for finding active users.

III. Proposed Method and Algorithm

A. Recommender System

A recommendation system can also be called a decision support system that can direct users to personalize items of interest or be liked by users. The recommendation system can provide users with inner directions by finding items that match user preferences [23]. The recommendation system's basic form works with two methods, specifically user-item interaction, such as rating or buying behavior, and attribute information about users and items such as profiles or search keywords. The first method is user-item interaction called collaborative filtering, while the second method is called content-based [24]. The content-based method works by checking the attributes of the recommended item. For instance, if Netflix users have watched many movies with the cowboy genre, then the next film to recommend is a cowboy. Meanwhile, collaborative filtering recommends items based on similarities between users or between items. In other words, items recommended to users are likes by other similar users [25].

B. Collaborative Filtering

Collaborative filtering (CF) is the most popular method of finding recommendations. CF works based on predictions and ratings or the behavior of other users in the system. The fundamentals behind this method are opinions of other users can be selected and aggregated in such a way as to provide a reasonable prediction of the preferences of the active user. It

can be intuitively assumed that if users agree about some items' quality or relevance, they will probably agree about other items. Another example, if a group of users likes the same things as Mary, then Mary is likely to like the things they like [26]. This approach is based on a simple idea; users will prefer items recommended by others who have something in the standard [27].

An example is like giving recommendations for places to eat that we like to our friends. What distinguishes other approaches is that CF only considers the utility matrix. CF is a stand-alone method because it does not know the item except the user [27].

C. Neighborhood-Based Collaborative Filtering

The neighborhood-based method, or it can be called memory-based, is the earliest method developed for collaborative filtering. This method is based on the fact that the users form a similar rating pattern on similar items [24], unlike the model-based approach, which is difficult to provide recommendations for films that do not match the film's information. As a result, a model-based recommendation system recommends films that are not according to user preferences [28]. Meanwhile, the neighborhood-based method can address those weaknesses [28].

D. User-Based Collaborative Filtering

This method is designed to find similarities from users who have similar rating patterns to other users and rated the item in question [27]. For instance, if Alice and Bob have rated films the same way in the past, other users could use Alice's ratings in film A to predict Bob's non-rated ratings on film A. In general, most similar users to Bob can be used to make ranking predictions for Bob. The similarity function is calculated between rating rows to find similar users. Pearson Correlation Coefficient and Cosine Similarity could be used to calculate similarity [29].

Pearson Correlation Coefficient (PCC) is a correlation search method developed by Karl Pearson. Meanwhile, correlation is a measurement technique that determines how close the relationship between the two variables is. The measurement results of the PCC can be either positive or negative. A positive relationship shows that the two variables have a parallel (linear) increase in value. Meanwhile, a negative relationship shows that the two variables have a parallel (linear) decrease in value. A parallel is an increase or decrease in value that follows between two variables. Equation (1) is used to calculate the similarity between users or items [24].

$$sim(u,v) = \frac{\sum_{i \in C}(r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in C}(r_{u,i} - \bar{r}_u)^2}\sqrt{\sum_{i \in C}(r_{v,i} - \bar{r}_v)^2}} \quad (1)$$

Where, $sim(u,v)$ is the similarity value between user u and user v, $r_{u,i}$ and $r_{v,i}$ are the values of user u and user v against the item, and $\bar{r}_u$ and $\bar{r}_v$ are the mean values of user u and user v against the item.

E. BAT Algorithm

Bat Algorithm (BA) is a metaheuristic method based on swarm intelligence proposed by Xin-She [30] and BA the proposed work is formulated as a non-linear optimization problem [16]. BA was inspired by small bats (microbats) that use echolocation/sonar to detect prey [12]. Most of the bat species are insectivorous. Besides using echolocation to catch food, bats also use it to avoid obstacles and find perches in the dark. Echolocation works by emitting sound, and then when the sound hits the object, the sound will return to the source [30]. Bats fly randomly and have position x_i, velocity v_i, frequency f_i, pulse rate r_i and loudness A_i, to find the optimal solution. The bat moves each iteration to come up with a new solution that may be more optimal [12]. The following are the steps of the BA algorithm. A New Discretization of Bat algorithm in Equation (2), (3), and (4).

$$f_i = f_{\min} + (f_{\max} - f_{\min})\beta \quad (2)$$

$$v_i^{t+1} = v_i^t + (x_i^t - x_{best})f_i^{(t)} \quad (3)$$

$$x_i^{t+1} = x_i^t + v_i^t \quad (4)$$

Where, β is the random vector between $[0,1]$, and $t+1$ is a number of iteration.

Update the best position of *x* best using (5).

$$x_{new} = x_{old} + \in A^t \quad (5)$$

Where, $\in$ determines the measurement and helps in getting the convergence velocity from BA.

Save the best solutions. This step save the best current solution as well as update the loudness A_i and pulse rate r_i Equation (6) and (7).

$$A_i^{t+1} = \alpha A_i^t \quad (6)$$

$$r_i^{t+1} = r_1^0(1 - e^{(-\gamma)}) \quad (7)$$

F. Firefly Algorithm

Firefly algorithm is a SI method inspired by nature, namely the firefly lifestyle. Xin-She developed this method in 2007 [31] [24]. As with other SI methods, the Firefly algorithm aims to solve optimization problems like the firefly lifestyle. Fireflies produce short, rhythmic lights that are different from one another. Firefly populations each have their lighting characteristics. In this method, the firefly is compared, and the less bright fireflies move towards, the brighter fireflies [32]. The firefly chosen as the most attractive is the optimal response to the problem [20], dynamic adaptive [22], and hybrid firefly algorithm [16].

Fireflies' attraction is their brightness, the light intensity at a certain distance from the light source as an inverse-square law. That means that the light intensity will decrease with increasing distance. The fireflies' attractiveness is directly

proportional to the nearby fireflies' light intensity and is measured depending on the length of the fireflies from one another. The equation of variation of attractiveness β with distance r is defined (8) [33].

$$\beta = \beta_0 . e^{-\gamma . r^2} \tag{8}$$

Where β_0 is the value of the firefly attractiveness when $r = 0$ and γ is the light absorption coefficient. The position shift of the firefly i attracted to the firefly j is determined by (9) [20].

$$x_i = x_i + \beta_0 . e^{-\gamma . r_{i,j}^2 (x_j - x_i)} + \alpha . (rand - 0.5) \tag{9}$$

G. Rating Prediction

Collaborative filtering aims to predict an empty rating of the utility matrix. There are various methods to do this, and one of those is the k-Nearest Neighbor (k-NN). k-NN approach is one of the data mining methods considered among the top 10 data mining techniques [20]. The k-NN approach uses the well-known concept of *Cicero pares cum paribus facility congregant* (birds of a feather flock together or equivalent to equals easily associated). It attempts to identify an unknown sample based on the known classification of its neighbors. k-NN is used to predict the consumer's rating, following the k-NN method [14](10).

$$\hat{r}_{ui} = \frac{\sum_{v \in N_i^k(u)} sim(u,v) . r_{vi}}{\sum_{v \in N_i^k(u)} sim(u,v)} \tag{10}$$

Where $sim(u,v)$ is the similarity value between user u and user v, $\hat{r}_{ui}$ is predicted values of user u for the items i and r_{vi} are the values of user v for the item i.

H. Evaluation

The recommendation engine's output is also calculated as a rating Root Mean Squared Error (RMSE) or Mean Absolute Error (MAE), calculating the delta between real and expected ratings. For more significant errors, the RMSE metric is used, while MAE benefits from simple understanding.

Whenever a new known rating is captured in the system, the deployed recommendation engine's accuracy must be continuously measured. The new actual rating is paired against the recommender's previous prediction [27]. MAE and RMSE formula can be seen at (11) and (12), respectively.

$$MAE = \frac{\sum_{i=1}^{n} | p_i - a_i |}{n} \tag{11}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} | p_i - a_i |^2}{n}} \tag{12}$$

To determine whether the differences between the two proposed models are statistically significant or not, we applied an independent t-test for an unpaired sample [16], [20]. The Independent sample t-test is defined by (13).

$$t = \frac{(\bar{x}_a - \bar{x}_b)}{\sqrt{\frac{S_a^{\,2}}{n_a} + \frac{S_b^{\,2}}{n_b}}} \tag{13}$$

This study used data from an open dataset at movielens.org, which contained 100,000 users who rated the film [27]. The distribution of the dataset can be seen in Table I.

TABLE I. DATA DISTRIBUTION IN DATASET

Data	Ratings	Users	Movies
Training	72456	943	1639
Testing	18114		
Totals	90570		

These stages begin with data collection from the MovieLens 100K site. A preprocessing process consists of features extraction to obtain movies id, users id, ratings, and timestamp. Furthermore, the data is transformed into a utility matrix to look for similarities using the Pearson Correlation Coefficient (PCC). The similarity matrix results are optimized using SI to obtain each user's weight. After that, data is divided into training data and testing data. The rating prediction was calculated using k-NN method by considering the user weight. The steps above could see in Fig. 1.

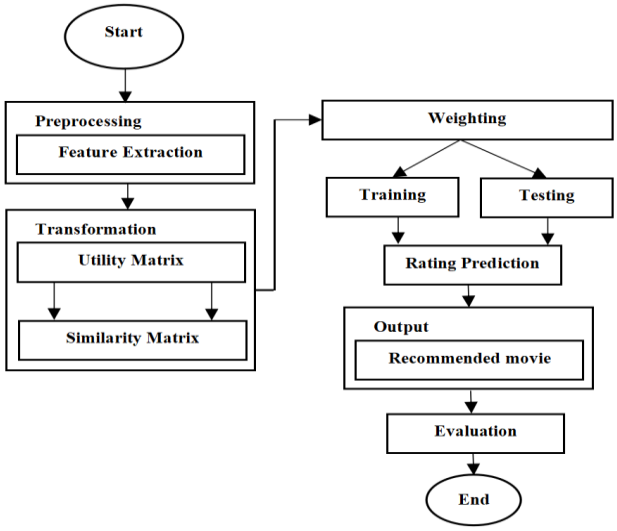


Fig. 1. Our Proposed Approach.

IV. PROPOSED METHOD

This research used data in the MovieLens 100K dataset, consisting of 943 users and 1680 films with a spread rate of 99.1%. The first stage is the data pre-processing. The purpose of data pre-processing is to extract features from the rating data in order to get useful features before making predictions. The next step is to carry out each user's weighting process using the

BA and Firefly algorithm. The BA and Firefly algorithm do the process repeatedly to get the best solution.

In the BA implementation, the initial alpha value is 0.95, and the gamma is 0.05. With a bat's population is 50, times the number of dimensions in the similarity matrix is 943, and a generation is initiated by multiplying population and dimension (50*943). Different from BA, there is no initial dimension at parameter initiation in the Firefly algorithm.

Parameter alpha in the Firefly algorithm was set at 1.0, beta at 1, and gamma at 0.95. The generation and the number of dimensions in the Firefly similarity matrix algorithm were set the same with BA in 943. Furthermore the sis initiation parameters can be seen in Table II.

TABLE II. PARAMETER INITIATION OF EACH SIS

Parameters	Bat Algorithm	Firefly Algorithm
Dimension	943	-
Population	50	943
Generation	50*943	943
Alpha	0.95	1.0
Beta	-	1
Gamma	0.95	0.01

When a new active user enters the system whose suggestions are to be made, BA and Firefly iteratively optimize the feature (item) weight by searching in the search space dimension m (item total). The results of this calculation can be seen in Table III.

TABLE III. WEIGHTING RESULT COMPARISON OF BAT AND FIREFLY

Methods	Max Weights	Min Weights	Avg Weights	Std Weights
BA	4.600450	-4.731380	-0.048353	1.695925
Firefly	0.257814	0.001194	0.006614	0.017754

After the feature weighting has been studied, active user neighbours are then formed using PCC to measure the distance between two similar users using (1). PCC is modified by multiplying the actual rank by the weight calculated by the BA and Firefly algorithms.

V. RESULT AND DISCUSSION

This section describes the results of the experiments that have been carried out and the work's findings. Several previous studies have stated that adding SI to traditional methods such as PCC can provide better predictive values. Therefore, in this experiment, we also add the conventional method's experimental results without SI modification.

An active user that was generated with BA and Firefly picked up from the highest weight. In this sense, we compare the error calculation results between the models trained using 50,100 and 200 active users in this test. The mean absolute error and RMSE values obtained using PCC, BA, and Firefly algorithms are shown in Table IV. The results show the active users generated by SI can improve the recommendation system's quality by 20%.

TABLE IV. RESULT OBTAINED USING BA AND FIREFLY ON MOVIELENS DATA SET

Methods	Active Users	MAE	RMSE	Time (s)
PCC	50	0.7331	1.0111	2.17
	100	0.7438	1.0392	5.25
	200	0.7231	1.0014	20.5
BA	50	0.6033	0.8669	10.8
	100	0.6287	0.8839	4.53
	200	0.6506	0.9010	17.7
Firefly	50	0.6010	0.8680	6.69
	100	0.6049	0.8695	4.9
	200	0.6163	0.8727	11.7

Fig. 2 and Fig. 3 graphically show the MAE variation for different numbers of active users randomly selected using both algorithms.

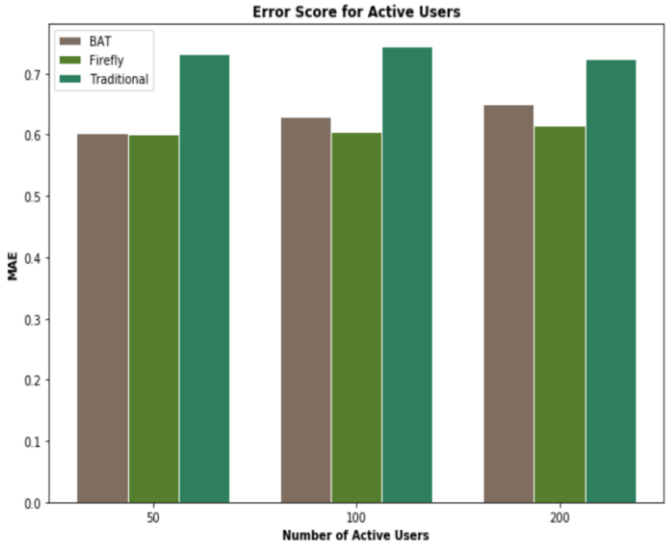


Fig. 2. Comparison MAE of user Active and non user Active.

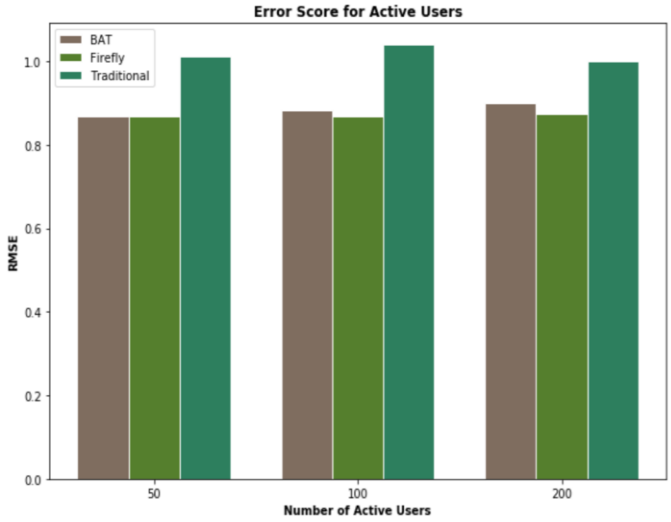


Fig. 3. Comparison RMSE of user Active and non user Active.

Thereafter, based on the experiment results using the active user, the recommender system that uses a certain optimal swarm intelligence weight has decreased in the error calculation. For MAE and RMSE scores, the Firefly Algorithm has a slightly better score than the BA. Meanwhile, the weightless recommendation system shows that the MAE and RMSE results are still relatively high. It concludes using the weightiest users affects the quality of the recommendation system. However, the execution time on the weightless recommendation system has a shorter time than the weighted ones. It could see in Fig. 2 and 3 sections 50 users that there is a significant difference in execution time, although, at 200 users, the time from the traditional recommendation system has increased. It concludes the more users there are, the more execution time is needed and does not affect the increase or decrease in the error calculation result.

In this research, we applied an independent sample t-test to saw the significance of the BAT and Firefly algorithm. The results of the independent t-test can be seen in Table V.

TABLE V. THE RESULT OF INDEPENDENT T-TEST ON MOVIELENS DATA SET

User Active	**MAE**	
	BAT	**Firefly**
50	0.6033	0.601
100	0.6287	0.6049
200	0.6506	0.6163
Mean	0.62753333	0.6074
Standard deviation	0.023671572	0.007950472
α	0.05	
Degree of freedom	4	
t-Value calculated using (3)	1.140233717	

Tests to determine the null hypothesis are $H_0 : \mu_1 = \mu_2$. it means that there is no difference between BAT (μ_1) and Firefly algorithm (μ_2). Meanwhile, the alternative hypothesis was $H_1 : \mu_1 \neq \mu_2$ which is explains there is a difference between BAT (μ_1) and Firefly algorithm (μ_2).

Refer to the test results in Table IV, the t-value = 1.140233717 and df = 4. Using the Two Tails T Distribution Table, the t-table value = 2.776 for the error rate of 5% and 4.604 for the error rate of 1%. Since the t-value is in the area of acceptance for the H_0 hypothesis at an error level of 5% or 1%, it can be concluded that there is no statistically significant difference between BAT and Firefly algorithms.

VI. CONCLUSION

The neighborhood method is a promising approach to perform predictions, resulting in a high accuracy based on the common items. This method, furthermore, could affect the resulting accuracy value because when each user provides limited data and sparsity, the accuracy of value might be narrow down as a consequence. The recommender system can use SI technique to overcome this problem. In this work, we proposed an approach to giving weights to the items in the user-item rating matrix to find the active user's better neighborhood using the BA and firefly algorithm. This method helped provide personalized recommendations to all users as it generated a different set of weights for each user.

The MAE value shows that SI can improve the recommender system's quality by 20%. For MAE scores, the Firefly Algorithm has a slightly better score than the BA. BA had MAE score 0.6033, 0.6287 and 0.6506 for k=50, 100 and 200 respectively. Meanwhile, Firefly Algorithm had an MAE score of 0.601, 0.6049, and 0.6163 for k=50, 100, and 200, respectively.

The test result using the independent t-test method got t-value=1.140233717. Since the t-value accepted the null hypothesis at an error level of 5% or 1%, the conclusion was there is no statistically significant difference between BA and Firefly algorithms.

VII. FUTURE WORK

In our future work, we would like to add implicit feedback as an additional feature in IS. This thinking stems from the assumption that users' preferences and preferences may change over time. Thus, the addition of the time parameter may make the resulting prediction more relevant and follow the user's interests.

In addition, it is necessary to think of a way to reduce the scalability of the recommendation system in the future. It is a typical case that adding IS slows down the recommendation calculation process. Therefore, it is necessary to think of a way to reduce this problem. The application of the clustering method at the beginning is probably one way that can do.

REFERENCES

[1] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, “Recommendation systems: Principles, methods and evaluation,” Egypt. Informatics J., vol. 16, no. 3, pp. 261–273, 2015, doi: 10.1016/j.eij.2015.06.005.

[2] H. Hamdani, R. Wardoyo, and K. Mustofa, “A method of weight update in group decision-making to accommodate the interests of all the decision makers,” Int. J. Intell. Syst. Appl., vol. 9, no. 8, 2017, doi: 10.5815/ijisa.2017.08.01.

[3] Y. Afoudi, M. Lazaar, and M. Al Achhab, “Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network,” Simul. Model. Pract. Theory, vol. 113, no. July, p. 102375, 2021, doi: 10.1016/j.simpat.2021.102375.

[4] B. Walek, V. Fojtik, “A hybrid recommender system for recommending relevant movies using an expert system”, Expert Systems with Applications 158, 113452, 2020.

[5] E. Faizal and H. Hamdani, “Weighted Minkowski similarity method with CBR for diagnosing cardiovascular disease,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 12, 2018, doi: 10.14569/IJACSA.2018.091244.

[6] H. Khojamli and J. Razmara, “Survey of similarity functions on neighborhood-based collaborative filtering,” Expert Syst. Appl., vol. 185, no. June 2020, p. 115482, 2021, doi: 10.1016/j.eswa.2021.115482.

[7] X. Yuan, L. Han, S. Qian, G. Xu, H. Yan, “Singular value decomposition based recommendation using imputed data”, Knowledge-Based Systems, Volume 163, 1 January 2019, Pages 485-494.

[8] R. Zhang, Q. D. Liu, Chun-Gui, J. X. Wei, and Huiyi-Ma, “Collaborative Filtering for Recommender Systems,” Proc. - 2014 2nd Int. Conf. Adv. Cloud Big Data, CBD 2014, pp. 301–308, 2015, doi: 10.1109/CBD.2014.47.

[9] F. Ricci, L. Rokach, B. Shapira, P. B. Kantor, and F. Ricci, Recommender Systems Handbook. 2011.

[10] L. Xiaojun, "An improved clustering-based collaborative filtering recommendation algorithm," Cluster Comput., vol. 20, no. 2, pp. 1281–1288, 2017, doi: 10.1007/s10586-017-0807-6.

[11] G. Guo, J. Zhang, and N. Yorke-Smith, "A Novel Recommendation Model Regularized with User Trust and Item Ratings," IEEE Trans. Knowl. Data Eng., vol. 28, no. 7, pp. 1607–1620, 2016, doi: 10.1109/TKDE.2016.2528249.

[12] S. Yadav, Vikesh, Shreyam, and S. Nagpal, "An Improved Collaborative Filtering Based Recommender System using Bat Algorithm," Procedia Comput. Sci., vol. 132, pp. 1795–1803, 2018, doi: 10.1016/j.procs.2018.05.155.

[13] S. Ujjin and P. J. Bentley, "Particle swarm optimization recommender system," 2003 IEEE Swarm Intell. Symp. SIS 2003 - Proc., pp. 124–131, 2003, doi: 10.1109/SIS.2003.1202257.

[14] R. Katarya and O. P. Verma, "A collaborative recommender system enhanced with particle swarm optimization technique," Multimed. Tools Appl., vol. 75, no. 15, pp. 9225–9239, 2016, doi: 10.1007/s11042-016-3481-4.

[15] J. Sobecki, "Comparison of selected swarm intelligence algorithms in student courses recommendation application," Int. J. Softw. Eng. Knowl. Eng., vol. 24, no. 1, pp. 91–109, 2014, doi: 10.1142/S0218194014500041.

[16] Z. Cheng, H. Song, J. Wang, H. Zhang, T. Chang, and M. Zhang, "Hybrid firefly algorithm with grouping attraction for constrained optimization problem," Knowledge-Based Syst., vol. 220, p. 106937, 2021, doi: 10.1016/j.knosys.2021.106937.

[17] S. Yadav, V. Kumar, S. Sinha, and S. Nagpal, "Trust aware recommender system using swarm intelligence," J. Comput. Sci., vol. 28, pp. 180–192, 2018, doi: 10.1016/j.jocs.2018.09.007.

[18] V. Vellaichamy and V. Kalimuthu, "Hybrid collaborative movie recommender system using clustering and bat optimization," Int. J. Intell. Eng. Syst., vol. 10, no. 5, pp. 38–47, 2017, doi: 10.22266/ijies2017.1031.05.

[19] M. Wasid and V. Kant, "A Particle Swarm Approach to Collaborative Filtering based Recommender Systems through Fuzzy Features," Procedia Comput. Sci., vol. 54, pp. 440–448, 2015, doi: 10.1016/j.procs.2015.06.051.

[20] F. Shomalnasab, M. Sadeghzadeh, and M. Esmaeilpour, "An Optimal Similarity Measure for Collaborative Filtering Using Firefly Algorithm," J. Adv. Comput. Res., vol. 5, no. August, pp. 101–111, 2015.

[21] Y. Saji and M. Barkatou, "A discrete bat algorithm based on Lévy flights for Euclidean traveling salesman problem," Expert Syst. Appl., vol. 172, no. January, p. 114639, 2021, doi: 10.1016/j.eswa.2021.114639.

[22] J. Liu, Y. Mao, X. Liu, and Y. Li, "A dynamic adaptive firefly algorithm with globally orientation," Math. Comput. Simul., vol. 174, pp. 76–101, 2020, doi: 10.1016/j.matcom.2020.02.020.

[23] A. Felfernig, L. Boratto, and W. Gan, "Group Recommender Systems," Springer, 2018, pp. 1–176.

[24] L. Luo, H. Xie, Y. Rao, F.L. Wang, "Personalized Recommendation by Matrix Co-Factorization with Tags and Time Information", Expert Systems With Applications, Volume 119, 1 April 2019, Pages 311-321.

[25] J. Leskovec, A. Rajaraman, and J. D. Ullman, "Mining of Massive Datasets," Min. Massive Datasets, 2020, doi: 10.1017/9781108684163.

[26] M. D. Ekstrand, J. T. Riedl, and J. A. Konstan, "Collaborative filtering recommender systems," Found. Trends Human-Computer Interact., vol. 4, no. 2, pp. 81–173, 2010, doi: 10.1561/1100000009.

[27] V. Kotu and B. Deshpande, Recommendation Engines. 2019.

[28] C. Desrosiers and G. Karypis, "Recommender Systems Handbook," Recomm. Syst. Handb., pp. 107–108, 2011, doi: 10.1007/978-0-387-85820-3.

[29] Z. Tan and L. He, "An Efficient Similarity Measure for User-Based Collaborative Filtering Recommender Systems Inspired by the Physical Resonance Principle," IEEE Access, vol. 5, pp. 27211–27228, 2017, doi: 10.1109/ACCESS.2017.2778424.

[30] X. S. Yang, "A new metaheuristic Bat-inspired Algorithm," Stud. Comput. Intell., vol. 284, pp. 65–74, 2010, doi: 10.1007/978-3-642-12538-6_6.

[31] X. Yang, "Metaheuristic Optimization : Nature-Inspired," Artif. Intell., Evol. Comput. Metaheuristics, pp. 405–420, 2013.

[32] H. Wang et al., "Firefly algorithm with neighborhood attraction," Inf. Sci. (Ny)., vol. 382–383, pp. 374–387, 2017, doi: 10.1016/j.ins.2016.12.024.

[33] A. B. Downey, "Think Stats: Probability and Statistics for Programmers," Psychol. Bull., vol. 70, no. 2, p. 140, 2011.

Modeling a Fault Detection Predictor in Compressor using Machine Learning Approach based on Acoustic Sensor Data

Divya M.N[1]
Research Scholar, VTU Research Centre, MSRIT, VTU, Belagavi, Karnataka, India
School of ECE, REVA University, Bengaluru, India

Narayanappa C.K[2]
Department of Medical Electronics, M.S. Ramaiah Institute of Technology, Bangalore, Karnataka, India

Gangadharaiah S.L[3]
Department of Electronics and Communication, M. S. Ramaiah Institute of Technology, Bengaluru, India

***Abstract*—Proper functioning of the air compressor ensures stability for many critical systems. The ill-effect of the breakdown caused by the wear and tear in the system can be mitigated if there exists an effective automated fault classification system. Traditionally, the simulation-based methods help to extend to identify the faults; however, those systems are not so effective enough to build real-time adaptive methods for the fault detection and its type. This paper proposes an effective model for the fault classification in the air compressor based on the real-time empirical acoustic sensor time-series data were taken on a sampling frequency of 50Khz. In the proposed work, the time-series datais transformedinto the frequency domain using fast Fourier transforms,where half of the signals are considered due to its symmetric representation. Afterward, a masking operation is carried out to extract significant feature vectors fed to the multilayer perception neural network. The uniqueness of the proposed system is that it requires less trainable parameters, thus reduces the training time and imposes lower memory overhead. The model is benchmarked with performance metric accuracy, and it is found that the proposed masked feature set-based MLP-ANN exhibits an accuracy of 91.32%. In contrast, the LSTM based fault classification model gives only 83.12% accuracy, takes more training time, and consumes more memory. Thus, the proposed model is realistic enough to be considered a real-time monitoring system of the fault and control. However, other performance metrics like precision, recall, and F1-Score are also promising with the LSTM based fault classifier.**

***Keywords*—Air-compressor; fault detection; LSTM; multi-layer perception; ANN; acoustic sensor data**

I. INTRODUCTION

The air compressors (AC) play a significant role in essential functions like fuel injection and metal finishing in the aircraft's design [1]. The ACs are used widely in thermal plants [2], power generation systems [3], vehicle propulsion [4] and pipeline systems [5], etc. In building an effective quality control system, the compressor simulation plays an essentialrole in evaluating the tolerable pressure by the different components of the aircraft while in transit [6]. Specifically, the aircraft manufacturer depends on the very high quality of the compressors for every phase of the production for the operative success of the functions. Another important aspect is that the aircraft components thwart the contamination due to mixing the air with the lubricants in ACs. Fig. 1 demonstrates the common applications in aircraft ACs.In order to design the complete line of the product of anair compressor, a compress air system is used. Two main air compressors are widely used: i) Rotary screw AC, and ii) Reciprocating ACs used depending upon the application requirements.

The automated fault classification (AFC) problem in compressors has attracted researchers to address the issues and build solution paradigms in the recent past so that early detection can minimize the damage caused to the overall system. The early warning systems are a step towards preventive maintenance [7], which is broadly classified into two categories,namely: i) maintenance on breakdown [8] and ii) condition-based maintenance (CBM) [9-10]. However, the CBMhas taken an edge over the maintenance on breakdown because the CBMperforms both the detection and seclusion of the faults that occurred in the early stage of the breakdown itself. The early and intelligent fault detection system [11] synchronizes with the conditions [12] of the various machine aspects dynamic changes in pressure, temperature, vibration, or acoustics [13]. Dhosi et al. [14] reviewed the correlation of vibrations and fault for various machines like pumps, turbines, and compressors. However, the most predominant measurement for fault detection is an acoustic signal.Some of the recent studies include the fault diagnosis in the planetary gearbox [15], Polak et al. [16], have highlighted the fault detection in the compressor, combustion engines using acoustic signals, last but not least, Ahmed et al. [17] reviews the use of acoustic measurement for the power unit fault detection in aircraft.

This paper proposes fault classification in the air compressor based on the real-time empirical acoustic sensor data. However, the challenging point in designing such an intelligent fault detection system (IFDS) is to identify the sensitive points from where the signals are acquired, and that is performed by the mechanism of the sensitive point or position analysis (SPA) [18-19]. The raw data collected from all the

sensitive positions are exposed to various noises requiring appropriate denoising treatments [20]. The learning model does not take these de-noised data directly into its computing model. Therefore, the significant information representation of the acoustic signals using mathematical models plays a vital role in accuracy for both dimensionality reduction and predictor performance. Various representations of the signals, including time, frequency, and time & frequency domains, are found in the designing of the fault detection systems for pipeline leakage [21], mechanical compound fault [22], centrifugal compressor [23], and reciprocating compressor [24]. Theoretically, the detection probability and accuracy are assumed to be higher if more features are considered while designing the detection models [25]. However, in contrast, the machine learning-based model performances degrade with the higher number of features [26] due to ill-effect posed by high polynomiality complexities and approximation.

Furthermore, this is handled by using an effective feature selection mechanism towards dimension reduction techniques like Partial least squares (PLS) used in the gas turbine's compressor blade [27]. Many other popular ones used in machines related faults detection systems are PCA [28], the variance of the ICA [29], etc. Moreover, there exist various types of faults which is trained to the model using these feature sets and many of the learning models as a function approximator such that if Yn represents all the feature set, then a function Fn emerge out as Fn(Yn) → C, where C is the class or the type of the faults. Popular learning models used in the fault classification in the machine are linear-SVM used for vehicle power systems [30], CNN for bearing fault classification [31], and LSTM for compressor valves [32]. Towards achieving better classification performance, this paper has presented an effective model for the fault-classification in the air compressor using a multilayer perceptron neural network that overcomes memory overhead, unlike Long Short Term Memory (LSTM) based approaches. The paper utilizes the autistic signal data prepared by Verma et al. [33] from 24 sensor positions based on SPA from an air compressor. The entire process flow of this method is described in the respective sections in this paper as a snap-view given in Fig. 2.

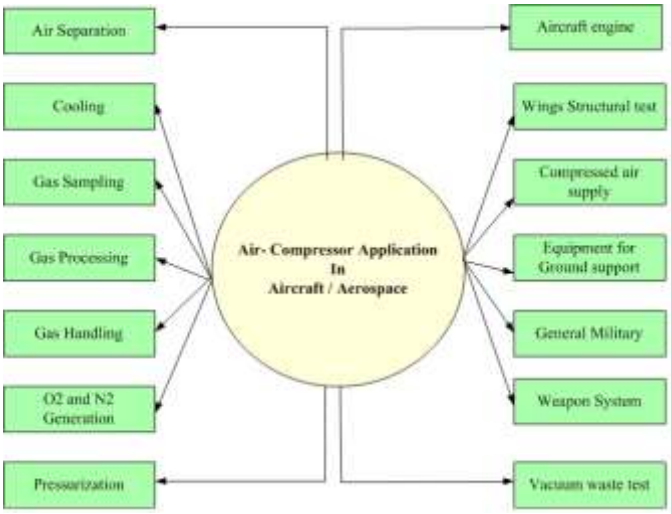


Fig. 1. Air Compressor Application in Aircraft.

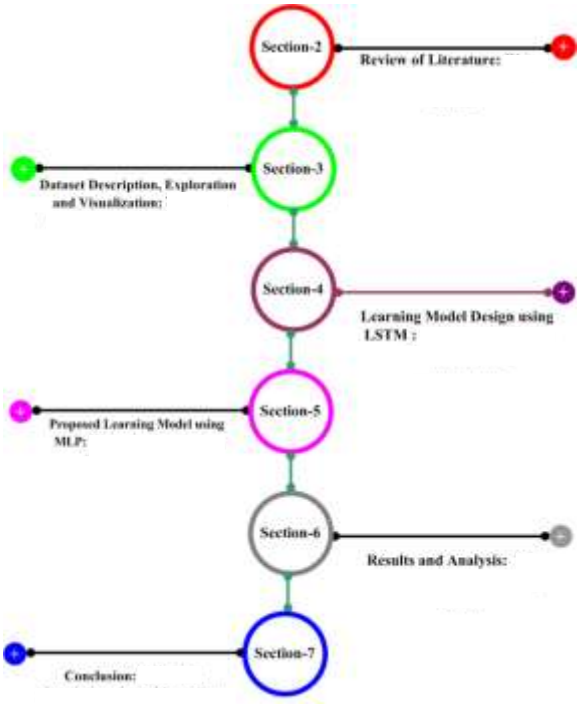


Fig. 2. Snap View of the Paper Organization.

As mentioned in the above Fig. 2 the remaining part of the paper are organized as follows: Section II discusses the related work in the context of AC fault prediction, Section III presents dataset visualization and analysis; Section IV discusses the implementation of LSTM model; Section V discusses implementationof the proposed model based on MLP; Section VI presents the result and performance analysis of the proposed system and finally entire work of this paper is concluded in Section VII.

II. Review of Literature

This section discusses the related work carried out towards air compressor fault prediction. In the existing literature, there are two kinds of approaches used to predict fault AC classification. The first one is the model-based approach, and the second is the data-driven approach [34]. The model-based approaches utilize mathematical modeling for machine life estimation and fault prediction [35-36]. In contrast, data-driven approaches are based on statistical analysis and soft computing approaches like machine learning, deep learning, and evolutionary. However, model-based approaches involve complicated procedures to describe attributes of the mechanical system [37]. Conditions of a mechanical system like air compressors can be analyzed by processing sensory data using data-driven and soft-computing approaches [38]. The work carried out by Ouadine et al. [39] has applied a neural network optimized using a genetic algorithm for predicting Aircraft AC-bearing fault. The dataset used in this study consists of vibratory signals captured from different bearings defects. The features are determined based on spectral density estimation, and prediction outcomes were compared with discriminant analysis classifier.Li et al. [40] introduced an intelligent fault detection system for mechanical rotating systems. The study uses a recurrent neural network (RNN) with fuzzy logic. The

RNN filters the input signal, and then the filtered signal is fed to fuzzy logic to detect faults. Ghorbanian and Gholamrezaei [41] investigated the application of different machine learning mechanisms in the context of analyzing performance compressors. The authors have utilized general regression neural (GRN) and MLP to simulate the performance of these models. The result indicated GRN is less associated with mean error and performed well with the experimental data but limited to interpolationfactor. On the other hand, MLP was evaluated, and the result indicated the most favorable outcome to analyze compressor performance.The work of Aravinth and Sugumaran [42] adopted a statistical feature extraction approach and random forest (RF) classifier to monitor and predict the fault in the AC to avoid regular failure in industrial and domestic applications. In this study,the accelerometer sensory signal is processed via a statistical approach, and RF is applied to detect the type of fault in AC.Fan et al. [43] have considered the case study of vehicle communication and presented their work on predicting AC breakdown using data streamed by the vehicles. The authors have used histogram analysis to model the signal. However, the histogram is a more straightforward approach that can determine the deviation in the signal to some extent. The study of Cui et al. [44] suggested an intelligent model for the early detection of faults in AC. The approach used in this study is based on the construction of an adaptive matrix based on the PCA and backpropagation techniques. This matrix is constructed to store the signals and determine a function of deviation in the signal pattern.

Further, identify early fault signature, a threshold is computed based on the mechanism of the sliding statistical window method. Work towards evaluating trustworthiness and prediction reliability on the AC in the Ammonia Plantis considered by Musyafa et al. [45]. Chen et al. [46] presented an LSTM-oriented approach for classifying compressor breakdown using aggregated sensory data.The performance of the presented model is evaluated using information captured from large heavy-duty vehicles. The authors have formulated a classification task to identify whether a compressor fault will occur within the specified horizon. An LSTM learning model is used to predict, and its performance is evaluated against the RF classifier. The experimental outcome exhibited that RF slightly outperforms LSTM regarding AUC. However, the prediction outcome from LSTM shows stability over time, maintain stability in the trend of healthy faulty classification. Another work carried out in a similar direction by Yang et al. [47] suggested an AC fault classification mechanism using on lifting wavelet approach. Initially, this study has decomposed the vibration signal of the AC wavelet; and further statistical features of decomposition are computed as the AC faults. In the classification process,the probabilistic model-based supervised classifier is employed to predict the fault class. The study outcome suggested that faulty features determined using a wavelet-based approach provide comprehensive fault features that lead to higher accuracy by the supervised classifier.

III. Dataset Description, Exploration, and Visualization

This section describes the process of data creation, its description, and an exploratory analysis and Visualization that decides the line of research for the data presentation for the learning model to overcome the memory constraints of the traditionally applied LSTM based fault classification system for higher accuracy.

A. Data Description

The Department of Electrical engineering of the Indian Institute of Technology (IIT), Kanpur, India having an air compressor of the single-stage reciprocating type. An effort by Dr. N.K Verma and his team to provide an open-source dataset [33][34] on the following specification as in Table I.

In the work of Verma et al. [33], all sensors, as in Fig. 3, records raw data every 5 seconds at the sampling rate of each sensor at 50 kHz and gets stored into the respective eight files as a structure shown in Fig. 3.

A closer analysis of each record shows that it consists of precisely 225 'dat' files in all the folders. Since there are 24 sensors at the sensitive points and an additional one sensor kept at a far distance to record the acoustic data at every 5 seconds, there are nine different timeslots; thus, as per Fig. 4, there are 9 x 25 = 225 files for each fault.

TABLE I. Specifications

SI. No	Specification	Values
1	Air pressure range	0-500Ib/m^2, 0-35 Kg/cm^2
2	Induction Motor	5Hp, 415V, 5Am,50Hz,1440rpm
3	Presure Switch	Type Pr-15, Range 100-213PSI

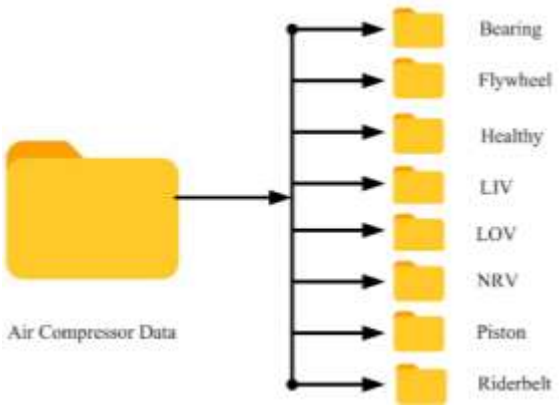


Fig. 3. A File Structure for the Data.

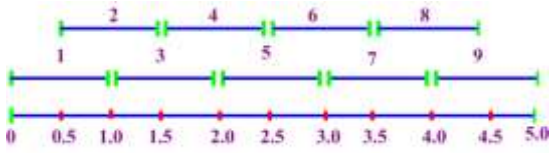


Fig. 4. Time Slot for the Data Records from the Sensors.

TABLE II. CLASSES OF THE FAULT

Sl. No	Fault class	Abbreviation
1	Bearing fault	BF
2	Piston Ring fault	PRF
3	Flywheel fault	FWF
4	Rider Belt fault	RBF
5	Leakage Inlet Value fault	LIVF
6	Leakage Outlet Valve fault	LOVF
7	Non-Return Valve fault	NRVF

The raw data undergoes various pre-processing stages, including filtering to eliminate the undesirable frequency component using FIR filter at 400 Hz cut-off frequency (COF) threshold and low pass filter with COF of 12 kHz to obtain the valuable information.Further, clipping, smoothing, and normalization operations are performed to obtain the pre-processed data. The operation for extracting the name of the fault classes is applied and as in Table II. Moreover, the fault class is categorized as seven faults classes and consists of the normal class.

1) Bearing fault: Bearing fault in the compressor arises when there is malfunctioning in the bearings, which are meant to make the compressor wheels running smoothly. Either bearing may break or may get imbalanced due to wear and tear. Due to this, friction in the machine will increase, and noise will arise.

2) Piston fault: Piston is the major part of the mechanism which converts rotatory motion into linear or vice versa. If there is a fault in the piston RPM of the entire machine may reduce. Moreover, due to this, the full sound of the machine will get less loud.

3) Flywheel: Flywheel is the main storage of kinetic energy in any machine. The main source of rotary motionis its motor or IC engine, which may not provide continuous energy to the machine. Hence if there is a fault in the flywheel due to wear and tear, the wheel spins faster; however, it can store less kinetic energy. Since it spins faster, the frequency of the sound may increase.

4) Leakage in inlet valve: This fault occurs when the inlet valve of the compressor leaks, the pressure in the cylinder also reduces significantly. The noise becomes lessloud since the compressor is no longer working at optimal efficiency. The speed of the piston will reduce, and the frequency of the noise also reduces.

5) Leakage in outlet valve: Contrary to the previous problem, high frequency and loud noises will appear when the leakage is in the outlet valve. This is because the pressure in the cylinder and speed of the piston remains the same, but still, the air escapes from the outlet valve with high pressure. This causes extra noises of various frequencies to appear.

6) NRV fault: NRV refers to a Non-return valve, which means that the valve will close when air tries to flow in the opposite direction. The fault arises when the air starts hitting the NRV valve in the opposite direction, which might be caused due to blockage or damage. In either case, there will be an impulsive load on the valve. This will induce noises of low frequency to appear along with the rest of the noises.

All these faults including normal is encoded as {0,1,2,3,4,5,6,7}. In order to take the recording, typically, faults were inducted into the AC. Fig. 5 illustrates the placement of the 24 acoustic sensors (AS) setup at four typical locations as {top of the piston: 6, NRVF: 6, Opp. of NRV Side:6, Opp. Of FW Side: 6}, where 6 → number of AS.

B. Data Exploration and Visualization

Initially, all the 225 data files stored into the respective directories are read and converted into the 2D vector of size 50,000 x 1, as each file contains reading at every '1' sec at a 50KHz sampling rate. The explicit procedure for this operation is as in algorithm 1. That can be understood using the flow chart in Fig. 6.

Algorithm-1: 2D Vector representation of the processed data

Input: C
Output:P, R
Start
Initialize P, R
for ∀ fault_Dir ∈ C
f_{read}(fault_Dir)
for ∀ file ∈ fault_Dir
Fn←f_{ioin}(fault_Dir, fault, file)
Text←f_{read}(Fn)
CSV←f_{split}(Text : ,)
V←f_{float}(*CSV*)
Append V→*P* &fault→*R*
End
End
End

Top Of Piston | NRV Side | Opp. of NRV Side | Opp.of Flywheel Side

Fig. 5. Position for the Sensitive Position Analysis on the different Sides of the Air Compressor.

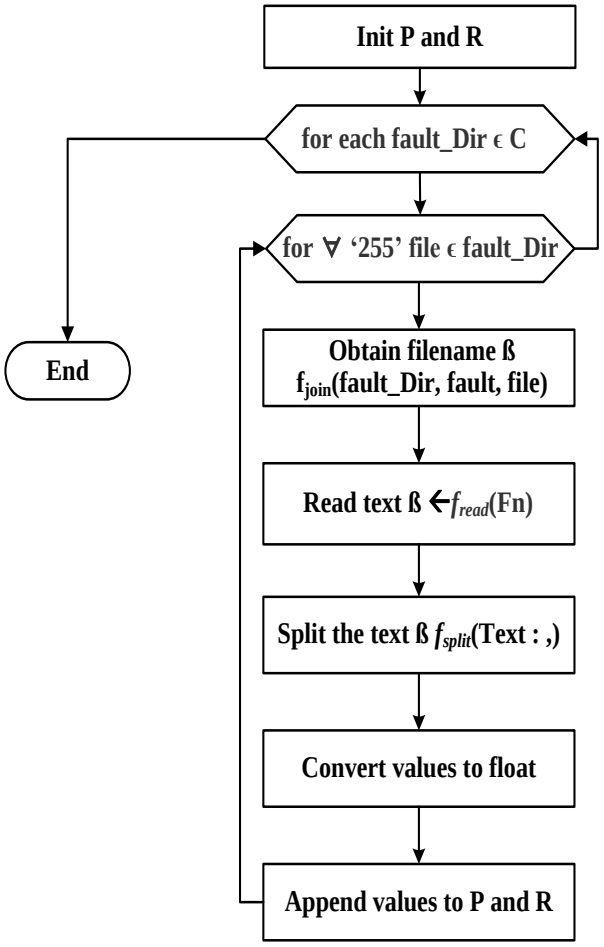


Fig. 6. Process Flow of the 2D Data Representation of the Pre-process Sensor Data of Faults.

In the process of this stage of the data representation, all the data stored into respective folders(fault_Dir) ∈, the main folder (C) are read.Then ∀'255' file ∈ fault_Dir are read to obtain the file names (Fn) of the data by joining the strings: {fault_Dir, fault, and the file name.wav}, however when the 'Fn' is read, it is in the string format that gets converted into the comma-separated string that gets converted into the number types as values(V). Further,the value and the fault are updated to the initialized prediction vector (P) and response (R).

C. Signal Transformation

This section presents the transformation of the time-domain signal into the frequency domain, as demonstrated in Fig. 7.

The time-domain audio signal is transformed into the frequency domain using the numerical expression given in equation 1.

$$x[k] = \sum_{n=0}^{N} x[n] \cdot e^{\frac{-2\pi j}{N} \cdot Kn} \quad (1)$$

Where x[k] refers to the frequency domain using fast Fourier transformation, N denotes samples per second (sampling frequency), $n \in [0, 50000)$, x[n] indicates samples in the time domain, $e^{\frac{-2\pi j}{N} \cdot Kn}$ is the Euler's formula, and coefficient of e, $\frac{-2\pi j}{N} \cdot n$ denotes rotation, and finally, K indicates the amplitude of AS signal at a particular frequency.

1) Time domain analysis: The time-domain signal represents variation in quantity concerning time as a waveform. The advantage of considering AS signals in the time domain are highlighted as follows:

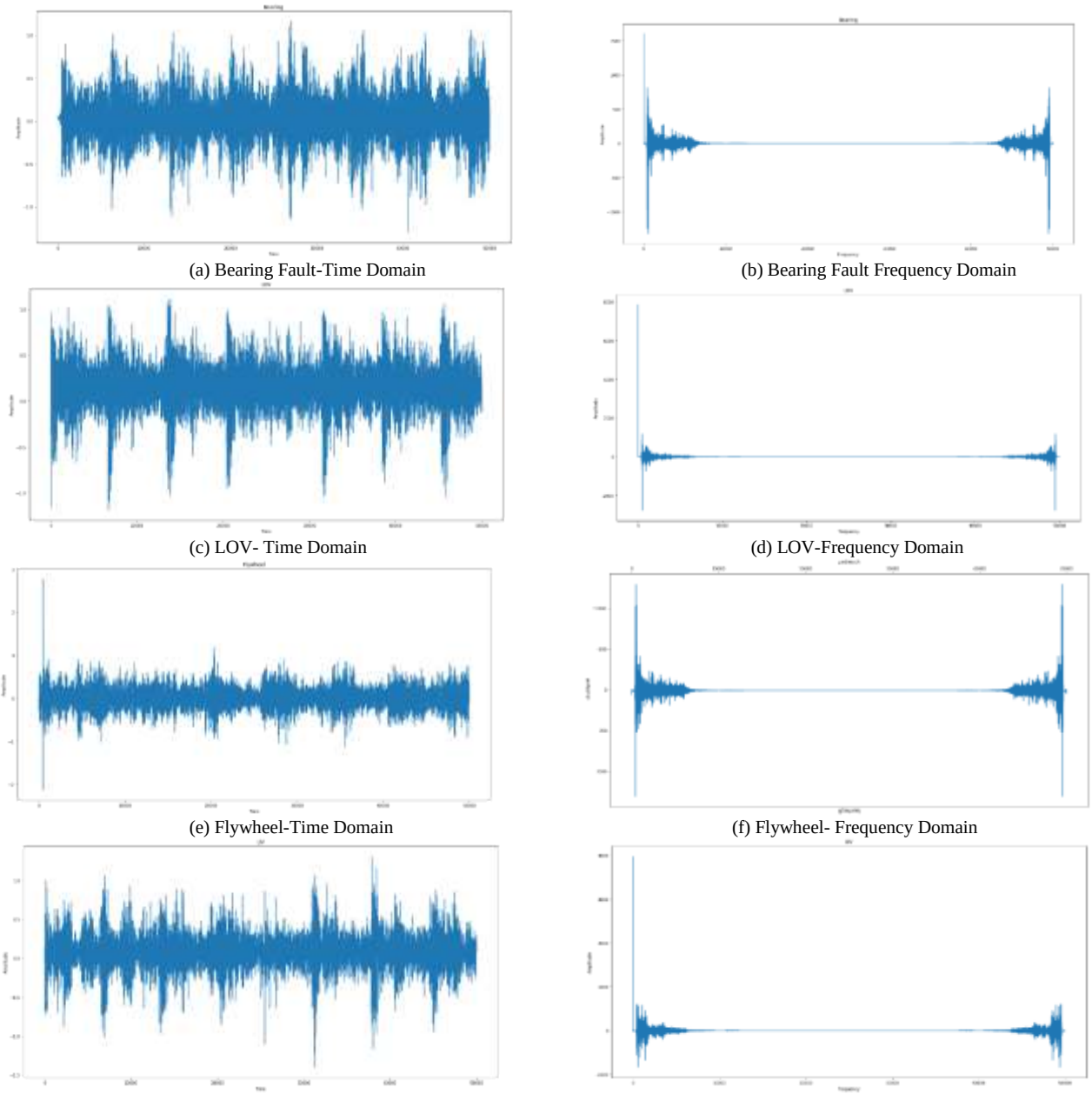
Advantages

- Minor changes in the AS signaling pattern can be represented in the time domain.
- If there is time-sensitive data, a particular noise occurring only in the first few seconds of the machine's starting can be represented in the time domain.
- Phase shifts can be easily recognized in the time domain representation.

However, even considering these advantages time domain may not be a suitable representation of the data since the readings are taken after sometime of the machine being started. In a mechanical system like AC, there are absolutely no changes of occurrences of phase shifts. Phase shifts occur only in electronic systems. In a complex machine-like AC, many types of audio signals are mixed. There may be sound coming from the main cylinder, sound from the flywheel, and minor sounds due to friction between moving parts. The disadvantages of AC signal analysis in the time domain are described as follows:

Disadvantages

- AC signals are captured at different frequencies and often change depending on the sampling frequency of the sensors.
- Time-domain analysis may not be suitable to determine the fault accurately because of a high number of captured signal overlapping.
- The ambient noises are removed in the adopted dataset [BP]. However, there are common sounds recorded by all the sensors, for example, the sound of the motor being captured by ASs since it transmits efficiently through the metal shell of the AC.



(a) Bearing Fault-Time Domain

(b) Bearing Fault Frequency Domain

(c) LOV- Time Domain

(d) LOV-Frequency Domain

(e) Flywheel-Time Domain

(f) Flywheel- Frequency Domain

(g) LIV-Time Domain.

(h) LIV-Frequency Domain

Fig. 7. Represents the Acoustic Signals∀ Fault Classes in a Time Domain and Frequency Domain Such that AS ∀ Fault Classes ∈ Time-Domain = {7a, 7c, 7e, 7g, 7i, 7k,7m,7o}, Whereas AS ∀ Fault Classes ∈ Frequency Domain = {7b, 7d, 7f, 7h, 7j, 7l,7n,7p}.

2) Frequency domain analysis: The AS signals in the frequency domain represent the amplitude of the quantity over various frequencies. The signal in the frequency domain is called a spectrum. There are many advantages of representing AS signal in frequency domain described as follows:

Advantages

- Any frequency domain transformation works as a frequency un-mixer.
- Easier to find out which instrument faults by looking at variations in the natural frequencies in the spectrum. For example, when there is a fault in the bearing, the friction increases and produces high-frequency noise from rubbing the metal pieces. So, the higher frequency noise becomes more dominant when there is a bearing fault present. Hence, such fault can be easily recognized.

Disadvantages

- In the frequency domain representing phase, shifts are quite challenging tasks. However, phase shits are not important in a mechanical system.

Further, a descriptive statistical analysis is performed better to understand the overall data through the summarization

process and generate actionable information from the signal representation data. It provides '225 x 8 = 1800' samples.

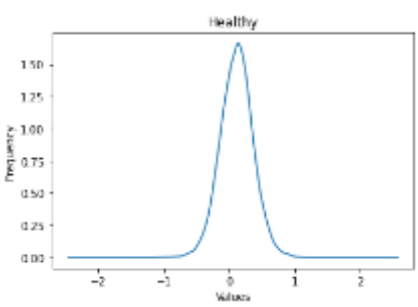
Each has nine statistically significant computations belonging to a set {min, Q1, Q2, Q3, max, count, standard deviation, mean, fault-type}. Table III provides some random samples subset from the complete descriptions.

The count of all the samples is 50,000, indicating that there is no need to work on the cleaning process as there are no missing values in the value point in the sample or data point.

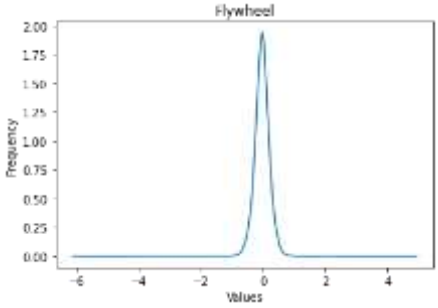
However, a better correlation is analyzed through the histogram of a sample for each fault type as shown in Fig. 8 provides a better visual perception of the data pattern.

TABLE III. A SUBSET OF A SAMPLE OF DESCRIPTIVE STATISTICAL ANALYSIS OF THE DATA

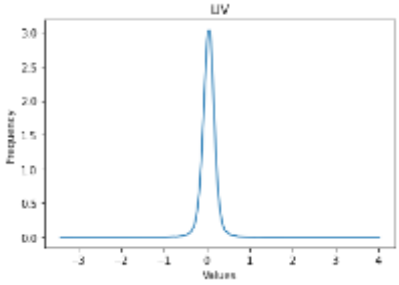
Sample No	count	Min	Q1	Q2	Q3	Max	Std Dev.	Mean	Fault type
0	50,000	-1.5920	-0.049589	0.049434	0.147045	1.3448	0.186775	0.048568	1
1	50,000	-1.2099	-0.082617	0.034085	0.149682	1.2805	0.206034	0.032862	1
2	50,000	-1.1444	-0.070414	0.059485	0.188635	1.1400	0.223090	0.058771	1
3	50,000	-1.1913	-0.095192	0.045985	0.188565	1.1978	0.242442	0.044639	1
4	50,000	-1.1168	-0.025945	0.082214	0.190672	1.2462	0.194634	0.080915	1
…	…	…	…	…	…	…	…	…	…
1795	50,000	-1.3833	-0.105585	0.000489	0.107895	1.3087	0.192076	0.000854	8
1796	50,000	-1.2674	-0.111110	0.002802	0.116820	1.3919	0.206447	0.003726	8
1797	50,000	-1.5291	-0.052908	0.049118	0.155342	1.2516	0.194609	0.049839	8
1798	50,000	-1.2897	-0.166673	-0.074163	0.019368	1.6754	0.167009	-0.074875	8
1799	50,000	-1.4939	-0.075400	0.035780	0.148352	1.5342	0.204002	0.035833	8



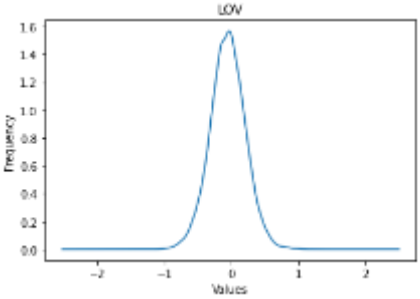
8 (a) Healthy



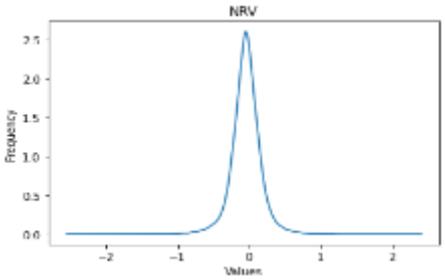
8(b) Flywheel



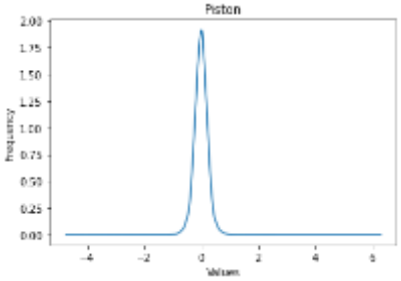
8 (c) LIV



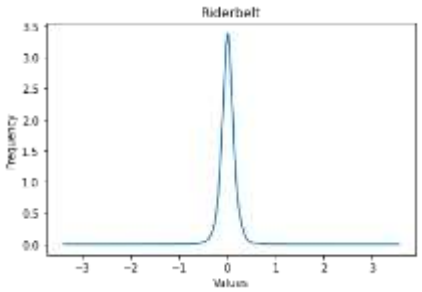
8 (d) LOV



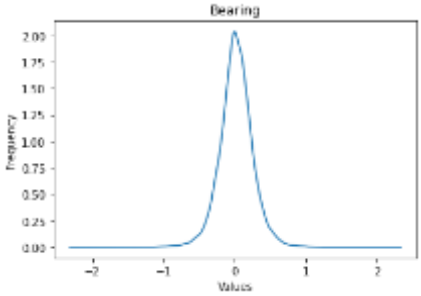
8 (e) NRV



8 (f) PISTON



8 (g) Ridger Belt



8(h) Bearing

Fig. 8. Histogram Plot of a Sample Set of each Category of Fault.

As shown in Fig. 8, the amplitude ranges from a higher value to a lower value depending on the fault type. The histograms for the respective fault types, as in Fig. 8(a) to Fig. 8(h), indicates the repetitions of amplitudes with thecentral tendency of each curve to zero. However, Fig. 8(d) and 8(h) show multiple central tendencies with zero and a little higher peak in Bearing and a lower peak in the case of LOV fault. The detailed observatory description for the rest ofthe distribution is as below:

- *Flywheel:* As it can be observed that the flywheel curve is wider compared to both above histograms. (When we call it wider, observe the x-axis. The curve is landing at -2,+2) due to this, it can be concluded that when there is a flywheel fault, the noise of a particular frequency from the machine gets louder. This is an important indicator.
- LIV: It can be observed here that the noise will become less loud compared to normal operation. This is quite understandable since LIV stands for leakage in the inlet valve. Moreover, due to this, pressure will reduce, and the loudness of the machine will also reduce.
- LOV: Contrary to the previous example, when there is a leakage in the outlet valve (LOV), it will induce another high-frequency noise. Upon closer inspection of the peak, there are two peak points present. The lower one is for the formal operation, and the higher one is for the noise. The air will escape with a much higher velocity from the outlet valve. Moreover, due to this, high-frequency noise is induced.
- NRV: NRV or non-return valve occurs only when the air hits the NRV with the impulsive load. The purpose of NRV is to ensure the unidirectional flow of air. Except for this, the machine is in normal working condition. Hence, this is very similar to a normal operation. However, due to lack of pressure in output, some noises are not present.
- PISTON: In this case, however, the outside piston is malfunctioning. Since both flywheel and piston are external components to the main turbine, this histogram looks very similar to flywheel fault.
- Ridge Belt: Ridge belt is the belt connecting the flywheel to some machine tool or energy converter. If this is at fault similar effect of flywheel fault is produced in the case of acoustics.
- Healthy: In a healthy air compressor, it can be observed that the central tendency of the KDE plot is little towards the positive side of the plot. This means that there is a very low-frequency noise is present when the compressor is working normally. This could be due to the rotation of the wheels and bearings.

IV. Learning Model Design using LSTM

LSTM is a specific Recurrent Neural Network (RNN) class, which is most suitable for predicting time-domain sequences and their long-term dependencies more accurately than ordinary machine learning models.RNN considers that the association amongcellsisformulated as a directed graph. The previous state of the cell may be recurrent, which gives the network the ability to "remember" the information. With this exclusive structure, RNN can make decisions based on previous output value and current Input. However, RNN encountered the issue of exploding andgradient vanishing during the training phase.LSTM has been conceptually designed to address the issue of vanishing and exploding gradients. The LSTM network has a unique structure called a cell (neuron), allowing it to control the flow of information in the network. The elementary unit structure of the LSTM cell is shown in Fig. 9.

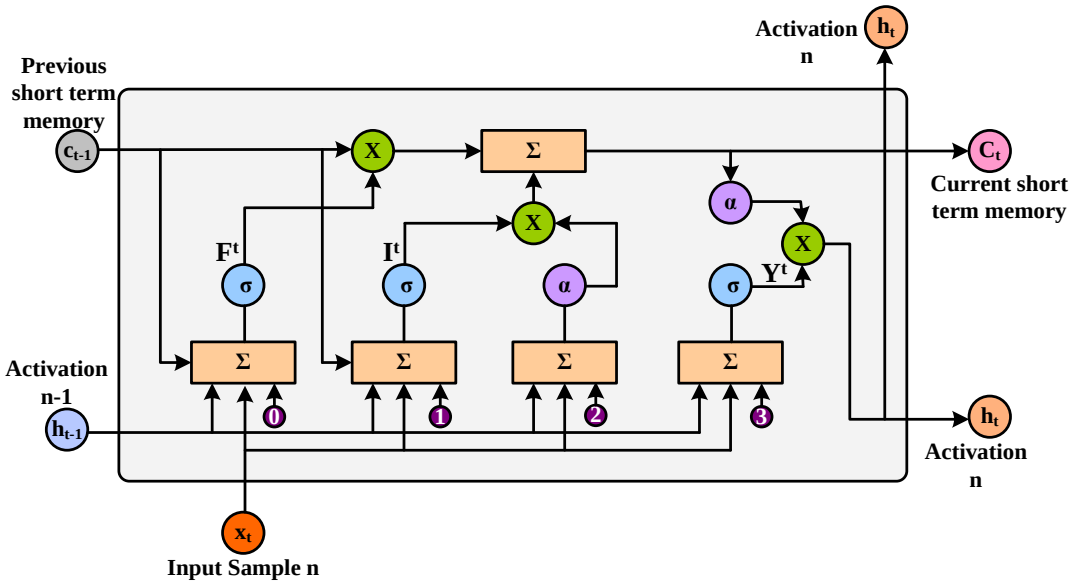


Fig. 9. Basic Structure of LSTM Cell.

In Fig. 9, the basic structure of the LSTM cell is demonstrated that utilizes vector connection by different functions such as sigmoid 'σ'and hyperbolic tangent 'α' with point-by-point addition '∑' and multiplication '⊗' operations. The cell has knodes such as input nodes (x_t) that takes input samples in the form of vector to the LSTM, activation-n (h_t) shows the output of a node, the current short-term memory, or current state of cell (C_t) where both $\{h_t, C_t\} \epsilon\ \mathbb{R}^{k\times 1}$, previous short-term memory (C_{t-1}) indicates the previous state of cell and activation n-1 (h_{t-1}) Shows the output of the previous node. Moreover, to have better control and memorize the flow of information, the LSTM cell utilizes gating mechanisms such as input gate I^t, forget gateF^tAnd the output gateY^t, where each cell gate such that $\{I^t, F^t, Y^t\} \epsilon\ \mathbb{R}^{k\times 1}$. The I^t utilizing x_t and h_{t-1} determines what value to use to decide the value of C_t. The operation of updatingC_t by I^tgate numerically expressed as follows:

$$I^t = \sigma(w_i \otimes [h_{t-1}, x_t] + b^I) \quad (1)$$

$$C_t^\alpha = \alpha(w_C \otimes [h_{t-1}, x_t] + b^C) \quad (2)$$

Where, in equation (1) I^t denotes input gate of the cell at timestep 't' (occurrence of LSTM cell), the variable w_i and b^I are the weights and bias of 'σ' sigmoid operation between I^t and Y^t. In equation (2) C_t^α denotes values of cell state generated by α, the variable w_C and b^C denotes weight and bias of α operation between C_t and h_t.The next F^t gate decides what information from C_{t-1} to be considered to update C_t. The operation of information flow by $F^t$the gate is numerically expressed as follows:

$$F^t = \sigma(w_F \otimes [h_{t-1}, x_t] + b^F) \quad (3)$$

Where, in equation (3) F^t denotes forget gate of the cell at 't', the variable w_F and b^F are the weights and bias of 'σ' sigmoid operation between F^t and I^t. Using equations (1), (2), and (3), the operation of C_t can be numerically expressed as follows:

$$C^t = F^t \otimes (C_{t-1}) + I^t \otimes C_t^\alpha \quad (4)$$

Further, the Y^t gate determines what information in C_t become value of h_t. The operation of information flow by $Y^t$the gate is numerically expressed as follows:

$$Y^t = \sigma(w_Y \otimes [h_{t-1}, x_t] + b^Y) \quad (5)$$

$$h_t = Y^t \otimes \alpha(C_t) \quad (6)$$

Where, in equation (5) Y^t denotes output gate of the cell at 't', the variable w_Y and b^Y are the weights and bias of 'σ' Y^t. In equation (6) h_t denotes the output of LSTM network computed point-by-point multiplication of previous equation (5) and function α with the input argumentC_t. In the proposed work, the task of compressor fault prediction using time-domain AS signals is regarded as a sequence classification problem.Therefore, the proposed study explores implementinga deep learning mechanism, particularly LSTM, for large-scale time-domain AS signals modeling for fault prediction in AC.The proposed architecture learning model for AC fault classification is demonstrated in Fig. 10.

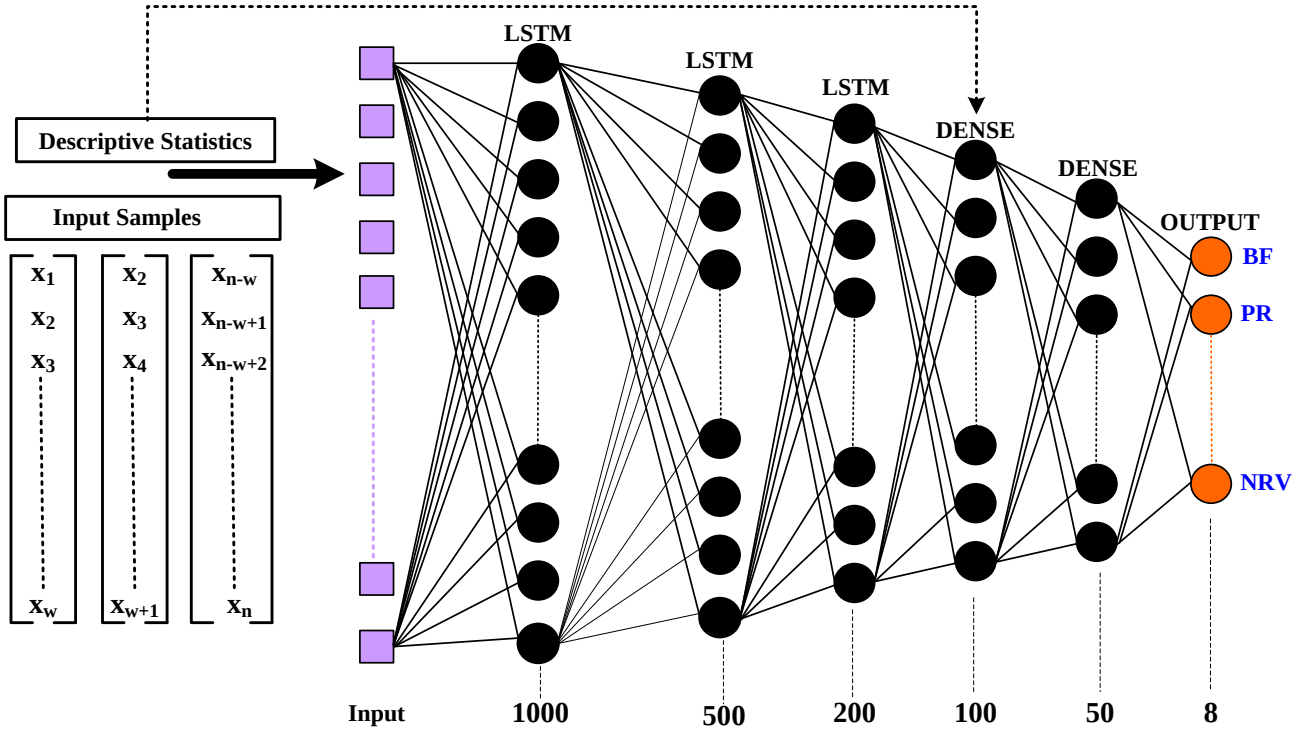


Fig. 10. LSTM Model for AC Fault Classification.

The architecture of the learning model consists of a total three-layer such as i) sequence input layer (I), ii) 5 hidden layers (H), and iii) one output layer (Y), such that learning model $M \in \{I = 1, H = 5, Y = 1\}$. Three hidden layers are configured with 1700 LSTM cells,and the remaining two hidden layers are dense with 158 (regular,deeply connected neurons). The configuration description of the LSTM learning model is presented in Table I. The learning model takes Input as time-domain sensory signals in sequence such that $X_n \in \{x_1, x_2, x_3 \cdots, x_t, \ldots x_L\}$where t denotes timestep,$'L'$length of time-domain signal samples x and maps to output class $C \in \{C_1, C_2, C_3 \cdots C_8\}$ via hidden layer $H \in \{H_1, H_2, H_3 \cdots H_5\}$ using function $H_k = f\{H_{k-1}, x_k\}$, where function*f* refers to the LSTM cell method discussed above that generalizes the long-term dependence between the time domain relationship of the Input X_n signal. The LSTM is trained considering the Input X_n in the form of vectors using the sliding windowing (w) approach,where the Input is a sequence of time-domain signals with length L and w+1 window length. The process of window sliding is illustrated in Fig. 11 with window length (w=1000 AS signal samples).

In the above illustrated, the model takes Input as the first window having the first 1000 AS signal. Then, the next window is selected from the second signal sample of the first window, i.e., from the 2nd sample to the 1001st sample. This process is recurrent until all-time-domain input signals are windowed and fed to the LSTM model. The process flow of AC faults prediction using LSTM is shown in Fig. 12.

The system initially imports the dataset, consistingof 1800 AS signals captured at a 50,000 Hz sampling frequency. To execute the sequence classification tasks, splitting the dataset into two sub-datasets, i.e., training and testing sets. The training dataset is used to train the model, and the testing dataset is used to evaluate the model performance. This allows understanding characteristics of the trained model and provides scope for minimizing the effects of overfitting and underfitting of the model. The dataset split is carried out with a ratio of 80%-20% for training and testing, respectively. Therefore, the training dataset consists of 1440 AS signal samples and a testing dataset composed of 360 AS signaling samples. The study further considers feature selection, where descriptive statistics were analyzed in the time domain. In the training phase, a sliding windowing operation is carried to represent AS signals into the fixed-sized frame, which is further processed via the LSTM layer. Its output is then accumulated with the operation of the dense layer that considers descriptive statistics as Input.

The adoption of a dense layer enhances the generalization of the learning model minimizes the issue of overfitting and underfitting during the learning process. In the proposed LSTM architecture, Adam optimizer is used with a categorical cross-entropy loss function to reduce training loss by adjusting learning attributes such as weights, biases, and learning rate. The configuration details of the LSTM model implemented in this study are mentioned in Table IV. Moreover, Softmax activation is used at the output layer of the LSTM model, as it is designed to address multiclass sequence classification problems (i.e., multiple AC faults). A similar procedure is carried out during model testing, and its effectiveness is assessed regarding the accuracy, precision, recall, and F1-score.

Fig. 13 illustrates the learning curve of the LSTM model training performance over 1000 epochs. As the epochs pass by, the reduction in learning also reducesin the LSTM. Generally, learning rate reduction happens when the error is not reducing for more than 5 epochs. The learning rate is reducing rapidly the error is not converging quickly enough in LSTM. It can beobserved thata sharp exponential decrease in learning rate in the LSTM model. When evaluated with the testing dataset, the model achieved an accuracy rate of 83.12% in AC fault classification. The following section discussed the proposed learning model based on a multilayer perceptron neural network.

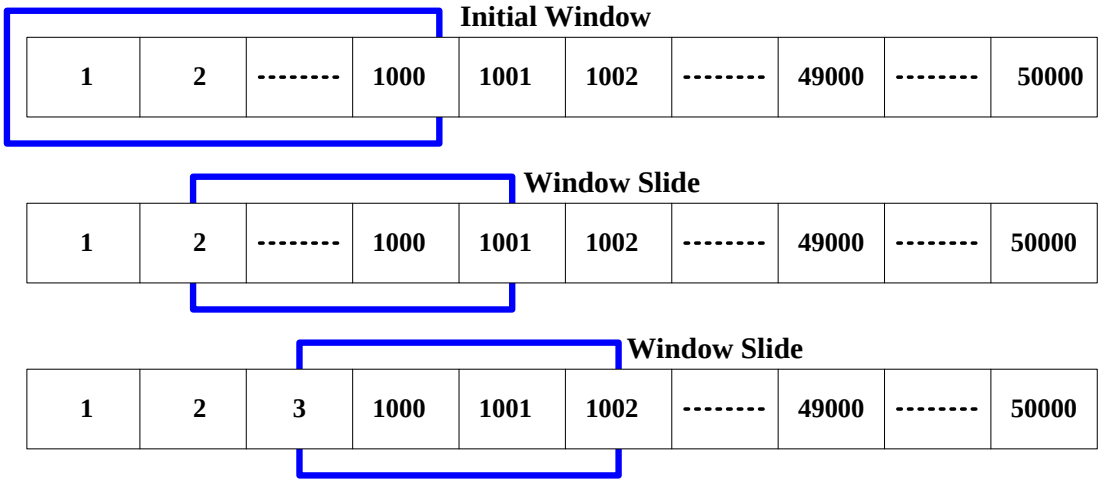


Fig. 11. Process of Sliding Window.

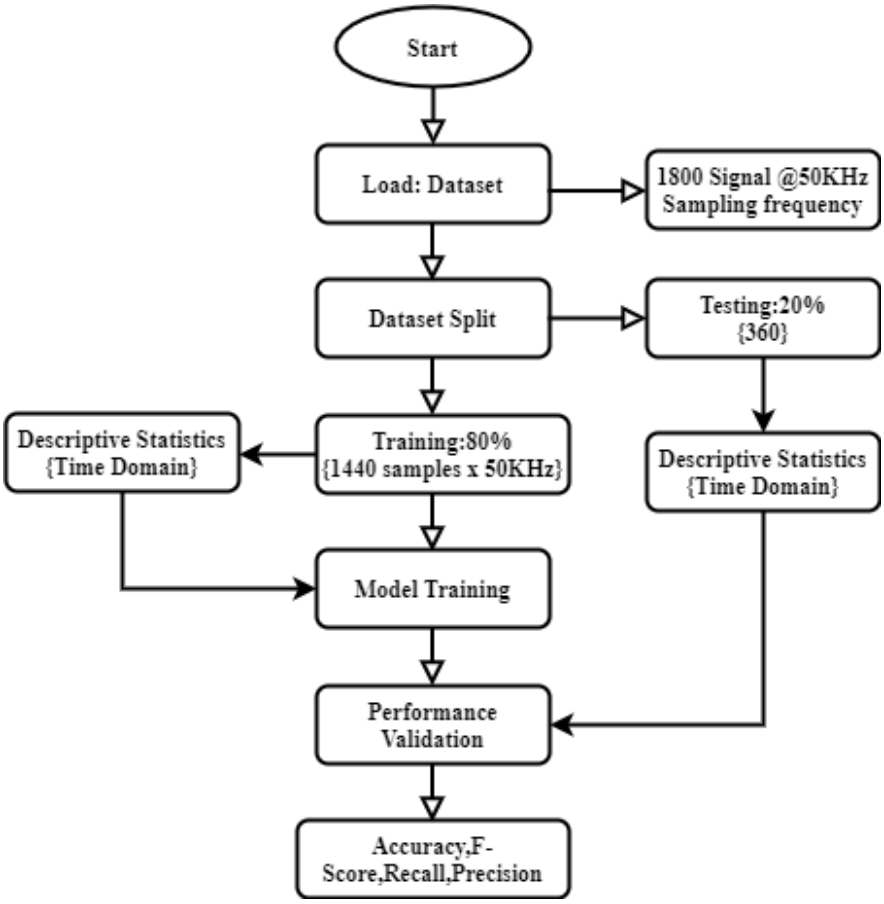


Fig. 12. Illustration of the Process Flow of the LSTM for AC Fault Classification.

TABLE IV. LSTM MODEL CONFIGURATION DETAILS

Input layer	1	Fixed window length of 1000 AS signal
LSTM layer	3	Specific to time-domain AS Signal
Dense Layer	2	Specific to Descriptive Statistics
Optimizer	Adam	Stochastic optimizer
Activation for output	Softmax	Multiclass classification
Loss Function	Categorical cross-entropy	Computesthe difference between two probability distributions
Min Learning rate	10^{-5}	Minimum permitted learning rate for reduce_lr
Max Learning rate	10^{-3}	Initial learning rate
Epochs	1000	The optimal number of epochs for best generalization

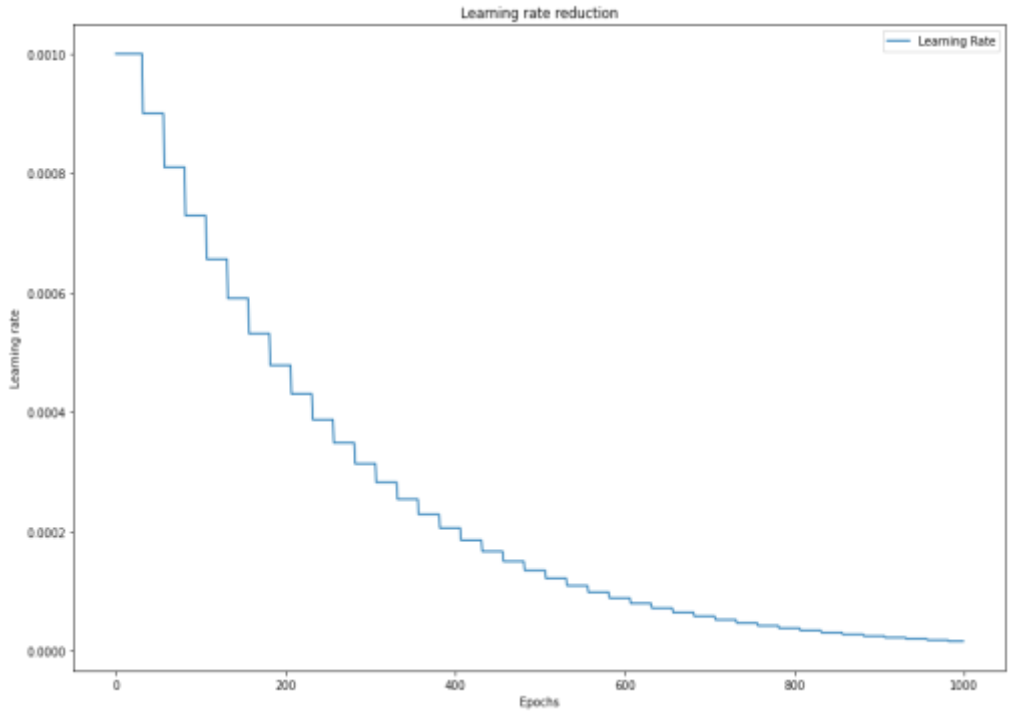


Fig. 13. Learning Curve of LSTM.

V. Proposed Learning Model using Multilayer Layer Perceptron

The neural network consists of an artificial neuron interconnected together by synaptic weights to form a network. Each neuron is modeled by the linear threshold unit, which maps single Input to single output using mathematical operation described as follows:

$$y = f(\sum_{i=1}^{n} W_i xi + \theta) \quad (7)$$

where y denotes the output of the neuron, W_i indicated synaptic weight, $i \in \{1,2,3 \cdots N\}$, xi is the Input $i \in \{1,2,3 \cdots N\}$, and θ indicates a threshold function. A non-linear f (x) can be a sigmoid function or a hyperbolic tangent function.

Amultilayer perceptron (MLP)class of NN. In MLP, the signal travels only in a forward direction; numerically, it can be represented as follows:

$$y = f(A_Z + \theta_y) \quad (8)$$

$$z = f(B_u + \theta_z) \quad (9)$$

where, y is an $m \times 1$ vector refers to the output of the neurons at the output layer; Zis a $p \times 1$ vector, indicates the outputs of neurons at the hidden layer; u is an $n \times 1$ vector, indicates the feature vector of the input signal;θ_y and $\theta_z$arethe threshold vector for the neurons at the output and hidden respectively;the size of θ_y is $m \times 1$ and θ_z is $P \times 1$,A and B are the matrices of size$m \times p$ and $p \times n$, respectively. Both refer to synaptic weights connecting the hidden layer neuron to the output and the Input and hidden layer neurons. The nonlinearity function to be a sigmoid function, i.e.,

$$f(\alpha) = \frac{1}{1+e^{-\alpha}} \quad (10)$$

The unknown parameters A, B,θ_y and θ_z can be determined viareducing an error criterion such that:

$$J = \sum_{i=1}^{P} (d_i - y_i)^2 \quad (11)$$

Where d_i indicates expected outputs which are required to MLP learn and $i \in \{1,2,3 \cdots P\}$, P indicated a total number of instances.

The proposed system implements MLP to classify frequency domain AS signals to predict AC faults because MLP can address complex non-linear problems. It works with both large and small input data and offers quick prediction after training. All these factors are highly significant to the real-time scenario. Although the LSTM is suitable for time sequence data prediction, it is prone to computational overhead and sometimes overfitting problems. The architecture of the proposed AC fault classification system using MLP is shown in Fig. 14.

The MLP can perform better if the AS signal is better exposed to the MLP; as observed in the data exploration, the frequency of audio samples provides a better insight into AS signal representing the faults. Hence, in the proposed study, the frequency domain AS signal is used to train the MLP model to get better accuracy in classification fault classes. The proposed model is composed of two core modules such as i) Adaptive filter and ii) MLP module. Adaptive filter as a frequency domain bandpassfilter, also known as the digital filter,restricts some frequencies from being givenInput to MLP. The functional process of an implemented digital filter with MLP is shown in Fig. 15.

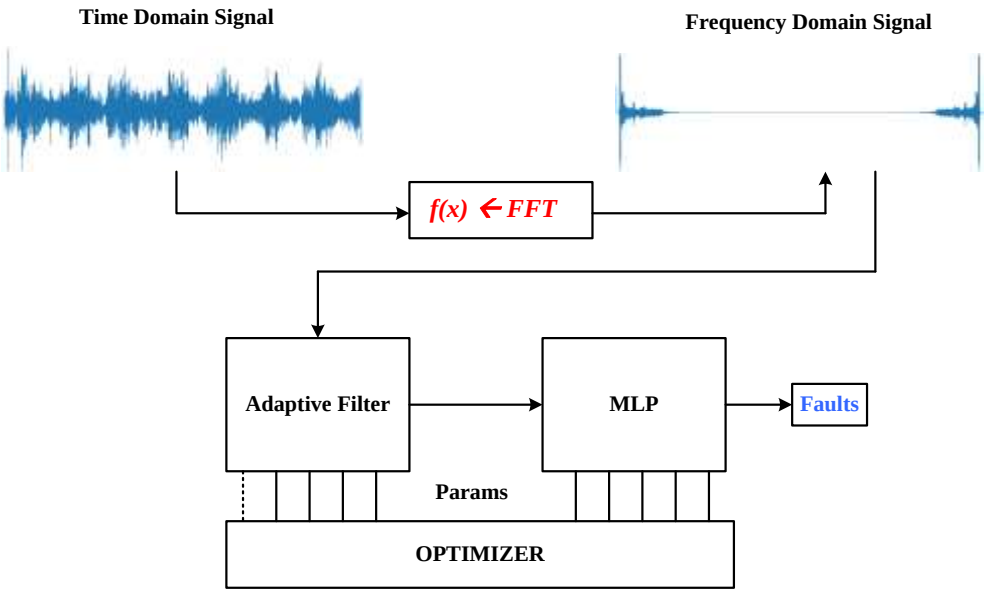


Fig. 14. Schematic Architecture of Proposed Learning Model.

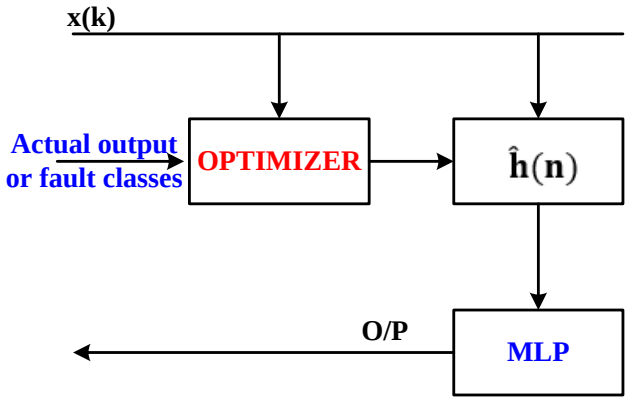


Fig. 15. Synchronization of Frequency Domain Bandpass Digital Filter with MLP.

In the real-time scenario, AS-generated signals often may associate with noisy environmental factors, consisting of recursive or redundant frequencies. Since the proposed learning model takes Input in the frequency domain, it is essential to ensure that the Input AS frequency domain signal does not associate with any irrelevant factors to achieve higher accuracy in the output (O/P). The implementation of adaptive digital filter $\hat{h}$ (n) restricts the irrelevant and redundant frequencies before it is being introduced to the MLP. As a result, reduction in the number of input frequency domain signal samples x(k) reduces computational complexity, thereby reducing feature space complexity by removing irrelevant frequency domain AS signaling features. The processed information by $\hat{h}$ (n) representing a precise input, which providesbetter generalization ability to the MLP in the training phase. The architecture of MLP for AC fault prediction is shown in Fig. 16.

The MLP architecture proposed in the current study consists of the single input layer, with input frequency domain sensory signals such that $x_n \in \{x_1, x_2, x_3 \cdots, x_N\}$ each at 25000 sampling frequency (Nyquist frequency) and mapped to output class $C \in \{C_1, C_2, C_3 \cdots C_8\}$ at output layer via a hidden layer of type dense$D \in \{D_1, D_2, D_3\}$. Since the time domain AS signal is transformed into a frequency domain signal, the theoretical maximum frequency using FFT can be detected always half of the sampling frequency. In the current study, since AS signal sampling frequency is 50 KHz, after transforming to the frequency domain using FFT, Nyquist frequency is 25 KHz. In the proposed MLP architecture, a linear activation function is used at each hidden layer. In the output layer, SoftMax activation is used to deal with the prediction of multiclass AC faults. Therefore, the output layer contains only 8 neurons signifying 8 different outputs. The SoftMax function ensures the sum of all outputs is always 1; hence only the maximum output is selected as the final output with the help of argmax function. A common optimizer is used for both ANN as well as the filter. The optimizer sets the h(n),known as the filter's impulse response. Fig. 17 exhibits the Nyquist frequency sampling process.

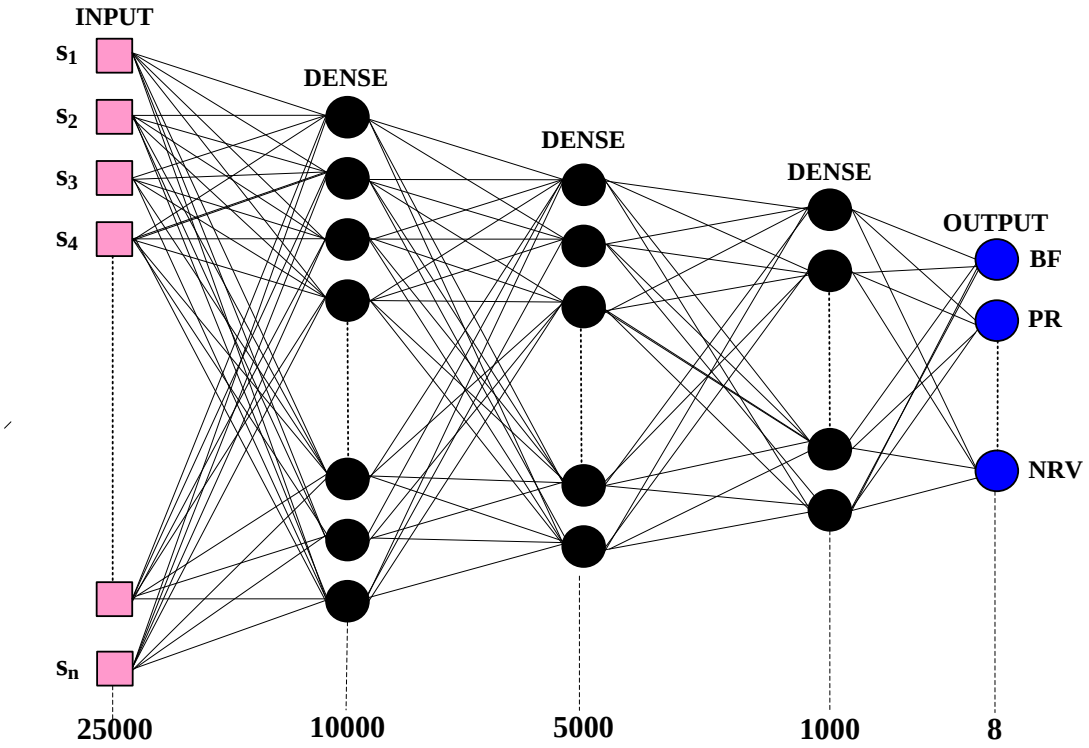


Fig. 16. Architecture of Implemented MLP Model.

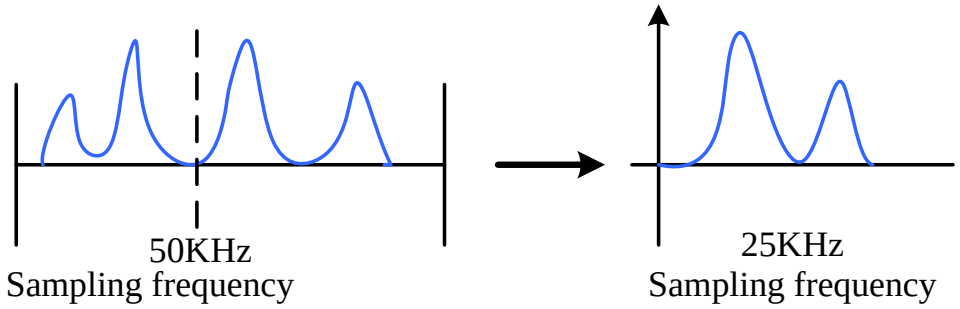


Fig. 17. Illustration of Nyquist Frequency Sampling.

The process flow of AC faults prediction using MLP is shown in Fig. 18, where the system initially imports the dataset,consistingof 1800 AS signals captured at a 50,000 Hz sampling frequency. To execute the sequence classification task, splitting the dataset into two sub-datasets, i.e., training and testing sets. The training dataset is used to train the model, and the testing dataset is used to evaluate the model performance. The input AC signal is converted to the frequency domain using FFT. In this present study, since the sampling frequency is 50 KHz, using the Nyquistmechanism,the sampling frequency of AS signalis computed at 25 KHz, which is the theoretical maximum frequency of FFT. Next, descriptive statistics of frequency domain AS signalsare computed and processed with a domain bandpass adaptive filter. As a result, redundant frequencies from the given Input AS signals being restricted. The filtered AS signal is further fed to MLP, where training is carried out using linear activation functions at each dense layer. After training the model, the testing dataset is used to evaluate the model. Fig. 19 illustrates the learning curve of the MLP model training performance over 1000 epochs.

In Fig. 19, the learning curve trend exhibits a reduction in learning rate is slower compared to LSTM. This indicates that the error reduces rapidly, and effective generalization of MLP. The next section discusses the performance metrics considered for the proposed learning model performance analysis.

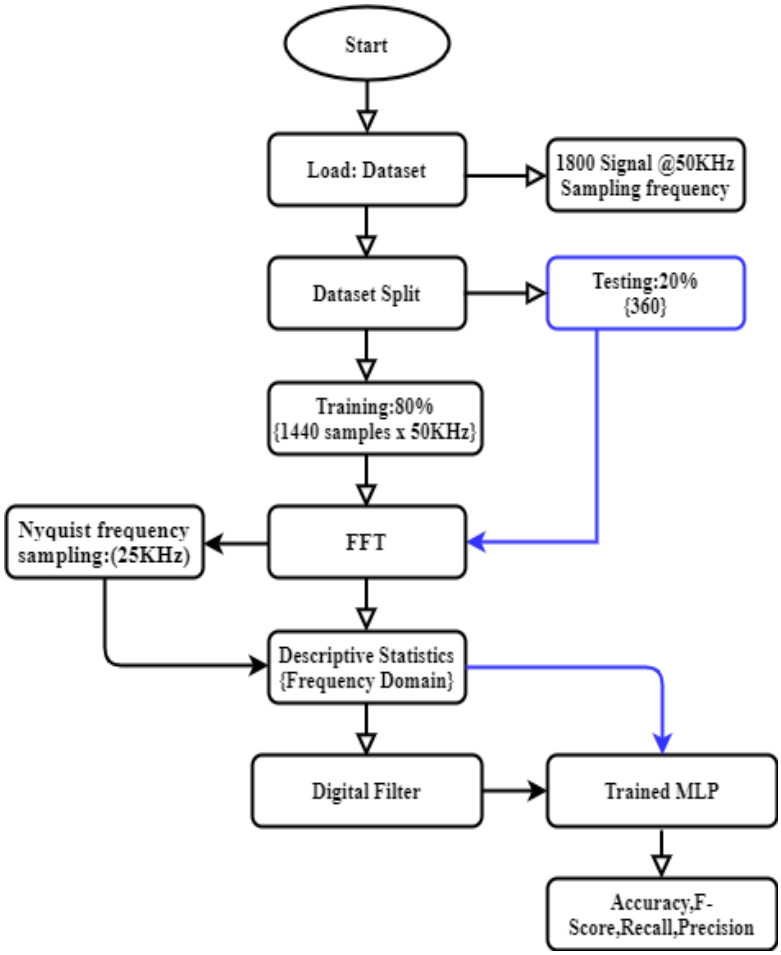


Fig. 18. Illustration of the Process Flow of the MLP for AC Fault Classification.

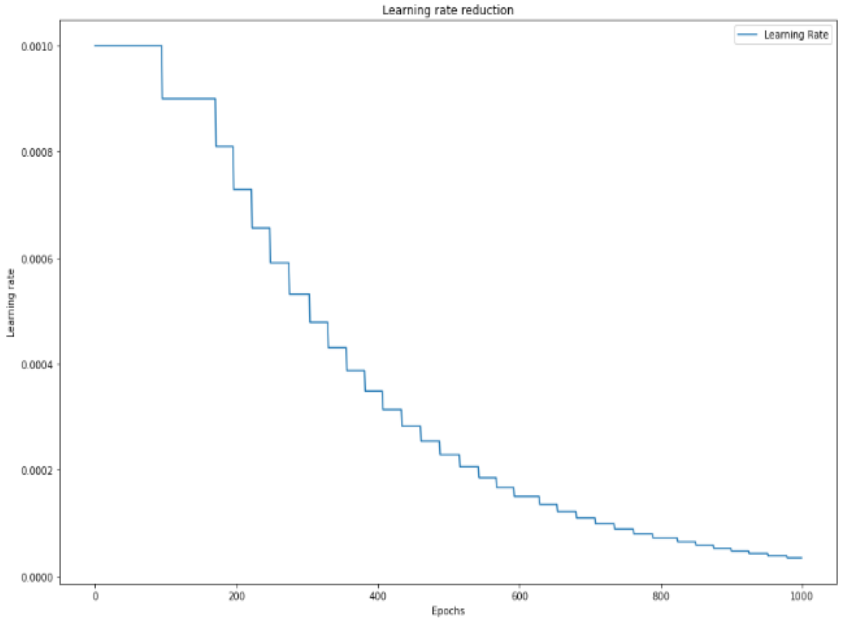


Fig. 19. Learning Curve of MLP.

VI. Performance Metrics

Both learning models' outcome and performance evaluation is carried out concerning multiple performance parameters such as accuracy, precision, recall, and F-1 score.

Accuracy (A): Accuracy can be defined as the ratio of correct predictions over a total number of predictions. Therefore, in the current context of the case study, accuracy can be described as follows:

$$A = \frac{\text{Number of correct predictions of AC faults}}{\text{Total number of AC fault classes}}$$

Precision (P): Precision is the ratio of the number of correct predictions over a total number of predictions made to the current fault class.

$$P = \frac{\text{Total number of accurately identified AC fault classes}}{\text{Total number of detected AC faults classes}}$$

Recall (R): Recall is the ratio of correctly predicted values over the number of expected faults, i.e., the total fault classes present in the test dataset. The lower recall represents the inability of the system to detect the particular class. Like precision, even for recall,the weighted average is taken.

$$R = \frac{\text{Total number of accurately identified AC fault classes}}{\text{Total number of expected AC faults classes}}$$

F1 score: This performance metricdescribes the harmonic mean of precision and recall, which truly represents the system's performance.

$$\text{F1\_Score} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right)$$

VII. Results and Performance Analysis

This section presents the outcome obtained from both implemented learning model and performance analysis and discussesAir Compressor faults classification using acoustic sensor signals. The entire modelling and development of the proposed system are carried out using Python.

A. Analysis of Learning Rate

The comparative analysis concerning the learning rate reduction to access training performance of both LSTM and the proposed learning model.

Fig. 20 presents a comparison of implemented LSTM and Proposed MLP regarding learning rate. It can be analyzed from the learning curve trend that at the beginning, the proposed MLP method takes a little longer time to reduce the learning rate compared to LSTM. However, the proposed MLP model maintains a significant reduction in error during its initial stage of the training process. This indicates that the proposed method has a better optimization in learning and generalization compared to LSTM. It is to be noted that the more the area between the curves better the improvement will be, and if the error in the training phase is not reduced for continuous five epochs, then the learning rate will reduce. The proposed model's learning rate is more, which signifies that the proposed MLP learns faster than the LSTM.

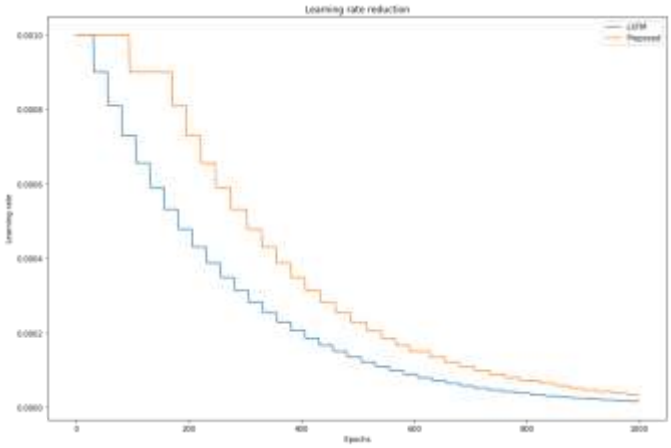


Fig. 20. Comparative Analysis Regarding Learning Rate.

B. Analysis of Classification Performance

The performance analysis is carried out considering multiple evaluation metrics. This is because the system's accuracy is always not a good metric to measure performances, especially when using learning models. The accuracy may not represent the performance of the system completely. If the modelcorrectly predicts fault classes, it indicates a higher accuracy even if it cannot predict negative values. Therefore, performance evaluation of the implemented learning model, i.e., LSTM and the proposed MLP, is carried out considering other two evaluation metrics such as precision and recall rate. However, there is always a trade-off between precision and recall. This is because the precision focuses more on the exactness of the learning models.On the other hand, recall rate focuses more on measuring the completeness of the learning model. For example, suppose the model recognizes the air compressor as faulty. When there are no faults in the air compressor, it is said to have a lower precision rate, which indicates many false positives or biases in the air compression fault predictions. If the model has a higher precision score, then the model is subjected to a low false-positive rate. A low recall rate indicates higher false negatives, and a higher recall rate indicates low false negatives in the prediction result. Since the precision represents the correctness of positive results and recall represents the correctness of negative results, the model should be built to balance both. In order to measure the balance between precision and recall, the F1_score metric is evaluated, which shows the harmonic mean of precision and recall. The harmonic mean is used instead of the regular average since the harmonic mean reduces the effect of extreme values. Table V presents the numerical outcome obtained, followed by the graphical outcome in Fig. 21 to evaluate the implemented learning models.

TABLE V. Numerical Outcome

Performance Metrics	LSTM	Proposed
Accuracy	0.83123	0.913234
Precision	0.863425	0.962342
recall	0.815464	0.892342
F1 score	0.838759	0.926021

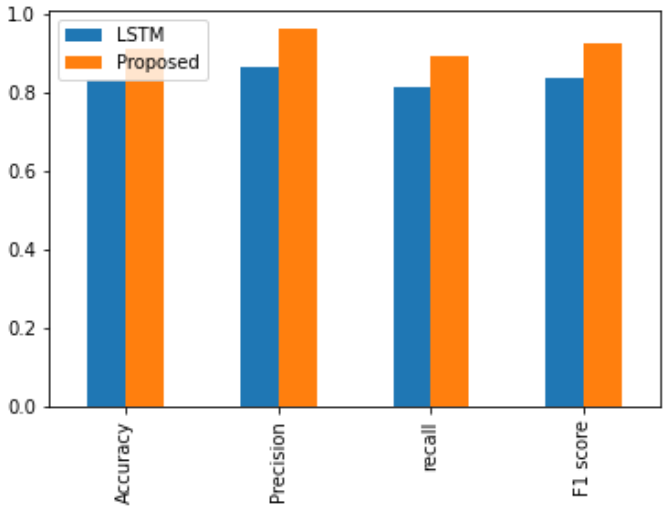


Fig. 21. Comparative Analysis of Classification Outcome.

In Fig. 21, the comparative analysis exhibits that the proposed learning model outperforms LSTM in all performance metrics. The LSTM achieved an accuracy rate of 83.12%, whereas the proposed model achieved a 91.32% accuracy rate. In the case of precision metric, LSTM has scored 86.34%, and the proposed model has attained 96.23 % of the precision rate. The proposed model achieved an 89.23% recall rate, whereas LSTM achieved an 81.54% recall rate. Also, the proposed model has a higher F1_score than LSTM, i.e., 92.60% and 83.87, respectively. Based on the observation, it can be analyzed that the LSTM is biased to an extent towards the faulty results. This indeed came as no surprise since the data contains very few healthy signal samples than faulty signals. The proposed method also has a difference of approximately 7% between precision and recall. However, these are within acceptable limits, which indicates that the proposed model is better at detecting air compressor faults. However, even then, the system is more reliable when both precision and recall are balanced.

VIII. Conclusion

In the proposed work, the study aimed to predict different types of air compressor faults. The analysis was carried out using sensory signals captured from the Acoustic sensors mounted on the Air compressor. The proposed study carried out data visualization and exploratory analysis to characterize the signal features and faults in time and frequency domains. The proposed study is concerned with two aspects of the classification process: the first classification of air compressor faults using the LSTM learning model where the time-domain signal is used as input. On the other hand, the frequency-domain signal is used with a digital filter in the proposed MLP learning model. The result indicated that the proposed learning model outperforms LSTM in accuracy, precision, recall rate, and F1_Score. The outcome shows 83.12% and 91.32% of accuracy achieved by LSTM and MLP, respectively. Also, the learning performance of both models is evaluated. The analysis exhibited that the proposed MLP has less training time compared to LSTM. Therefore, the proposed learning can be claimed to be efficient and suitable for real-time implementation. It has less training time, does not suffer from feature extraction problems, has less memory overhead, and has good generalization ability due to preciseness in the input signal leads to achieving higher accuracy.

References

[1] Sastry YS, Kiros BG, Hailu F, Budarapu PR. Impact analysis of compressor rotor blades of an aircraft engine. Frontiers of Structural and Civil Engineering. 2019 Jun;13(3):505-14.

[2] Zhang L, Cui J, Zhang Y, Yang T, Li J, Gao W. Performance analysis of a compressed air energy storage system integrated into a coal-fired power plant. Energy Conversion and Management. 2020 Dec 1;225:113446.

[3] Razmi AR, Afshar HH, Pourahmadiyan A, Torabi M. Investigation of a combined heat and power (CHP) system based on biomass and compressed air energy storage (CAES). Sustainable Energy Technologies and Assessments. 2021 Aug 1;46:101253.

[4] Fang Y, Lu Y, Roskilly AP, Yu X. A review of compressed air energy systems in vehicle transport. Energy Strategy Reviews. 2021 Jan 1;33:100583.

[5] Li J, Yu T. A new adaptive controller based on distributed deep reinforcement learning for PEMFC air supply system. Energy Reports. 2021 Nov 1;7:1267-79.

[6] Sayed E, Abdalmagid M, Pietrini G, Sa'adeh NM, Callegaro AD, Goldstein C, Emadi A. Review of Electric Machines in More/Hybrid/Turbo Electric Aircraft. IEEE Transactions on Transportation Electrification. 2021 Jun 15.

[7] Cui C, Lin W, Yang Y, Kuang X, Xiao Y. A novel fault measure and early warning system for air compressor. Measurement. 2019 Mar 1;135:593-605.

[8] Aikin AR. The process of effective predictive maintenance. Tribology & Lubrication Technology. 2021 Feb 1;77(2):34-40.

[9] Ye Y, Grossmann IE, Pinto JM, Ramaswamy S. Modeling for reliability optimization of system design and maintenance based on Markov chain theory. Computers & Chemical Engineering. 2019 May 8;124:381-404.

[10] Qu Y, Hou Z. Degradation principle of machines influenced by maintenance. Journal of Intelligent Manufacturing. 2021 Feb 11:1-0.

[11] Lv Q, Yu X, Ma H, Ye J, Wu W, Wang X. Applications of Machine Learning to Reciprocating Compressor Fault Diagnosis: A Review. Processes. 2021 Jun;9(6):909.

[12] Alshorman O, Alshorman A. A review of intelligent methods for condition monitoring and fault diagnosis of stator and rotor faults of induction machines. International Journal of Electrical & Computer Engineering (2088-8708). 2021 Aug 1;11(4).

[13] Połok B, Bilski P. Intelligent Diagnostic system for the rachet mechanism faults detection using acoustic analysis. Measurement. 2021 Jun 24:109637.

[14] Doshi, S., Katoch, A., Suresh, A. et al. A Review on Vibrations in Various Turbomachines such as Fans, Compressors, Turbines and Pumps. J. Vib. Eng. Technol. (2021). https://doi.org/10.1007/s42417-021-00313-x.

[15] Yao J, Liu C, Song K, Zhang X, Jiang D. Fault detection of complex planetary gearbox using acoustic signals. Measurement. 2021 Jun 1;178:109428.

[16] Połok B, Bilski P. Intelligent Diagnostic system for the rachet mechanism faults detection using acoustic analysis. Measurement. 2021 Jun 24:109637.

[17] Ahmed U, Ali F, Jennions I. A review of aircraft auxiliary power unit faults, diagnostics and acoustic measurements. Progress in Aerospace Sciences. 2021 Jul 1;124:100721.

[18] Lou F, Key NL. Reconstructing Compressor Non-Uniform Circumferential Flow Field From Spatially Undersampled Data—Part 1: Methodology and Sensitivity Analysis. Journal of Turbomachinery. 2021 Aug 1;143(8):081002.

[19] Verma NK, Salour A. Faults and Data Acquisition. InIntelligent Condition Based Monitoring 2020 (pp. 7-88). Springer, Singapore.

[20] Verma N.K., Salour A. (2020) Pre-processing. In: Intelligent Condition Based Monitoring. Studies in Systems, Decision and Control, vol 256. Springer, Singapore. https://doi.org/10.1007/978-981-15-0512-6_3.

[21] Wang F, Lin W, Liu Z, Wu S, Qiu X. Pipeline leak detection by using time-domain statistical features. IEEE Sensors Journal. 2017 Aug 15;17(19):6431-42.

[22] Yi Z, Pan N, Guo Y. Mechanical compound faults extraction based on improved frequency domain blind deconvolution algorithm. Mechanical Systems and Signal Processing. 2018 Dec 1;113:180-8.

[23] Sun Z, Zou W, Zheng X. Instability detection of centrifugal compressors by means of acoustic measurements. Aerospace Science and Technology. 2018 Nov 1;82:628-35.

[24] Mondal D, Zhen D, Gu F, Ball AD. Fault diagnosis of reciprocating compressor using empirical mode decomposition-based Teager energy spectrum of airborne acoustic signal. InAdvances in Asset Management and Condition Monitoring 2020 (pp. 939-952). Springer, Cham.

[25] Hammond TT, Curtis MJ, Jacobs LE, Tobler MW, Swaisgood RR, Shier DM. Behavior and detection method influence detection probability of a translocated, endangered amphibian. Animal Conservation. 2020 Sep.

[26] McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction.

[27] Hartwig L, Bestle D. Compressor blade design for stationary gas turbines using dimension reduced surrogate modeling. In2017 IEEE Congress on Evolutionary Computation (CEC) 2017 Jun 5 (pp. 1595-1602). IEEE.

[28] Liang X, Duan F, Mba D, Ian B. Centrifugal Compressor Diagnosis Using Kernel PCA and Fuzzy Clustering. InAsset Intelligence through Integration and Interoperability and Contemporary Vibration Engineering Technologies 2019 (pp. 373-381). Springer, Cham.

[29] Li X, Duan F, Loukopoulos P, Bennett I, Mba D. Canonical variable analysis and long short-term memory for fault diagnosis and performance estimation of a centrifugal compressor. Control Engineering Practice. 2018 Mar 1;72:177-91.

[30] Gong CS, Su CH, Tseng KH. Implementation of machine learning for fault classification on vehicle power transmission system. IEEE Sensors Journal. 2020 Jul 20;20(24):15163-76.

[31] Sobie C, Freitas C, Nicolai M. Simulation-driven machine learning: Bearing fault classification. Mechanical Systems and Signal Processing. 2018 Jan 15;99:403-19.

[32] Wang Y, Wang J, Sun J, Liang E, Wang T. Investigation on recognition method of acoustic emission signal of the compressor valve based on CNN and LSTM network. InE3S Web of Conferences 2021 (Vol. 252, p. 02023). EDP Sciences.

[33] N. K. Verma, R. K. Sevakula, S. Dixit and A. Salour, "Intelligent Condition Based Monitoring Using Acoustic Signals for Air Compressors," in IEEE Transactions on Reliability, vol. 65, no. 1, pp. 291-309, March 2016, doi: 10.1109/TR.2015.2459684.

[34] https://www.iitk.ac.in/iil/datasets/ [Accessed on 22-06-2021].

[35] Schwabacher, M., Goebel, K.: A survey of artificial intelligence for prognostics. In: AAAI Fall Symposium, pp. 107–114.

[36] Omer Faruk Eker, Faith Camci, and Ian K Jennions. 2014. A similarity-based ¨ prognostics approach for remaining useful life prediction. (2014).

[37] RachaKhelif, Simon Malinowski, BrigieChebel-Morello, and Noureddine Zerhouni. RUL prediction based on a new similarity-instance based approach. In IEEE International Symposium on Industrial Electronics' 14.

[38] Zhong, G., Yang, G.: Fault detection for discrete-time switched systems in finitefrequency domain. Circuits Syst. Signal Process. 34(4), 1305–1324.

[39] Rigamonti, M., Baraldi, P., Zio, E.: Echo state network for the remaining useful life prediction of a turbofan engine. In: Third European Conference of the Prognostics and Health Management Society.

[40] Ouadine AY, Mjahed M, Ayad H, El Kari A. Aircraft Air Compressor Bearing Diagnosis Using Discriminant Analysis and Cooperative Genetic Algorithm and Neural Network Approaches. Applied Sciences. 2018 Nov;8(11):2243.

[41] Li, X.; Palazzolo, A.; Wang, Z. Rotating machinery monitoring and fault diagnosis with neural network enhanced fuzzy logic expert system. In Proceedings of the ASME Turbo Expo 2016: Turbomachinery Technical Conference and Exposition, Seoul, South Korea, 13–17 June 2016.

[42] Ghorbanian K, Gholamrezaei M. An artificial neural network approach to compressor performance prediction. Applied Energy. 2009 Jul 1;86(7-8):1210-21.

[43] Aravinth S, Sugumaran V. Air compressor fault diagnosis through statistical feature extraction and random forest classifier. Progress in Industrial Ecology, an International Journal. 2018;12(1-2):192-205.

[44] Fan Y, Nowaczyk S, Antonelo EA. Predicting air compressor failures with echo state networks. InPHM Society European Conference 2016 (Vol. 3, No. 1).

[45] Cui C, Lin W, Yang Y, Kuang X, Xiao Y. A novel fault measure and early warning system for air compressor. Measurement. 2019 Mar 1;135:593-605.

[46] Musyafa A, Kusumawardhani S, Noriyati RD, Justiono H. Evaluation of the Reliability and Prediction Maintenance on the Air Compressor System in Ammonia Plant PT. Petrokimia Gresik. Australian Journal of Basic and Applied Sciences. 2015 May;9(11):853-62.

[47] Chen K, Pashami S, Fan Y, Nowaczyk S. Predicting air compressor failures using long short term memory networks. InEPIA Conference on Artificial Intelligence 2019 Sep 3 (pp. 596-609). Springer, Cham.

[48] Yang WS, Su YX, Chen YP. Air compressor fault diagnosis based on lifting wavelet transform and probabilistic neural network. InIOP Conference Series: Materials Science and Engineering 2019 Oct 1 (Vol. 657, No. 1, p. 012053). IOP Publishing.

Organisational Information Security Management Maturity Model

Mazlina Zammani[1]
National Cyber Security Agency
National Security Council
Jalan Impact, 63000, Cyberjaya, Malaysia

Rozilawati Razali[2], Dalbir Singh[3]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
43600, Bangi, Selangor, Malaysia

***Abstract*—Information Security Management (ISM) is a systematic initiative in managing the organisation's information security. ISM can also be defined as a strategic approach to addressing information security (IS) risks, breaches, and incidents that could threaten the confidentiality, integrity, and availability of information. Although organisations have complied with ISM requirements, security incidents are still afflicting numerous organisations. This issue shows that the current implementation of ISM is still ineffective. The ineffective ISM implementation illustrates the low maturity level. To achieve a higher level of maturity, organisations should always evaluate their ISM practices. Several maturity models have been developed by international organisations, consultants, and researchers to assist organisations in assessing their ISM practices. However, the current models do not evaluate ISM practices holistically. The measurement dimensions in current models are more focused on assessing certain factors only. This caused the maturity assessment to be not executed comprehensively. Therefore, this study aims to address this shortcoming by proposing a comprehensive maturity assessment model that takes into account ISM success factors to evaluate the effectiveness of the implementation. This study adopted a mixed-method approach, which comprises qualitative and quantitative studies to strengthen the research finding. The qualitative study analyses the existing literature and conducts interviews with nine industry practitioners and six experts while the quantitative study involves a questionnaire survey. The data obtained from the qualitative study were analysed using content analysis while the quantitative data employed statistics analysis. The study identified fourteen success factors and fifty-seven maturity dimensions, which each contains five maturity levels. The proposed model was evaluated through experts' reviews to ensure its accuracy and suitability. The evaluation shows that the model can identify the ISM maturity level systematically and comprehensively. This model will ultimately help the organisations to improve the weaknesses in the implementations thus diminishing security incidents.**

Keywords—*Information security*; *information security management; maturity models; information security management maturity model*

I. INTRODUCTION

Now-a-days, organisations' reliance on Information and Communication Technology (ICT) has increased severely due to the rapid development of technology [1],[2],[3],[4],[5]. ICT plays an imperative role in organisations daily operations to ensure the smoothness of the services [6],[7]. In line with the increasing use of ICT in daily operations, organisational information is extremely exposed to security threats and risks [8], [9], [10].

Various efforts have been done to ensure the information is protected. One of the efforts is establishing Information Security Management (ISM). ISM is a strategic approach to addressing information security risks and incidents that could threaten the confidentiality, integrity, and availability of information [10],[11],[12],[13]. However, security incidents endure occurring in organisations [14],[15]. For example, in October 2020, hackers targeted government agencies and telecommunications operators in Iraq, Kuwait, Turkey, and the UAE as part of a cyber espionage campaign [16]. In the latest statistical report released by the National Cyber Coordination and Command Centre, National Cyber Security Agency (NACSA) stated that a total of 4,194 security incidents against public and private organisations were reported in 2020 [17]. This issue shows that the current implementation of ISM is still ineffective [14]. The ineffective ISM implementation illustrates the low maturity level.

Although organisations have complied with ISM requirements set by the industry standards, there is a lack of objective mechanisms to gauge the maturity of the implementation [18]. Even though there are attempts on ISM maturity models [19],[20],[21],[22], they mainly appear as abstract concepts. The current maturity models are typically process-oriented, focusing on measuring security activities and technology aspects without giving much attention to the people aspect, which also contributes to the effectiveness of the ISM implementation [23]. This caused the maturity assessment not executed comprehensively. Thus, the maturity of ISM implementation remains low.

A comprehensive maturity model should consider all aspects in ISM and should not limit to certain aspects only. This study aims to fulfil these needs by proposing a holistic maturity model that considers ISM success factors from four major aspects; People, Process, Organisational Document, and Technology to measure the implementation's effectiveness.

This paper is organised as follows. Section II discussed a review of ISM success factors and the current maturity models. Section III provides the methodology used in this study. Section IV presents the findings and lastly, Section IV summarises the findings.

II. BACKGROUND

A. ISM Success Factors

ISM provides a strategic direction for implementing security processes and activities to assure security objectives are met, consistent risk management, and effective use of information resources [11],[24]. ISM is likewise a multi-disciplinary discipline that should be given due attention to ensuring an appropriate and secure environment in protecting organisational information [25]. Previous studies have indicated that the success of ISM implementation depends on technical and non-technical factors. Those factors are organised into four aspects: People, Organisational Document, Process, and Technology as listed in Table I.

The people aspect consists of individuals or parties directly involved in the ISM. The organisational document refers to strategic and operational documents that need to be developed and adhered to during ISM implementation. Meanwhile, the process aspects consist of ISM key activities and finally, the technology aspect comprises the use of ICT Infrastructure to support the ISM operations. A comprehensive explanation of the factors and their elements can be found in [26].

B. ISM Maturity

ISM maturity guarantees the successful management of information security [27]. A maturity model is a staged structure where particular security aspects are measured, with the postulation that organisations develop and enhance their ISM implementation from the lowest level to the highest level [27],[28]. Thus far, industries and researchers have developed a few maturity models to assist the organisation in measuring the level of ISM implementation [12],[29].

Control Objectives for Information and Related Technology version 4.1 (COBIT 4.1) is widely used for IT governance [21]. It was developed by IT Governance Institute (ITGI) in the year 2007. This model helps measure an organisation's Information Technology (IT) processes, define a designated maturity level, and improve the process to achieve the preferred maturity level [30]. COBIT 4.1 has six maturity levels, which are from maturity level 0 to maturity level 5.

Another maturity model is Cybersecurity Capability Maturity Model (CMM), developed by Global Cyber Security Capacity Centre in 2014. This model was later revised and improved in 2016 and with a new name Cybersecurity Capability Maturity Model for Nations (CMM). The model allows the organisation to self-assess its current cybersecurity capacity [31]. Conversely, the Open Information Security Management Maturity Model (O-ISM3) by The Open Group assesses maturity based on management processes in four components; general, strategic, tactical, and operational [32]. O-ISM3 has five maturity levels, which look for evidence of the processes in those four components.

Many researchers have adopted the above models in their research work. For example, a study presents a cyclical maturity evaluation model [56] where the maturity level is adopted from COBIT 4.1. The model is based on ISO/IEC 27002 security controls where each implementation of the controls will be assessed. The model outlines eight steps to be followed throughout the assessment. A different researcher proposes a model for measuring ISM performance [46]. The proposed model evaluates the performance based on critical factors, namely, human, processes, risk assessment, and technology. The model contains three maturity levels; basic, intermediate, and advance.

TABLE I. ISM SUCCESS FACTORS

Aspects	*ISM Success Factors*	*Sources*
People	Top Management • knowledge • leadership • commitment	[11],[25],[26],[33],[34],[35] [36],[37],[38],[39],[40],[41] [42],[43],[44]
	IS Coordinator Team • knowledge • commitment • communication skill	[15],[26]
	ISM Team • knowledge • commitment • technical skills • willingness • cooperation	[26],[33],[36],[40],[42],[43],[45]
	IS Audit Team • knowledge • auditing skills • commitment • cooperation • communication skills	[26],[37],[38],[42],[43]
	Employees • awareness • compliance • motivation	[5],[26],[35],[36],[37],[38] [39],[45]
	Third Parties • awareness • compliance	[26],[38],[42],[43],[46]
Organisational Document	IS Policy • clear • comprehensive • communicated • reviewed	[5],[25],[26],[33],[34],[35], [36],[37],[38],[39],[41],[42], [43],[45],[47],[48]
	IS Procedures • clear • complete • communicated • reviewed	[26],[36],[37],[49]
Process	Resource Planning • financial resources • human resource	[26],[33],[34],[35],[38],[42],[43],[45],[50]
	Competency Development Awareness • awareness programs • training programs	[26],[33],[34],[35],[37],[38],[39],[42],[43],[45],[48]
	Risk Management • risk assessment • risk treatment	[25],[26],[35],[36],[37],[38], [41],[42],[43],[45],[48],[51]
	Business Continuity Management • plan • simulation	[26],[37],[38],[41],[49],[52]
	IS Audit • audit program • audit finding & reporting • follow-up audit	[26],[36],[37],[38],[42],[43], [53]
Technology	IT Infrastructure • software • hardware	[5],[26],[36],[38],[42],[43], [45],[50],[54],[55]

On the other hand, a maturity model developed by [57] aims to assess the organisation's ability to meet security objectives. The model defines the process of managing, measuring, and controlling security based on four aspects; governance, security management, system architecture, and service management. Each aspect has its indicators [12]. This model has five levels of compliance which starting from non-compliance to full compliance.

The comparison of the mentioned models is summarised in Table II. Table II shows several ISM success factors are being considered as the maturity dimensions in the existing model. However, the existing models are typically process-oriented which focus more on the process and technology factors and have less emphasis on the people factors. This causes the implementation of ISM is evaluated less comprehensively. People factors play a significant role in ISM [58]; thus, need to be emphasized as well [59]. Therefore, a holistic maturity model is required by incorporating all ISM success factors and their elements to ensure the effectiveness of the ISM implementation.

TABLE II. COMPARISON OF MATURITY MODELS

Model/ Basis of comparisons		COBIT 4.1 [21]	CMM [31]	O-ISM3 [32]	Cyclical evaluation model [56]	IS Assessment Model [46]	IS Maturity Model [57]
The objective of the model		Measure the current maturity of an organisation's Information Technology (IT) processes	Measure the current cybersecurity capacity	Measure ISM maturity based on management processes in four aspects; general, strategic, tactical, and operational.	As a means to measure the current situation of IS management based on ISO/IEC 27002 security controls.	Assessing information security implementation levels in organisations.	Assessing the ability of the organisation in meeting security objectives
Scope of coverage		34 processes	5 dimensions	45 processes	133 controls	4 factors / aspects/ domains	4 factors / aspects / domains
Maturity levels		Six levels ranking of 0-5	Five levels ranking of 1-5	Five levels ranking of 1-5	Six levels ranking of 0-5	Three levels ranking of 1-3	Five levels ranking of 1-5
ISM success factors involved in assessment							
People	Top Management	√			√	√	√
	IS Coordinator Team						
	ISM Team	√	√		√		
	IS Audit Team						
	Employees	√				√	
	Third Parties						√
Organisational Document	IS Policy	√	√	√	√	√	
	IS Procedures	√	√	√	√		
Process	Resource Planning	√	√	√	√	√	√
	Competency Development & Awareness	√	√	√			√
	Risk Management	√	√	√		√	
	Business Continuity Management		√	√	√		
	IS Audit	√	√	√			√
Technology	IT Infrastructure	√	√	√		√	

III. Methodology

This study adopts the mixed-method approach, which comprises both qualitative and quantitative data collection and analysis. This approach involves four main phases: theoretical, empirical, model development, and model validation. Fig. 1 illustrates the research design.

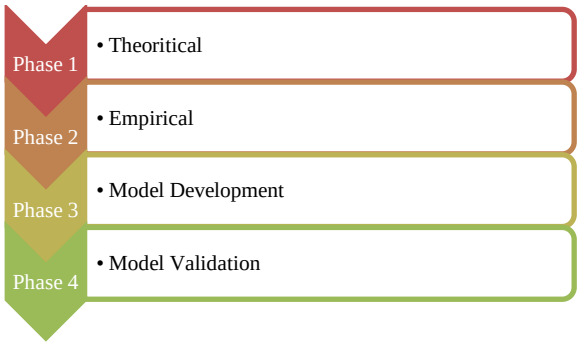


Fig. 1. Research Design.

A. Phase 1: Theoretical

The theoretical study reviewed published and unpublished documents in multiple online databases such as ACM Digital Library, Web of Science, Science Direct, Google Scholar, Pro-quest, IEEE Explorer, Mendeley and CiteSeer to identify the ISM success factors and ISM maturity models. The selected documents were then analysed qualitatively using content analysis. The preliminary findings of this study have been reported in [44].

B. Phase2: Empirical

The empirical study is to verify the success factors and identify each success factor's maturity dimension and levels. As it involves various aspects, it is thus divided into three parts:

- Empirical I: The purpose of Empirical I is to verify the ISM success factors derived from the theoretical study and discover other relevant factors from practitioners' views. This study used semi-structured interviews. A series of individual and focus group interviews with experienced ISM practitioners was conducted. The findings of this study have been reported in [26].
- Empirical II: The purpose of Empirical II is to confirm and refine the findings of Empirical I through a large-scale survey. A total of 400 questionnaires were sent to respondents in public and private agencies. The data collected from the survey were analysed using Statistical Analysis. The findings of this empirical II have been reported in [60].
- Empirical III: A series of interviews with six experts were conducted to identify the ISM maturity dimensions and levels. The selection of experts was based on their experience, knowledge, and expertise in ISM. Contents analysis technique was used to analyse the data.

C. Phase 3: Model Development

The ISM maturity model was developed using the findings from Empirical I, II, and III. The identified success factors, dimensions, and levels were used as the components in the maturity model.

The development of this maturity model is guided by the International Standards ISO / IEC 33004: 2015 Information technology - Process assessment - Requirements for process reference, process assessment and maturity models [61]. In addition, the measurement theory of [62] and [63], which introduced the ordinal scale, was also used as a basis in the development of this ISM maturity model.

D. Phase 4: Model Validation

This phase evaluates the accuracy of the proposed model through expert review. A series of interviews with three experts were conducted to evaluate the accuracy and suitability of the proposed model. Based on the review, the proposed model was improved.

IV. Result and Finding

Based on the experts reviewed, the final Organisational ISM Maturity Model has 4 aspects, 14 factors, 42 elements, and 57 maturity dimensions. The 14 factors are grouped under four main aspects namely People, Organisational Document, Process and Technology. Each factor has its own elements. Each element has specific dimensions. Each dimension has five levels of maturity; maturity level 1 to maturity level 5 where Level 1 is the lowest level of maturity while Level 5 is the highest level of maturity. The finalised Organisational ISM maturity model is shown in Table III.

This study has produced a comprehensive model of measuring organisational ISM maturity. In contrast to the existing model, this Organisational ISM Maturity Model contains factors from process and technology aspects and contains factors from non-technical aspects, namely People and Organisational Document. Every identified factor was then sorted according to its categories and subsequently determined its maturity dimensions. Based on the arrangement of categories and factors generated, this study helps the organisations to self-assessing the maturity level of their ISM implementation systematically. Through the assessment conducted, the organisation can identify their ISM maturity level while further improving the implementation of their ISM.

TABLE III. ORGANISATIONAL INFORMATION SECURITY MANAGEMENT MATURITY MODEL

Aspects	Factors	Elements	Maturity Dimensions	Level 1	Level 2	Level 3	Level 4	Level 5
People	Top Management	Leadership	Personnel/unit involvement in ISM	ISM implementation involves only the ICT unit.	ISM implementation involves ICT unit and process owners.	ISM implementation involves the ICT unit, process owners, and administrative unit.	ISM implementation involves the ICT unit, process owner, administrative unit, and responsible units.	ISM implementation involves the ICT unit, process owners, administrative unit, responsible units, and stakeholders.
		Knowledge	The percentage of understanding the objectives and security issues.	Less than 25% of objectives and security issues are understood.	At least 25% of the objectives and security issues are understood.	At least 50% of the objectives and security issues are understood.	At least 75% of the objectives and security issues are understood.	100% security objectives and issues are understood.
		Commitment	The response rate on the ISM issue.	The response to the ISM issues is very slow.	The response to the ISM issues is slow.	The response to the ISM issues is fairly fast.	The response to the ISM issues is fast.	The response to the ISM issues is very fast.
	IS Coordinator Team.	Knowledge	The percentage of IS Coordinator Team members understand the needs, governance, and processes of ISM.	Less than 25% of IS Coordinator Team members understand the needs, governance, and processes of ISM.	At least 25% of the IS Coordinator Team members understand the needs, governance, and processes of ISM.	At least 50% of the IS Coordinator Team members understand the needs, governance, and processes of ISM.	At least 75% of the IS Coordinator Team members understand the needs, governance, and processes of ISM.	100% of the IS Coordinator Team members understand the needs, governance, and processes of ISM.
		Commitment	The percentage of the ISM planning schedule is achieved.	Less than 25% of the ISM planning schedule is achieved.	At least 25% of the ISM planning schedule is achieved	At least 50% of the ISM planning schedule is achieved.	At least 75% of the ISM planning schedule is achieved.	100% of the ISM planning schedule is achieved.
		Communication Skills	The clarity of the information presented.	Very unclear.	Unclear.	Quite clear.	Clear.	Very clear.
			The attitude of IS Coordinator Team members when communicating	Being not open and not persuasive.	Being less open and less persuasive.	Being a little open and a little persuasive.	Being open and persuasive.	Being very open and very persuasive.
	ISM Team	Knowledge	The percentage of ISM team members are knowledgeable in IS domain.	Less than 25% of ISM team members are knowledgeable in IS domain.	At least 25% of ISM team members are knowledgeable in IS domain.	At least 50% of ISM team members are knowledgeable in IS domain.	At least 75% of ISM team members are knowledgeable in IS domain.	100% of ISM team members are knowledgeable in IS domain.
		Technical Skills	The average duration of ISM team members' involvement in implementing IS operations.	Less than 1 year.	Between 1 - 2 years.	Between 2 - 3 years.	Between 3 - 4 years.	Over 4 years.
			The capability of ISM team members to complete IS operations.	Unable to complete IS operations at a specific time without support from consultants.	Slightly capable to complete IS operations at specific times without support from consultants.	Moderately capable to complete IS operations at specific times without support from consultants.	Capable to complete IS operations at specific times without support from consultants.	Very capable to complete IS operations at specific times without support from consultants.
		Commitment	The percentage of ISM team members committed to implementing IS operations.	Less than 25% of ISM team members committed to implementing IS operations.	At least 25% of ISM team members committed to implementing IS operations.	At least 50% of ISM team members committed to implementing IS operations.	At least 75% of ISM team members committed to implementing IS operations.	100% of ISM team members committed to implementing IS operations.

			The percentage of ISM team members follow security procedures.	Less than 25% of ISM team members follow the security procedures.	At least 25% of ISM team members follow the security procedures.	At least 50% of ISM team members follow the security procedures.	At least 75% of ISM team members follow the security procedures.	100% of ISM team members follow the security procedures.
		Willingness	The percentage of ISM team members willing to accept and implement changes.	Less than 25% of ISM team members willing to accept and implement changes.	At least 25% of ISM team members willing to accept and implement changes.	At least 50% of ISM team members willing to accept and implement changes.	At least 75% of ISM team members willing to accept and implement changes.	100% of ISM team members willing to accept and implement changes.
		Cooperation	Level of understanding between ISM team members to achieve IS objectives.	There is no understanding to achieve IS objectives.	Lack of understanding to achieve IS objectives.	Quite understanding to achieve IS objectives.	Understanding to achieve IS objectives.	Very understanding to achieve IS objectives.
	IS Audit Team	Knowledge	The percentage of IS audit team members are knowledgeable in IS standards.	Less than 25% of IS audit team members are knowledgeable in IS standards.	At least 25% of IS audit team members are knowledgeable in IS standards.	At least 50% of IS audit team members are knowledgeable in IS standards.	At least 75% of IS audit team members are knowledgeable in IS standards.	100% of IS audit team members are knowledgeable in IS standards.
			The percentage of IS audit team members are knowledgeable in the ISM scope of the audited organisation.	Less than 25% of IS audit team members are knowledgeable in the ISM scope of the audited organisation.	At least 25% of IS audit team members are knowledgeable in the ISM scope of the audited organisation.	At least 50% of IS audit team members are knowledgeable in the ISM scope of the audited organisation.	At least 75% of IS audit team members are knowledgeable in the ISM scope of the audited organisation.	100% of IS audit team members are knowledgeable in the ISM scope of the audited organisation.
		Auditing skills	The frequency of audit team members' involvement in internal and external audit within 3 years.	1 time involved in internal/external audit.	2 times involved in internal/external audit.	3 times involved in internal/external audit.	4 times involved in internal/external audit.	More than 4 times involved in internal/external audit.
		Commitment	Level of detail in writing audit notes.	Not detailed.	Lack of detail.	Quite Detailed	Detailed.	Very detailed.
		Cooperation	The work culture of IS audit team members during audit findings discussion.	No cooperation during audit findings discussion.	Lack of co-operation during audit findings discussion.	Quite cooperate during audit findings discussion.	Cooperate during audit findings discussion.	Strongly cooperate during audit findings discussion.
		Communication Skills	The clarity of information delivery (oral and written).	Very unclear.	Unclear.	Quite clear.	Clear.	Very clear.
	Employee	Awareness	The percentage of employees' awareness toward IS policy.	Less than 25% of employees are aware of IS policy.	At least 25% of employees are aware of IS policy.	At least 50% of employees are aware of IS policy.	At least 75% of employees are aware of IS policy.	100% of employees aware of IS policy.
		Compliance	The percentage of employees' compliance with IS policy.	Less than 25% of employees comply with IS policy.	At least 25% of employees comply with IS policy.	At least 50% of employees comply with IS policy.	At least 75% of employees comply with IS policy.	100% of employees comply with IS policy.
		Motivation	The frequency of employees receiving appreciation.	Never received an appreciation.	Rarely receive an appreciation.	Quite often receive appreciation.	Often receive appreciation.	Very often receive appreciation.
	Third parties	Awareness	The percentage of third parties' awareness toward IS policy.	Less than 25% of third parties are aware of IS policy.	At least 25% of third parties are aware of IS policy.	At least 50% of third parties are aware of IS policy.	At least 75% of third parties are aware of IS policy.	100% of third parties are aware of IS policy.
		Compliance	The percentage of third parties' compliance with IS policy and contracts.	Less than 25% of third parties comply with IS policy and contracts.	At least 25% of third parties comply with IS policy and contracts.	At least 50% of third parties comply with IS policy contracts.	At least 75% of third parties comply with IS policy contracts.	100% of third parties comply with IS policy and contracts.

0rg. Document	IS Policy	Clear	The percentage of IS policy contents that outlines the objectives, controls, and responsibilities of the parties involved are understood by the reader.	Less than 25% of IS policy contents are understood by readers.	At least 25% of the IS policy contents are understood by the reader.	At least 50% of the IS policy contents are understood by the reader.	At least 75% of the IS policy contents are understood by the reader.	100% of the IS policy contents are understood by the reader.
		Comprehensive	The percentage of security controls established is based on the recommendations of international standards and IS requirements.	Less than 25% of security controls are established based on the recommendations of international standards and IS requirements.	At least 25% of security controls are established based on the recommendations of international standards and IS requirements.	At least 50% of security controls are established based on the recommendations of international standards and IS requirements.	At least 75% of security controls are established based on the recommendations of international standards and IS requirements.	100% security controls are established based on the recommendations of international standards and IS requirements.
		Communicated	The frequency of IS policy dissemination.	Once a year.	2 times a year.	3 times a year.	4 times a year.	More than 4 times a year.
			The number of IS policy dissemination mediums.	1 medium.	2 mediums.	3 mediums.	4 mediums.	More than 4 mediums.
		Reviewed	The percentage of IS policy contents is reviewed/ updated according to current needs.	Less than 25% of IS policy contents are reviewed/ updated according to current needs.	At least 25% of IS policies contents are reviewed /updated according to current needs.	At least 50% of the IS policy contents are reviewed/ updated according to current needs.	At least 75% of IS policy contents are rereviewed/ updated according to current needs.	100% IS policy contents are reviewed/ updated according to current needs.
	IS Procedures	Clear	The percentage of IS procedures understood by the personnel/ team in charge.	Less than 25% of IS procedures are understood by the personnel/ team in charge.	At least 25% of IS procedures are understood by the personnel/ team in charge.	At least 50% of IS procedures are understood by the personnel/ team in charge.	At least 75 % of IS procedures are understood by the personnel/ team in charge.	100% IS procedures are understood by the personnel/ team in charge.
		Complete	The level of IS procedures feasibility.	Most of the procedures are very difficult to implement/ follow.	Most of the procedures are difficult to implement/ follow.	Most of the procedures are quite easy to implement/ follow.	Most of the procedures are easy to implement/ follow.	Most of the procedures are very easy to implement/ follow.
		Communicated	The frequency rate of the IS procedures communicated.	Most of the procedures are not communicated to the responsible officer.	Most of the procedures are rarely communicated to the responsible officer.	Most of the procedures are communicated to the responsible officer regularly.	Most of the procedures are communicated to the responsible officer as required.	Most of the procedures are communicated to the responsible officer periodically and as required.
		Reviewed	The Percentage of IS procedures reviewed/ updated according to current needs.	Less than 25% of IS procedures are reviewed/ updated according to current needs.	At least 25% of IS procedures are reviewed/ updated according to current needs.	At least 50% of IS procedures are reviewed/ updated according to current needs.	At least 75% of IS procedures are rereviewed/ updated according to current needs.	100% content of IS procedures reviewed/ updated according to current needs.
Process	Resource Planning	Financial resources	The amount of financial allocation to support the implementation of the ISM.	Very insufficient.	Insufficient.	Quite sufficient.	Sufficient.	Very sufficient and is given priority in the allocation application every year.
		Human Resources	The number of officers performing security operations.	Very insufficient	Insufficient.	Quite sufficient.	Sufficient.	Very sufficient.
			The competency level of allocated officers.	Not competent.	Lack of competence.	Quite competent.	Competent.	Very competent.

	Competency Development & Awareness	Training Programmes	The suitability of the training programmes given to employees and team members.	Most of the training programs given to staff and team members do not suit the work scope.	Most of the training programs given to staff and team members are less suited to the work scope.	Most of the training programs given to staff and team members are quite suited to the work scope.	Most of the training programs given to staff and team members are suited to the work scope.	Most of the training programs given to staff and team members are well suited to the work scope.
			The knowledge of the employees and team members after attending training programs.	Very low.	Low.	Moderate.	Good.	Excellent.
		Awareness Programmes	The number of awareness programs mediums in a year.	Awareness programs are implemented through 1 medium.	Awareness programs are implemented through 2 mediums.	Awareness programs are implemented through 3 mediums.	Awareness programs are implemented through 4 mediums.	Awareness programs are implemented in more than 4 mediums.
			The frequency of awareness programs in a year.	Once a year.	Twice a year.	3 times a year.	4 times a year.	More than 4 times a year.
			The percentage of security incidents has been reduced.	Less than 25% of security incidents have been reduced.	At least 25% of security incidents have been reduced.	At least 50% of security incidents have been reduced.	At least 75% of security incidents have been reduced.	100% security incidents have been reduced.
	Risk management	Risk Assessment	The percentage of process owners, asset owners and IS team's involvement in risk assessment.	Less than 25%.	At least 25%.	At least 50%.	At least 75%.	100%.
			The percentage of assets (included in the scope) that have been assessed.	Less than 25% of assets have been assessed.	At least 25% of the assets have been assessed.	At least 50% of the assets have been assessed.	At least 75% of the assets have been assessed.	100% of the assets have been assessed.
		Risk Treatment	Level of treatment suitability in managing risk.	Not appropriate	Less appropriate	Quite appropriate.	Appropriate.	Very appropriate.
			Percentage of high-risk assets that have been depreciated.	Less than 25%.	At least 25%.	At least 50%.	At least 75%.	100%.
	Business continuity and incident management	Plan	The percentage of plan availability.	Less than 25%.	At least 25%.	At least 50%.	At least 75%.	100%.
			The percentage of incidents and disasters successfully handled (identified, reported, recovered) within a set time.	Less than 25%.	At least 25%.	At least 50%.	At least 75%.	100%.
		Simulation	Diversity of simulation implementation over 5 years.	The same simulation was implemented over 5 years.	At least 2 different simulations were implemented over 5 years.	At least 3 different simulations were implemented over 5 years.	At least 4 different simulations were implemented over 5 years.	More than 4 different simulations were implemented over 5 years.
	IS Audit	Audit program	The level of audit scope.	The scope of the audit is not comprehensive.	The scope of the audit is less comprehensive.	The scope of the audit is quite comprehensive.	The scope of the audit is comprehensive.	The scope of the audit is comprehensive and has value-added.

		Audit Findings and Reporting	The clarity percentage of the audit findings and reporting.	Less than 25% of audit findings are clearly reported.	At least 25% of audit findings are clearly reported.	At least 50% of audit findings are clearly reported.	At least 75% of audit findings are clearly reported.	100% of audit findings are clearly reported.
		Follow-up Audit	The level of follow-up audit review.	The revision of the corrective and preventive actions is carried out incomplete.	The revision of the corrective and preventive actions is carried out less completely.	The revision of corrective and preventive actions is carried out quite completely.	The revision of the corrective and preventive actions is carried out completely.	The revision of the corrective and preventive actions is carried out completely and thoroughly.
			The accuracy percentage of the implementation of the preventive and corrective actions.	Less than 25% of corrective and preventive actions are implemented appropriately.	At least 25% of corrective and preventive actions are implemented appropriately.	At least 50% of corrective and preventive actions are implemented appropriately.	At least 75% of corrective and preventive actions are implemented appropriately.	100% of corrective and preventive actions are implemented appropriately.
Technology	IT Infrastructure	Hardware	The percentage of hardware maintenance.	Less than 25% of hardware is maintained on schedule.	At least 25% of the hardware is maintained on schedule.	At least 50% of the hardware is maintained on schedule.	At least 75% of the hardware is maintained on schedule.	100% hardware is maintained on schedule.
			The percentage of latest hardware used.	Less than 25% of the latest hardware is used.	At least 25% of the latest hardware is used.	At least 50% of the latest hardware is used.	At least 75% of the latest hardware is used	100% up-to-date hardware is used.
		Software	The percentage of software maintenance (updated version/security features in software architecture).	Less than 25% of software is maintained on schedule.	At least 25% of the software is maintained on schedule.	At least 50% of the software is maintained on schedule.	At least 75% of the software is maintained on schedule.	100% software is maintained on schedule.
			The percentage of use of software security functions.	Less than 25% of software security functions are used.	At least 25% of software security functions are used.	At least 50% of software security functions are used.	At least 75% of software security functions are used.	100% software security functions are used.

V. CONCLUSION

ISM is a strategic approach to address IS risks and breaches as well as to reduce IS incidents that can compromise the confidentiality, integrity and availability of organisational information. These IS risks, incidents and breaches can be minimised if the organisation implements ISM effectively. The effectiveness of ISM can be achieved if organisations assess the maturity of their ISM practices using a holistic maturity model. A holistic maturity model needs to consider the ISM success factors in every aspect to ensure that the assessment is made comprehensively.

This study has successfully developed a holistic maturity model to help organisations in self-assessing the maturity level of their ISM implementation. This initiative encourages organisations to continue improving the implementation of their ISM from time to time. This model can also be used as guidelines and references to academicians and researchers involved in information security maturity.

Finally, here are some suggestions for further research that can be implemented in the future:

- Specialise the model according to the type of organisation.

This study does not specialise in any particular type of organisation, whether public or private organisation. The nature of service is quite different between those two sectors, and it is believed that organisations in both sectors have relatively slightly different information security controls. Accordingly, detailed studies by the type of organisation can be done in the future to produce a more accurate model.

- Automate the maturity model.

Further studies are proposed to automate the Organisational ISM Maturity Model. The automated ISM maturity model not only simplifies the evaluation process but can also be used for record-keeping and report generating. This allows the organisation to monitor the progress of the ISM, compare the maturity level obtained each year, as well as predict the level of maturity that will be obtained in subsequent years more easily.

ACKNOWLEDGMENT

The authors would like to thank Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia for supporting this research. The authors also thank the practitioners and experts who participated in this study.

REFERENCES

[1] Mirtsch, M., Blind, K., Koch, C. and Dudek, G., "Information security management in ICT and non-ICT sector companies: A preventive innovation perspective," Computers & Security 109, pp. 1-23, 2021.

[2] Chu, A.M. and So, M.K., "Organizational information security management for sustainable information systems: An unethical employee information security behavior perspective," Sustainability vol. 12 no. 8, pp. 3163 – 3187, 2020.

[3] Napitupulu, Darmawan. "A conceptual model of e-government adoption in Indonesia." International Journal on Advanced Science, Engineering and Information Technology 7, vol. 4, pp. 1471-1478, 2017.

[4] Kadhum, Ahmed Meri, and Mohamad Khatim Hasan. "Assessing the determinants of cloud computing services for utilizing health information systems: A case study." International Journal on Advanced Science, Engineering and Information Technology vol.7, no. 2, pp. 503-510, 2017.

[5] S. Woodhouse, "Critical Success factors for an Information Security Management System," in 5th International Conference on Information Technology and Applications ICITA 2008, 2008, no. Icita, pp. 244–249.

[6] Jere, Joseph N., and Nsikelelo Ngidi, "A technology, organisation and environment framework analysis of information and communication technology adoption by small and medium enterprises in Pietermaritzburg," South African Journal of Information Management, vol. 22 no. 1, pp. 1-9, 2020.

[7] Witarsyah, Deden, et al. "The critical factors affecting E-Government adoption in Indonesia: A conceptual framework." International Journal on Advanced Science, Engineering and Information Technology, vol. 7, no. 1, pp. 160-167, 2017.

[8] Ključnikov, A., Mura, L. and Sklenár, D, "Information security management in SMEs: factors of success," Entrepreneurship and Sustainability Issues vol. 6 no. 4, pp. 2081-2094, 2019.

[9] Khan, Navid Ali, Sarfraz Nawaz Brohi, and Noor Zaman, "Ten deadly cyber security threats amid COVID-19 pandemic," TechRxiv Powered by IEEE, 2020.

[10] Tu, C.Z., Yuan, Y., Archer, N. and Connelly, C.E., "Strategic value alignment for information security management: A critical success factor analysis," Information & Computer Security, pp. 1-28. 2018.

[11] Rahayu, H. and Rozilawati, R., "Contributing Factors for Successful Information Security Management Implementation: A Conceptual Model," International Journal of Innovative Technology and Exploring Engineering (IJITEE) vol. 9, no.2, pp. 4491-4499, 2019.

[12] Makupi, D. and Masese, N., "Determining Information Security Maturity Level of an organization based on ISO 27001," International Journal of Computer Science and Engineering vol. 6, no. 7, pp. 5-11, 2019.

[13] Singh, A.N. and Gupta, M.P., "Information security management practices: case studies from India," Global Business Review vol. 20, no. 1, pp. 253-271, 2019.

[14] Rahayu, H. and Razali, R., "Contributing Factors for Successful Information Security Management Implementation: A Preliminary Review." The Interdisciplinary Of Management, Economic And Social Research, pp. 12-22, 2020.

[15] R. Hashim and R. Razali, "Contributing Factors for Successful Information Security Management Implementation: A Preliminary Review," The Interdisciplinary Of Management, Economic And Social Research, vol. 9, no.2, p.12, 2019.

[16] Center for Strategic and International Studies (CSIS), "Significant Cyber Incidents Since 2006." 2021.

[17] National Cyber Security Agency. Cyber Security Incident Statistics. 2020.

[18] Schmid, M. and Pape, S., "A structured comparison of the corporate information security maturity level," IFIP International Conference on ICT Systems Security and Privacy Protection, pp. 223-237, 2019.

[19] M. F. Saleh, "Information Security Maturity Model," Int. J. Comput. Sci. Secur., vol. 5, no. 3, p. 21, 2011.

[20] T. Dirgahayu and D. Ariyadi, "Assessment to C OBIT 4 . 1 Maturity Model Based on Process Attributes and Control Objectives," in 2015 International Conference on Science in Information Technology (ICSITech), 2015, pp. 343–347.

[21] ITGI, The Control Cbjectives for Information and Related Technology (COBIT 4.1). 2007.

[22] V. C. Aceituno, "Ism3 1.0. Information Security Management Maturity Model," 2004.

[23] W. Sung and S. Kang, "An empirical study on the effect of information security activities: focusing on technology, institution, and awareness," Proceedings of the 18th Annual International Conference on Digital Government Research, pp. 84–93, 2017.

[24] A. S. Lima, J. N. de Souza, E. C. Branco, and M. Ribas, "Towards value-based information security management monitoring," Integr. Netw. Manag. (IM 2013), 2013 IFIP/IEEE Int. Symp., pp. 1260–1267, 2013.

[25] S. Dzazali and A. H. Zolait, "Assessment of information security maturity: An exploration study of Malaysian public service organizations," J. Syst. Inf. Technol., vol. 14, no. 1, pp. 23–57, 2012.

[26] M. Zammani and R. Razali, "An Empirical Study of Information Security Management Success Factors," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 6, no. 6, pp. 904–913, 2016.

[27] O. Matrane, M. Talea, C. Okar, and A. Talea, "Towards A New Maturity Model for Information System," 2015 Int. J. Comput. Sci. Issues, vol. 3, no. 12, pp. 268–275, 2015.

[28] K. Randeree, A. Mahal, and A. Narwani, "A business continuity management maturity model for the UAE banking sector," Bus. Process Manag. J., vol. 18, no. 3, pp. 472–492, 2012.

[29] J.V. Carvalho, A. Rocha, R. van de Wetering and A. Abreu, "A Maturity model for hospital information systems," Journal of Business Research, 94, pp. 388-399, 2019.

[30] C. S. Leem, B. W. Kim, E. J. Yu, and M. H. Paek, "Information technology maturity stages and enterprise benchmarking: an empirical study," Ind. Manag. Data Syst., vol. 108, no. 3, pp. 1200–1218, 2008.

[31] G. C. S. C. C. GSCSCC, Cybersecurity Capacity Maturity Model for Nations (CMM). Revised Edition, no. CMM. 2016.

[32] TOG, Open Information Security Management Maturity Model. The Open Group, 2011.

[33] N. Ibrahim and N. Ali, "The Role of Organizational Factors to the Effectiveness of ISMS Implementation in Malaysian Public Sector," Int. J. Eng. Technol., vol. 7, no. 4.35, pp. 544–550, Nov. 2018.

[34] P. K. Sari, N. Nurshabrina, and Candiwan, "Factor Analysis on Information Security Management in Higher Education Institutions," in 4th International Conference on Cyber and IT Service Management, pp. 1-5. IEEE, 2016., 2016, pp. 1–5.

[35] M. a. Alnatheer, "Information Security Culture Critical Success Factors," in 2015 12th International Conference on Information Technology - New Generations, 2015, pp. 731–735.

[36] P. Bowen, J. Hash, and M. Wilson, NIST Special Publication 800-100 - Information Security Handbook: A Guide for Managers, no. October. Maryland, USA: National Institute of Standards and Technology (NIST) in furtherance of its statutory responsibilities under the Federal Information Security, 2006.

[37] A. N. Singh, M. P. Gupta, and A. Ojha, "Identifying factors of 'organizational information security management,'" J. Enterp. Inf. Manag., vol. 27, no. 5, p. 8, 2014.

[38] M. Chander, S. K. Jain, and R. Shankar, "Modeling of information security management parameters in Indian organizations using ISM and MICMAC approach," J. Model. Manag., vol. 8, no. 2, pp. 171–189, 2013.

[39] M. Kazemi, H. Khajouei, and H. Nasrabadi, "Evaluation of information security management system success factors: Case study of Municipal organization," African J. Bus. Manag., vol. 6, no. 14, pp. 4982–4989, 2012.

[40] N. Maarop, N. Mustapha, R. Yusoff, R. Ibrahim, and N. M. M. Zainuddin, "Understanding Success Factors of an Information Security Management System Plan Phase Self-Implementation," Int. J. Soc. Behav. Educ. Econ. Bus. Ind. Eng., vol. 9, no. 3, pp. 884–889, 2015.

[41] A. Cartlidge et al., An Introductory Overview of ITIL® 2011. Orwich: TSO (The Stationery Office), 2012.

[42] COBIT v 5, COBIT for information security. Rolling Meadows, IL: ISACA, 2012.

[43] ISO, "ISO / IEC 27001: Information Technology – Security Techniques –Information Security Management System – Requirements," 2013.

[44] M. Zammani and R. Razali, "Information security management success factors," Adv. Sci. Lett., vol. 22, no. 8, 2016.

[45] MAMPU, CGSO, C. Malaysia, and MIMOS, "Rangka Kerja Keselamatan Siber Sektor Awam," 2016.

[46] M. A. Mohamad Stambul; and R. Razali, "An assessment model of information security implementation levels," in Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, 2011, pp. 1–6.

[47] A. Azhari, "Ke Arah Implementasi Sistem Polisi Keselamatan ICT Kajian Kes : Pusat Teknologi Maklumat dan Komunikasi," Universiti Teknologi Malaysia, 2008.

[48] Q. Hu, T. Dinev, P. Hart, and D. Cooke, "Managing Employee Compliance with Information Security Policies : The Critical Role of Top Management and Organizational Culture ∗," Decis. Sci. J., vol. 00, no. 00, pp. 1–45, 2012.

[49] ISO, "ISO / IEC 27002: Information Technology - Security techniques - Code of practice for information security controls," 2013.

[50] R. Diesch, M. Pfaff, and H. Krcmar, "A comprehensive model of information security factors for decision-maker," Computers & Security, 92, p. 101747, 2020.

[51] M. S. Saleh and A. Alfantookh, "A new comprehensive framework for enterprise information security risk management," Appl. Comput. Informatics, vol. 9, no. 2, pp. 107–118, 2011.

[52] N. Aisyah, S. Abdullah, N. L. Noor, E. Nuraihan, and M. Ibrahim, "Contributing Factor To Business Continuity Management (Bcm) Failure – a Case of Malaysia Public Sector," in Proceedings of the 5th International Conference on Computing and Informatics, ICOCI, 2015, no. 077, pp. 530–538.

[53] S. Islam, N. Farah, and T. F. Stafford, "Factors associated with security / cybersecurity audit by internal audit function An international study function," Manag. Audit. J., vol. 33, no. 4, pp. 377–409, 2018.

[54] A. A. Norman and N. M. Yasin, "Information Systems Security Management (ISSM) Success Factor: Retrospection From the Scholars," Proceedings of the 11th European Conference on Information Warfare and Security, no. July 2012. pp. 339–344, 2012.

[55] S. Chowdhury and K. M. Salahuddin, "A Literature Review of Factors Influencing Implementation of Management Information Systems in Organizations," vol. 12, no. 8, pp. 72–79, 2017.

[56] E. A. Rigon, C. M. Westphall, and D. R. Dos Santos, "A cyclical evaluation model of information security maturity," Inf. Manag. Comput. Secur., vol. 22, no. 3, pp. 265–278, 2014.

[57] Saleh, M. F. "Information Security Maturity Model," International Journal of Computer Science and Security (IJCSS), vol. 5, no. 3, p. 21, 2011.

[58] Y. Goksen, E. Cevik, and H. Avunduk, "A Case Analysis on the Focus on the Maturity Models and Information Technologies," in Procedia Economics and Finance, pp. 208–216, 2015.

[59] H. Stewart and J. Jürjens, "Information security management and the human aspect in organizations," Inf. Comput. Secur., vol. 25, no. 5, pp. 494–534, 2017.

[60] M. Zammani, R. Razali, and D. Singh, "Factors contributing to the success of information security management implementation," International Journal of Advanced Computer Science and Applications, vol. 10, no 11, pp. 384–391, 2019.

[61] ISO, "ISO/IEC 33004:2015 - Information technology — Process assessment — Requirements for process reference, process assessment and maturity models," 2015.

[62] Stevens, S.S., "On the theory of scales of measurement", Science, vol. 103, no. 2684, pp. 677-680, 1946.

[63] Sarle, W.S., "Measurement theory: Frequently asked questions," Disseminations of the International Statistical Applications Institute vol. 1, no. 4, pp. 61-66. 1995.

Applying Grey Clustering and Shannon's Entropy to Assess Sediment Quality from a Watershed

Alexi Delgado[1], Betsy Vilchez[2], Fabian Chipana[3], Gerson Trejo[4]
Renato Acari[5], Rony Camarena[6], Víctor Galicia[7], Chiara Carbajal[8]
Mining Engineering Section, Pontificia Universidad Católica del Perú, Lima, Peru[1]
Environmental Engineering Department, National University of Engineering, Lima, Peru[2, 3, 4, 5, 6, 7]
Administration Program, Universidad de Ciencias y Humanidades, Lima, Peru[8]

***Abstract*—The evaluation of the quality of sediments is a complex issue in the Peruvian reality, mainly because there is no sampling protocol or norm for comparison, which leads to the assessment of sediments without a comprehensive analysis of their quality. In the present study, the quality of the sediments in the upper basin of the Huarmey river was evaluated in 30 monitoring points and 7 parameters, which are: arsenic, cadmium, copper, chromium, mercury, lead and zinc, which were compared according to the standards recommended by the Environmental Quality Guidelines for Sediments in freshwater bodies of Canada (Canadian Environmental Quality Guidelines - CEQG, 2002. Sediment Quality Guidelines for Protection of Aquatic Life - Fresh water according to Canadian Council of Ministers of the Environment (CCME)). The results of the evaluation, by grey clustering method and Shannon entropy, showed that 13 monitoring points resulted in good sediment quality, 1 monitoring point had moderate quality and 16 monitoring points presented poor quality; therefore, it can be concluded that the effluents and discharges of the mining activities that take place in the aforementioned location have a negative impact on environmental quality. Finally, the results obtained can be of great help for OEFA, the regional government, the municipalities and any other body that has oversight functions, since they will allow them to be more objective and precise decisions.**

***Keywords*—Grey clustering; sediment quality; Shannon entropy**

I. Introduction

The water resource of the Huarmey river basin in Ancash, Peru, represents the vital element for the supply of population, agricultural, livestock, mining, energy and ecological use [1], being important its optimal, rational and sustainable use. However, due to the continuous complaints from the surrounding population expressing their discomfort over an alleged environmental impact on the water and sediment of the upper basin of the Huarmey river, an environmental monitoring was carried out, where there is a record of the existence of mining activity in exploration and operation stage (Minera Huinac SAC) [2].

In the upper basin of the Huarmey river, districts of La Merced, Aija, Huacllan and Succha, province of Aija, department of Ancash, Peru. The sediment quality assessment was carried out, for which there were 30 monitoring points carried out by the Organization for Environmental Assessment and Enforcement, a specialized technical organization in charge of compliance with Peruvian environmental regulations [2]. It is important to mention that, at present, Peru does not have a sediment sampling protocol and nor does it have regulations or quality standards to evaluate this component. Therefore, standards recommended by the Environmental Quality Guidelines for Sediments in freshwater bodies of Canada (Canadian Environmental Quality Guidelines - CEQG, 2002) will be used. Sediment Quality Guidelines for Protection of Aquatic Life - Fresh water according to the Canadian Council of Ministers of the Environment [3].

For the evaluation of the sediment quality, we will use the Grey Clustering method, as well as the Shanon entropy. Grey Clustering is a methodology that is based on a theory of fuzzy sets and can be applied by grey incidence of matrices or whitenization functions because unlike traditional statistical methods [4], this methodology considers the uncertainty of the fuzzy type which is present in the environment within their analysis. For the case study, the "Center–point Triangular Whitenization Functions - CTWF" will be used, since the CTWF is mainly applied to test if the observation objects belong to predetermined classes, known as grey classes [5] as evidenced in the studies of selection of innovative strategies [6] and in the evaluation of the quality of sediments by grey incidence [7]. On the other hand, the Shannon entropy method is also an artificial intelligence approach developed by Claude E. Shannon (Shannon and Weaver, 1994) that addresses the uncertainty due to the dispersion of the data [8], therefore this method was also used to determine the weights of the evaluation criteria within the CTWF method [9].

Therefore, the specific objective of the present work is to analyze and value the quality of sediments by Grey method Clustering and Shannon entropy in the upper basin of the Huarmey river monitored in May 2016 based on the standards recommended by the Guidelines Environmental Quality for Sediments in freshwater bodies of Canada (Canadian Environmental Quality Guidelines - CEQG, 2002. Sediment Quality Guidelines for Protection of Aquatic Life - Fresh water according to Canadian Council of Ministers of the Environment [3].

Thus, on the following the study is formed by Section II which summarizes the literature review; Section III, in which the CTWF method is explained in detail. After Section IV, will be the section where the case study is described, then the results and their discussion are presented in Section V. Finally, the conclusions are presented in Section VI.

II. Literature Review

Delgado et al., in 2017, developed a research that studied the water quality of the Santa River, in which different points were analyzed according to the parameters established by MINAM-Peru (DS No 015-2015). In this sense, 21 monitoring points of the Santa River basin were analyzed. It was concluded that the grey clustering method showed interesting results and that it could be applied to other studies on water quality or the environment in general. In this regard, the results showed that 47.6% of the monitoring points presented a good quality of water for human consumption, which could be purified by applying a disinfection; 33.3% of the monitoring points presented a moderate quality of water for human consumption, which could be purified with a conventional treatment; and 19.1% of the monitoring points presented a low quality of water for human consumption, which could be purified by applying a special treatment [10].

Delgado, in 2020, in Peru, mentions that evaluating the quality of surface waters is a complex issue that involves the comprehensive analysis of several parameters that are altered by natural or anthropogenic causes. In this sense, the Grey Clustering method, which is based on the Grey Systems theory [11], and the Shannon Entropy, based on the artificial intelligence approach, provide an alternative to evaluate water quality in a comprehensive manner considering the uncertainty within the analysis. In the mentioned study, the water quality in the upper basin of the Huallaga River was evaluated taking into account the results of the monitoring of twenty-one points carried out by the National Water Authority analyzing nine parameters of the Prati index. The results showed that all the monitoring points of the Huallaga River were classified as uncontaminated, which means that the discharges, generated by economic activities, are carried out through treatment plants that meet the quality parameters [12].

Environmental Assessment and Control Agency (OEFA by its Spanish acronym), in Peru, 2016 water and sediment quality monitoring was carried out from May 20 to 30 at 30 monitoring points belonging to the upper basin of the Huarmey River, which is formed by the Llactún and La Merced rivers with their respective tributaries, that when joined form the Santiago River which receives the contribution of the stream of the same name, in which there is a record of the existence of mining activity in exploration and operation stage (Minera Huinac S. A.C.). Finally, downstream it takes the name of the Aija River, which receives the contributions of the Mallqui and Allma rivers, where concentrations of total arsenic were exceeded in the 30 monitoring points, copper in 26 points, mercury in 10 points, lead in 18 points, zinc in 25 points of the reference values of the Canadian standard [2].

Chu and Tan, in 2014, in China, carried out the analysis of 39 samples of surface sediments, from the coastal ocean of Jiangsu to evaluate their quality. Making use of the Grey Clustering method for its evaluation and generating results classified into three categories (clean, light pollution and intense pollution). Of the thirty-nine samples, there are eleven clean samples, twenty light pollution samples, and eight heavy pollution samples. When analyzing the underlying reasons, pollutants dumped into the sea due to increased industrial and agricultural activities that contributed to the pollution. Therefore, more emphasis should be placed on the management of the surface tidal flat sediment environment, especially on the treatment of the pollution source to improve the sediment quality for the sustainable development of the coastal zone [7].

Delgado, in 2018, in Peru, states that the assessment of pollution and the quality of the air is a serious problem for big cities, considering the increasing pollution of the air. In this sense, the evaluation of this problem using the grey clustering method, which is based on the theory of the grey system, has great advantages since it considers this uncertainty within the analysis. In such study, an evaluation of air quality was carried out in three monitoring points located in three different districts of the city of Lima, Peru, which are San Martín de Porres, Carabayllo and Puente Piedra. The results revealed that the three monitoring points presented good air quality in accordance with Peruvian law. Nevertheless, this could be because the districts in which the monitoring points are located are relatively new. Finally, the results of this study could help local and central authorities to make the best decision on the evaluation of air quality [5].

National Water Authority (ANA by its Spanish acronym), in November 2015, evaluated the quality of the surface water of the Huarmey river basin, concluding that the effects on the bodies of water belonging to and / or tributaries to the Huarmey basin were located in the upper area of said basin, being the Montecristo, Huinac, Hercules and Santiago streams, as well as the Llactún river, which presented concentration values of the content of metals such as aluminum, arsenic, cadmium, copper, iron, manganese, lead and zinc, which exceed the value established in the ECA- Water, due to the fact that mining companies are installed in this area at the head of the basin [13].

The application of the Grey Clustering and Shannon Entropy methodology in sediment quality is an innovative method; this is due to the scarce existing bibliography that applies similar methodologies and the non-existence of sediment quality standards in Peru. That is why the importance of the present work lies in the application of these powerful methodologies in a new field of study, the Peruvian context, in addition, it allows perceiving the quality in an environmental component of the upper basin of the Huarmey River, Peru.

III. Methodology

In this section, we will proceed to describe the center-point triangular whitenization weight functions (CTWF) method, which can be described as follows: First assume that there is a set of m objects, a set of n criteria and a set of s grey classes, according to the sample value ($i = 1,2,\dots,m$; $j = 1,2,\dots,n$) Then the CTWF method is appreciated in a flowchart in Fig. 1 and are developed with the following steps [12], [14]–[16].

A. Step 1: Determining the Center Points

The criteria ranges are divided into three grey classes, which are: $\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_n$ and these values are determined using the Canadian sediment standard.

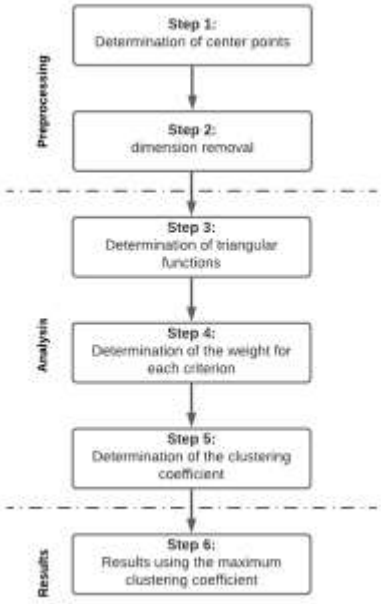


Fig. 1. Flowchart of the CTWF Method.

B. Step 2: Dimension Removal

At this point it is assumed that there are objects for evaluation and n criteria or grey classes, which form the following matrix $Z = \{Z_{i\,j};\ i = 1,2,\dots,m\ ;\ j = 1,2,\dots,n\}$. In this sense it is normalized by each criterion C_j $(j = 1,2,\dots,n)$. The normalized value $P_{i\,j}$, which is calculated by (1).

$$P_{i\,j} = \frac{Z_{i\,j}}{\frac{\sum_{j=1}^{n} Z_{i\,j}}{n}} \tag{1}$$

C. Step 3: Determination of Triangular Functions

The grey classes are expanded in the directions of each parameter used and for this purpose the Canadian standard for sediments will be used as a reference, which provides values to measure the quality. In this research the Canadian standard provides us with three quality levels for each parameter analyzed, so we will have three functions for each parameter analyzed. The new sequence of center points in $\lambda_1, \lambda_2\, y\, \lambda_3$. For class $k = 1,2y\ 3, j = 1,2,\dots,n$ for an observed value x_{ij}. The computation of the Central Point Triangular Whitenization Functions (CTWF) is shown by (2) – (4). A visual representation is shown in Fig. 2.

$$f_j^1(x_{ij}) = \begin{cases} 1 & x \in [0, \lambda_j^1] \\ \frac{\lambda_j^2 - x}{\lambda_j^2 - \lambda_j^1} & x \in < \lambda_j^1, \lambda_j^2 > \\ 0 & x \in [\lambda_j^2, +\infty] \end{cases} \tag{2}$$

$$f_j^i(x_{ij}) = \begin{cases} \frac{x - \lambda_j^{k-1}}{\lambda_j^k - \lambda_j^{k-1}} & x \in < \lambda_j^{k-1}, \lambda_j^k] \\ \frac{\lambda_j^{k+1} - x}{\lambda_j^{k+1} - \lambda_j^k} & x \in < \lambda_j^k, \lambda_j^{k+1} > \\ 0 & x \in [0, \lambda_j^{k-1}] \mathrm{U} [\lambda_j^{k+1}, +\infty] \end{cases} \tag{3}$$

$$f_j^n(x_{ij}) = \begin{cases} \frac{x - \lambda_j^{n-1}}{\lambda_j^n - \lambda_j^{n-1}} & x \in < \lambda_j^{n-1}, \lambda_j^n > \\ 1 & x \in [\lambda_j^n, +\infty > \\ 0 & x \in [0, \lambda_j^{n-1}] \end{cases} \tag{4}$$

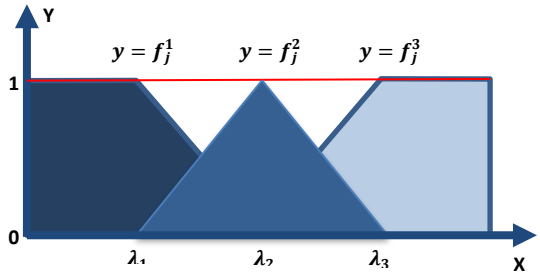


Fig. 2. CTWF Representation.

D. Step 4: Determining Weight for each Criterion

In this step, use is made of Shannon's Entropy weight method, which developed the measure "H", which satisfies the following properties [9], [12], [14], [15], [17]:

- H is a positive continuous function.
- If all p_i are equivalent $p_i = 1/n$, in that sense, H should be a growing monotonous function of n.
- For all $n \geq 2, H(P_1, P_2, \dots, P_n) = h(p_1 + p_2, p_3, \dots, P_n) + (p_1 + p_2)H(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2})$

Shannon shows that only functions satisfying this condition are computed by (5).

$$H_{Shannon} = -\sum_i^n p_i log(p_i) \tag{5}$$

Where: $0 \leq p_i \leq 1; \Sigma_{f=1}^n p_i = 1$

Regarding the entropy weight methodology, it can be demonstrated according to the following definition [9], [12], [15]. As shown above, m objects are shown for evaluation and n evaluation criteria, which form the following matrix $x = \{x_{i\,j};\ i = 1,2,\dots,m\ ;\ j = 1,2,\dots,n\}$. After that, the following steps will continue.

1) The matrix $x = \{x_{i\,j};\ i = 1,2,\dots,m\ ;\ j = 1,2,\dots,n\}$ is normalized by each criterion C_j. The normalization evaluates $P_{i\,j}$ and are calculated by (6).

$$f_j^1(x_{i\,j}) = P_{i\,j} = \frac{x_{i\,j}}{\Sigma_{i=1}^m x_{i\,j}} \tag{6}$$

2) The entropy of each criterion is calculated by (7).

$$H = -k \sum_{i=1}^n p_{i\,j}\ ln(p_{i\,j}) \tag{7}$$

Where, k is constant, $k = (ln(m))^{-1}$

3) The degree of divergence of intrinsic information of each criterion C_j is calculated by (8).

$$div_j = 1 - H_j \tag{8}$$

4) The weight of the entropy of each criterion C_j is calculated by (9).

$$w_j = \frac{div_j}{\Sigma_{j=1}^{n} div_j} \quad (9)$$

Where, w_j it's equal to the n_j

E. Step 5: Determining the Clustering Coefficient

The clustering coefficient σ_i^k per object $i, i = 1,2,\ldots,m$, with respect to the grey classes $k, k = 1,2,\ldots,s$, is calculated by (10).

$$\sigma_i^k = \Sigma_{j=1}^{n} f_j^k(x_{i\,j})\, n_j \quad (10)$$

Where $f_j^k(x_{i\,j})n_j$ is the CTWF of the k^{th}grey class of the j^{th} criterion, and n_j is the weight of criterion j, to establish these the Shannon Entropy method will be used.

F. Step 6: Results using the Maximum Clustering Coefficient

Finally, the value of $max_{1\le k\le s}\{\sigma_i^k\} = \sigma_i^k$ has to be calculated, based on that the object belongs to each grey class is opted. When there are several objects in some grey class, these objects can be ordered according to the magnitudes of their clustering coefficients.

IV. Case Study

A. Description of the Study Area

The study area is in the upper basin of the Huarmey River, district of La Merced, Aija, Huacllan and Succha, province of Aija, department of Ancash, Peru as shown in Fig. 3. This is due to the results of water and sediment quality information, carried out from May 20 to 30, 2016. For which there were 30 monitoring points conducted by the Organization for Environmental Evaluation and Enforcement [2].

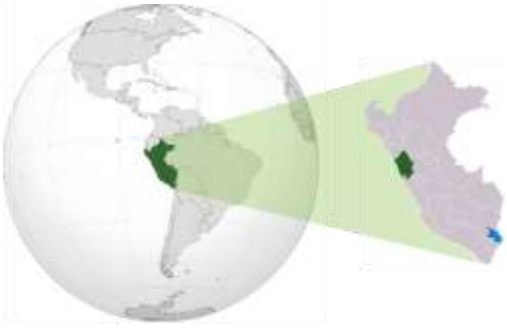


Fig. 3. Location of the Department of Ancash in Peru.

B. Description of Study Objects

For the evaluation of sediment quality in the upper basin of the Huarmey River La Merced, Aija, Huacllan and Succha districts, province of Aija, department of Ancash, information was collected from 30 monitoring points obtained from Report No. 266-2016 OEFA/DE- SDCA, carried out May 20 to 30, 2016 [2], and are shown in Table I and Fig. 4.

TABLE I. Monitoring Points in the upper Huarmey River Basin

Point	Code	Coordinates UTM - WGS 84 18S		
		East (m)	*North (m)*	*Elevation (m.a.s.l.)*
1	SED-HUA1	206381	8925199	3901
2	SED-QHuin1	206427	8925324	3906
3	SED-RLlac1	206461	8925143	3893
4	SED-RLlc2	206506	8924857	3856
5	SED-HUA2	206585	8924591	3841
6	SED-RLlac3	206711	8924450	3818
7	SED-HUA3	206948	8923944	3813
8	SED-RLlac4	207197	8924040	3781
9	SED-HUA4	207347	8923936	3789
10	SED-RLlac5	207254	8923762	3757
11	SED-HUA5	207220	8923686	3769
12	SED-RLlac6	207363	8923601	3748
13	SED-RLlac10	209764	8921211	3411
14	SED-RLlac7	208328	8922234	3618
15	SED-RLlac8	208479	8922051	3584
16	SED-HUA6	208442	8922203	3620
17	SED-HUA8	209528	8922005	3367
18	SED-HUA10	211354	8921662	3446
19	SED-RLMer3	212160	8920813	3158
20	SED-HUA9	212696	8922670	3292
21	SED-RLMer1	212970	8923072	3305
22	SED-RLMer2	212762	8922439	3269
23	SED-RAija1	209384	8916087	2774
24	SED-RSant2	209581	8916347	2832
25	SED-Rall2	214163	8915170	3290
26	SED-RLlac11	211204	8918533	3040
27	SED-RLMer4	211293	8918643	3049
28	SED-QSant5	211348	8918603	3066
29	SED-RSant1	211155	8918424	3066
30	SED-RMall1	215369	8914339	3417

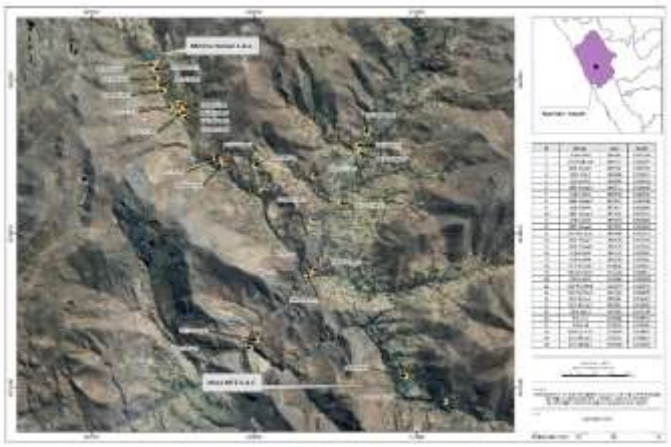


Fig. 4. Sediment Quality Monitoring Points in the upper Huarmey River Basin.

C. Description of Evaluation Criteria

The evaluation criteria for the present study are determined by the Canadian sediment quality parameters, which are presented in Table II:

TABLE II. SEDIMENT QUALITY ASSESSMENT CRITERIA

Criteria	Units	Notation
Total Arsenic	mg/Kg	C_1
Total Cadmium	mg/Kg	C_2
Total Copper	mg/Kg	C_3
Total chromium	mg/Kg	C_4
Total Mercury	mg/Kg	C_5
Total Plomium	mg/Kg	C_6
Total Zinc	mg/Kg	C_7

D. Definition of Grey Classes

The Grey Classes for the evaluation are 3 and are based on the sediment quality levels of the international standard - Canadian Sediment Quality Guidelines (CSQG) according to the Canadian Council of Ministers of the Environment [3] where it contemplates the limits to the ISQG (Interim Sediment Quality Guideline) values: the concentration below which no adverse biological effects are expected and PEL (Probable Effect Level): concentration above which adverse biological effects are frequently found. Proposing a third category considered as Moderate, which is presented in Table III.

E. Calculations using the CTWF Method

Step 1: Center points

Based on the international standard CSQG, the central values of the parameters to be analyzed were obtained and denominated as equivalent ISQG: Good and PEL: Poor. These values are shown in Table IV.

Step 2: Dimension removal

The non-dimension values for each parameter according to Table III on the average concentration of each parameter are shown in Table V.

TABLE III. QUALITY INTERVALS FOR THE INTERNATIONAL CSQG STANDARD

Parameter (mg/Kg)	Quality Index Condition		
	ISQG	*Moderate*	*Pel*
Total Arsenic	< 5.9	5.9 - 17	17 <
Total Cadmium	< 0.6	0.6 - 3.5	3.5 <
Total Copper	< 35.7	35.7 - 197	197 <
Total chromium	< 37.3	37.3 - 90	90 <
Total Mercury	< 0.17	0.17 - 0.486	0.486 <
Total Plomium	< 35	35 - 91.3	91.3 <
Total Zinc	< 123	123 - 315	315 <

TABLE IV. CENTRAL VALUES OF THE PARAMETERS FOR THE INTERNATIONAL STANDARD CSQG

Parameter (mg/Kg)	Quality Index Condition		
	Good	*Moderate*	*Bad*
Total Arsenic	5.9	11.45	17
Total Cadmium	0.6	2.05	3.5
Total Copper	35.7	116.35	197
Total chromium	37.3	63.65	90
Total Mercury	0.17	0.328	0.486
Total Plomium	35	63.15	91.3
Total Zinc	123	219	315

TABLE V. NON-DIMENSION VALUES FOR THE CSQG INTERNATIONAL STANDARD

Parameter (mg/Kg)	Quality Index Condition		
	Good	*Moderate*	*Bad*
Total Arsenic	0.515	1.000	1.485
Total Cadmium	0.293	1.000	1.707
Total Copper	0.307	1.000	1.693
Total chromium	0.586	1.000	1.414
Total Mercury	0.518	1.000	1.482
Total Plomium	0.554	1.000	1.446
Total Zinc	0.562	1.000	1.438

Similarly, based on the monitoring results obtained from Report No. 266-2016 OEFA/DE- SDCA, dimensionless values were obtained for the 30 selected monitoring points, as the example we show in Table VI for 5 monitoring points.

TABLE VI. DATA WITHOUT MONITORING DIMENSION IN THE CASE STUDY

Parameters	C1 - Total Arsenic	C2 - Total Cadmium	C3 - Total copper	C4 - Total Chromium	C5 - Total Mercury	C6 - Total Lead	C7 - Total Zinc
1	1.43	0.24	0.39	0.28	0.30	0.31	0.66
2	49.96	4.79	8.38	0.18	4.33	8.90	8.04
3	33.45	7.51	4.63	0.31	3.08	9.15	12.17
4	44.10	3.71	4.37	0.28	2.53	52.89	7.46
5	1.48	0.55	0.46	0.46	0.15	0.56	1.31

Step 3: Determining triangular functions and their functions

Replacing the values obtained from Table VI in the equations of the Whitenization functions, as an example for total arsenic functions equations are shown in (11) – (13). Its graphic representation is displayed in Fig. 5.

$$f_1^1(x) = \begin{cases} 1, x \leq 0.515 \\ \frac{1.00 - x}{0.485}, 0.515 \leq x \leq 1.00 \\ 0, x \geq 1.00 \end{cases} \quad (11)$$

$$f_1^2(x) = \begin{cases} \frac{x - 0.515}{0.485}, 0.515 \leq x \leq 1.00 \\ \frac{1.485 - x}{0.485}, 1.00 \leq x \leq 1.485 \\ 0, x \leq 0.515 \; U \; 1.485 \leq x \end{cases} \quad (12)$$

$$f_1^3(x) = \begin{cases} \frac{1.485 - x}{0.485}, 1.00 \leq x \leq 1.485 \\ 1, x \geq 1.485 \\ 0, x \leq 1.00 \end{cases} \quad (13)$$

Step 4: Determining weight for each criterion

The clustering weight (n_j) of each parameter was determined with Shanon's entropy. The following procedure was followed for this purpose.

1) Standardized parameter values from the Canadian standard, which are presented in Table VII.

2) The entropy (H_j) of each criterion (C_j) was calculated through (7). The results are shown in Table VIII.

3) Finally, the entropy weights wj were found according by using (9), and equated to the clustering weights n_i of each parameter. The values are presented in Table IX.

Step 5: Determining the clustering coefficient

The values of the clustering coefficients were calculated using (10). The results of the first 2 monitoring points are shown in Table X.

Step 6: Results using the maximum clustering coefficient

Finally, we calculate the value of $max_{1 \leq k \leq s}\{\sigma_i^k\} = \sigma_i^k$ for each grey class according to each monitoring point by adding a comparison by quality scale and with it we get Table XI.

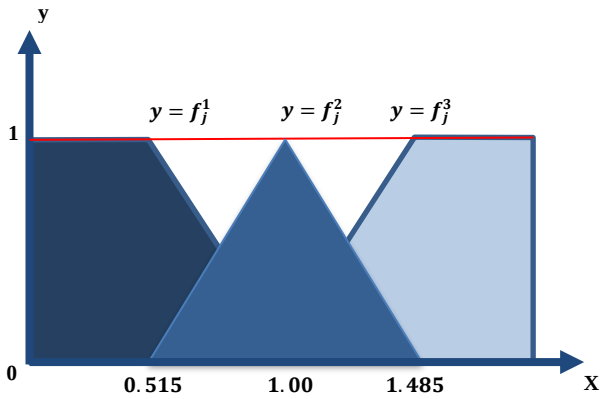


Fig. 5. Whitenization Functions for Total Arsenic.

TABLE VII. NORMALIZED VALUES FOR EACH PARAMETER

	C1	C2	C3	C4	C5	C6	C7
	Total arsenic	**total cadmium**	**total copper**	**total chromium**	**total mercury**	**total lead**	**total zinc**
Good	0.172	0.098	0.102	0.195	0.173	0.185	0.187
Moderate	0.333	0.333	0.333	0.333	0.333	0.333	0.333
Bad	0.495	0.569	0.564	0.471	0.494	0.482	0.479

TABLE VIII. DEGREE OF DIVERGENCE

Group	C1	C2	C3	C4	C5	C6	C7
H_j	0.93	0.83	0.84	0.95	0.93	0.94	0.94

TABLE IX. CLUSTERING WEIGHT OF EACH CRITERION

Group	C1	C2	C3	C4	C5	C6	C7
W_j	0.114	0.257	0.246	0.082	0.112	0.096	0.092

TABLE X. VALUES OF TRIANGULAR WHITENING FUNCTIONS (CTWF) FOR THE FIRST 2 MONITORING POINTS

P1	Criteria	C1	C2	C3	C4	C5	C6	C7	Results
	$f_1^1(x)$	0.000	1.000	0.882	1.000	1.000	1.000	0.771	0.114
	$f_1^2(x)$	0.108	0.000	0.118	0.000	0.000	0.000	0.229	0.021
	$f_1^3(x)$	0.892	0.000	0.000	0.000	0.000	0.000	0.000	0.050
P2	**Criteria**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**Results**
	$f_2^1(x)$	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.016
	$f_2^2(x)$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	$f_2^3(x)$	1.000	1.000	1.000	0.000	1.000	1.000	1.000	0.488

TABLE XI. COMPARISON OF SEDIMENT QUALITY ACCORDING TO THEIR MAX VALUE

Results	**Max**ik	**Quality**	**Quality scale**
P11	0.133	Good	Better sediment quality
P17	0.123	Good	
P22	0.123	Good	
P21	0.121	Good	
P19	0.120	Good	
P7	0.119	Good	
P20	0.118	Good	
P27	0.114	Good	
P1	0.114	Good	
P16	0.113	Good	
P9	0.111	Good	
P18	0.099	Good	
P5	0.088	Good	Good sediment quality
P30	0.080	Moderate	
P2	0.488	Bad	Lower sediment quality
P3	0.488	Bad	
P4	0.488	Bad	
P15	0.472	Bad	
P13	0.454	Bad	
P28	0.454	Bad	
P23	0.450	Bad	
p8	0.432	Bad	
P12	0.432	Bad	
P26	0.432	Bad	
P6	0.417	Bad	
P29	0.336	Bad	
P10	0.303	Bad	
P14	0.290	Bad	
P24	0.269	Bad	
P25	0.131	Bad	Poor sediment quality

V. RESULTS AND DISCUSSION

A. About the Case Study

It is observed in Table XII that 13 (43%) monitoring points resulted with the sediment quality good, while 1 monitoring point (3%) resulted with moderate quality and 16 (54%) monitoring points with poor quality. In addition, the comparison of quality level can be performed according to the maximum clustering coefficient (Max σ_i^k). In addition, we analyze according to each quality category into good, moderate and bad, respectively. For a better understanding further details regarding their location and differentiation are displayed in colors in Fig. 6.

- Good category: This means that no adverse biological effects are expected in these points and it is also observed that monitoring point P11 has the best water quality and point P5 the lowest quality within this category, this may happen because the points are located in tributaries to the main river and also the monitoring point P11 is more distant from the bridle path compared to the monitoring point P5 [2]. Likewise, the other points show good quality, possibly because they were sampled in tributaries of the main river.
- Moderate category: Means that at this point adverse biological effects are expected being monitoring point P30 the only point in this category, this may be because it is the farthest downstream point from the MTZ S.A.C mine compared to monitoring point P25 which is 100 m downstream of the mine [2].
- Bad category: It means that biological effects are frequently found in these points and also the monitoring point P25 presents the best sediment quality while point P2 presents the lowest sediment quality within this category, this could happen because the points are located within the main river where the mines discharge their effluents and also points P2, P3 and P4 are located downstream and closer to the Huinac mine, which is why they may have the lowest sediment quality compared to point P25, which has the best quality in the category since it is located 100 m downstream of the MTZ S.A.C. mine [2], but compared to points P2, P3 and P4, this mine has better control of its effluents.

Fig. 6. Results based on the Maximum Coefficient on the Case Study Map.

In relation to other studies, Chu & Tan [7] applied the Grey Clustering method to evaluate the surface sediment quality of the Jiangsu coastal ocean, and showed that more emphasis should be placed on the management of the surface tidal flat sediment environment, especially on the treatment of the pollution source to improve the sediment quality for the sustainable development of the coastal zone. In the water quality assessment conducted by Delgado et al., [12] in the Huallaga River basin, it was shown that of the 21 monitoring points analyzed all of these were found - according to the Pratti index - to be uncontaminated. A point to highlight from the previous study is that Shannon Entropy was used to determine the value of the weights, a feature that was not used in the study by Liping et al study [18] where the weights were determined using the arithmetic mean. Therefore, the present study is characterized by analyzing a topic that is little addressed from the perspectives of the Grey Clustering method, which is the analysis of sediments, but it also integrates a component that will provide greater objectivity and precision, Shannon Entropy, to determine the value of the weights.

B. Proposals for Poor Quality Points

According to the results and discussion on the case study, it is possible to evidence the contamination of sediment quality by heavy metals and that can be attributable to the effluents of the mining companies (mainly Huinac), being the monitoring points downstream of these, therefore, it is proposed that this incorporates constant monitoring of sediment quality and thus also take into account the natural state of the streams and rivers that may be mineralized areas, and if responsibility is warranted by this to proceed to better control their effluents by treatment plants.

C. About the Methodology

The Grey Clustering method is a useful methodology to analyze the environmental system with respect to environmental quality, since among the advantages of Fuzzy Sets it is considered that it can be applied when the internal mechanisms of the system are unknown or the concept to be measured is imprecise, in this case of the study area. It is a mathematical theory of uncertainty to model situations where traditional instruments do not lead to optimal results due to the existence of uncertainty problems [19]–[21].

Based on the study conducted, the Grey Clustering Analysis Methodology (GCA) was chosen as the method for Environmental Quality Assessment over the Delphi methods [22] and the AHP method [23], where the advantages and disadvantages involving these mentioned methodologies are shown in Table XII.

The Shannon entropy method allows determining the clustering weights (n_i) for each criterion in an objective way, without the need to consult experts, which makes this method a more efficient and integrated method for sediment quality assessment.

TABLE XII. COMPARATIVE TABLE OF ADVANTAGES AND DISADVANTAGES OF ENVIRONMENTAL QUALITY ASSESSMENT METHODS IN THE FIELD OF RESEARCH

Environmental Quality Assessment Method	Advantages	Disadvantages
Grey Clustering Methodology	-Defines a new and reduced set of indicators, linear combinations of the initial ones, which allows comparison at all levels evaluated. -High degree of flexibility, due to the membership functions, which allows it to include uncertainty in its analysis. its analysis. -It presents an objective weighting system. -It allows an integral analysis between the different thematic areas, by means of fuzzy inference rules. -Process a large set of data and reduce its dimensionality with a minimum loss of information [24].	The methodology has not been disseminated and its applications to environmental fields are recent.
Delphi Methodology	-Subjective experience and critical input. -Complex, large, multidisciplinary problems with considerable uncertainties. -Possibility of unexpected breakthroughs. -Particularly long-time frames. -Achieving consensus in areas of uncertainty or in situations lacking causality. -Focus on issues where multiple stakeholder groups are potentially involved. -Delphi studies can be relatively simple to design and flexible in the way those designs are combined [25].	-Causal models cannot be built or validated. -Opinions from a large group are required. -Deficiencies of the researcher or panel members may arise. -The researcher imposes his preconceived ideas on respondents.
Analytic Hierarchy Process (AHP) Methodology	-Allows a complex problem to be broken down and analyzed in parts. -Allows quantitative and qualitative criteria to be measured using a common scale. -It facilitates the understanding of the problem by the decision-maker or by those involved in the analysis stage. -It is easy to use and allows its solution to be complemented with mathematical optimization methods [26].	-Additional analysis is required to establish preconditions. Since minimum points of agreement among stakeholders regarding objectives, criteria, weights, etc. are required. -The analysis includes a certain degree of subjectivity. Because data from different sources are used, the analysis includes some degree of subjectivity.

VI. CONCLUSION

In the present quality study of the sediments of the surface water bodies of the upper basin of the Huarmey river, the 30 monitoring points of the basin could be classified, from which it was determined that 13 (43%) of the monitoring points presented good sediment quality, 1 (3%) moderate and 16 (54%) of poor quality. In the points that presented a good quality of sediments, it is attributed to the fact that they were located in tributaries of the main river, for the monitoring point that presented a moderate quality it can be attributed to being located at a significant distance from the MTZ SAC mine. Those that presented a poor quality of sediments, which corresponds to most of the points sampled, can be attributed to the fact that they are located within the main river, close to the activities of the mining industries (mainly Huinac) and the discharges of their effluents.

The study used the Grey Clustering method and Shanon's Entropy, as for the Grey Clustering methodology it turns out to be one of the most effective since it considers the uncertainty within the analysis; in addition, the analysis was enhanced with the Entropy methodology of Shannon, which allows developing the analysis process objectively, without the need for expert judgment. As a result of these, it allowed to generate classifications of the quality of sediments, which are pertinent in the application in the Peruvian context due to the lack of regulations regarding sediments.

This evaluation information obtained is relevant because it allows generating timely decision-making, in relation to the current context, by public entities and the central government with powers in the upper basin of the Huarmey River, Peru. And finally, the study serves as the basis for future research regarding the quality of sediments, in addition, to be complemented with subsequent studies of characterization of water, soil and sediments by natural formations in the Huarmey river basin.

REFERENCES

[1] INRENA and Administración Técnica del Distrito de Riego Casma - Huarmey, "Evaluación de los recursos hídricos en las cuencas de los ríos Casma, Culebras y Huarmey: estudio hidrológico en la cuenca del río Huarmey (Informe final)," Aut. Nac. del Agua, Dec. 2007.

[2] OEFA, "Informe del monitoreo ambiental de calidad de agua y sedimento, realizado del 20 al 30 de mayo de 2016, en la Cuenca alta del río Huarmey, distritos La Merced, Aija, Huacllan y Succha, provincia de Aija, departamento de Ancash," Lima, Informe Técnico W 2.66-201 6-0EFA/DE-SDCA, Dec. 2016.

[3] CCME, "Canadian Sediment Quality Guidelines for the Protection of Aquatic Life Freshwater," 2002.

[4] S. Liu and Y. Lin, "Grey Incidence and Evaluations," in Grey Systems: Theory and Applications, S. Liu and Y. Lin, Eds. Berlin, Heidelberg: Springer, 2011, pp. 51–105.

[5] A. Delgado, P. Montellanos, and J. Llave, "Air quality assessment in Lima city using the grey clustering method," 2018 IEEE Int. Conf. Autom. Congr. Chil. Assoc. Autom. Contorl(ICA-ACCA), 2018.

[6] Y. Zhang, J. Ni, J. Liu, and L. Jian, "Grey evaluation empirical study based on center-point triangular whitenization weight function of Jiangsu Province industrial technology innovation strategy alliance," Grey Syst. Theory Appl., vol. 4, no. 1, pp. 124–136, Jan. 2014, doi: 10.1108/gs-11-2013-0027.

[7] K. J. Chu and M. Tan, "Assessment of Sediment Quality in Jiangsu Coastal Ocean Based on Grey Clustering Method," Applied Mechanics and Materials, 2014.

[8] A. Delgado and I. Romero, "Environmental conflict analysis on a hydrocarbon exploration project using the Shannon entropy," in Proceedings of the 2017 Electronic Congress, E-CON UNI 2017, 2018-January, pp. 1-4, doi: 10.1109/ECON.2017.8247309.

[9] X.-S. He, Q.-W. Fan, and X.-S. Yang, "Chapter 6 - Probability theory for analyzing nature-inspired algorithms," in Nature-Inspired Computation

and Swarm Intelligence, X.-S. Yang, Ed. Academic Press, 2020, pp. 77–88.

[10] A. Delgado, A. Aguirre, E. Palomino, and G. Salazar, "Applying triangular whitenization weight functions to assess water quality of main affluents of Rimac river," in Proceedings of the 2017 Electronic Congress, E-CON UNI 2017, 2018-January, pp. 1-4, doi: 10.1109/ECON.2017.8247308.

[11] A. Delgado and I. Romero, "Social impact assessment on a hydrocarbon proyect using triangular whitenization weight functions," in CACIDI 2016 - Congreso Aergentino de Ciencias de la Informatica y Desarrollos de Investigacion, 7785998, doi: 10.1109/CACIDI.2016.7785998.

[12] A. Delgado, J. Vidal, J. Castro, J. Felix, and J. Saenz, "Assessment of Surface Water Quality on the Upper Watershed of Huallaga River, in Peru, using Grey Systems and Shannon Entropy," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, 2020, doi: 10.14569/IJACSA.2020.0111156.

[13] ANA, "Resultados del monitoreo participativo de la calidad de agua superficial de la cuenca del río y mar de Huarmey (realizado del 28 de octubre al 04 de noviembre de 2015): Informe técnico," Aut. Nac. del Agua, Dec. 2015.

[14] V. Bax, W. Francesconi, and A. Delgado, "Land-use conflicts between biodiversity conservation and extractive industries in the Peruvian Andes," Journal of Environmental Management, vol. 232, pp. 1028–1036, Feb. 2019, doi: 10.1016/j.jenvman.2018.12.016.

[15] S. Liu, Grey Systems: Theory and Applications, vol. 68. Berlin, Heidelberg: Springer, 2011.

[16] S. Liu, Y. Yang, and J. Forrest, Grey Clustering Evaluation Models. 2017.

[17] A. Shemshadi, H. Shirazi, M. Toreihi, and M. J. Tarokh, "A fuzzy VIKOR method for supplier selection based on entropy measure for objective weighting," Expert Syst. Appl., vol. 38, no. 10, pp. 12160–12167, 2011.

[18] W. Liping, L. Kunrong, and Z. Weibo, "Application of Grey Clustering method for water quality evaluation in fenchuan River Yan'an Baota Area," in 2011 International Symposium on Water Resource and Environmental Protection, 2011.

[19] R. E. Bellman and L. A. Zadeh, "Decision-Making in a Fuzzy Environment," Manage. Sci., vol. 17, no. 4, p. B-141, Dec. 1970, doi: 10.1287/mnsc.17.4.B141.

[20] L. A. Zadeh, "Fuzzy sets," Inf. Control, vol. 8, no. 3, pp. 338–353, Jun. 1965, doi: 10.1016/S0019-9958(65)90241-X.

[21] A. Delgado, F. Gil, J. Chullunquía, T. Valdivia, and C. Carbajal, "Water Quality Analysis in Mantaro River, Peru, Before and After the Tailing's Accident Using the Grey Clustering Method," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 11, no. 3, pp. 917–922, 2021, doi: 10.18517/IJASEIT.11.3.11928.

[22] J. R. Avella, "Delphi Panels: Research Design, Procedures, Advantages, and Challenges," Int. J. Dr. Stud., vol. 11, pp. 305–321, Sep. 2016.

[23] E. Vergara Maldonado, "Pautas para la selección de las técnicas AHP, PROMETHEE y ábaco de Régnier modificado," 2010.

[24] A. Delgado, J. A. G. Achata, J. A. B. Valdivia, J. C. J. S. Montes, and C. Carbajal, "Grey Clustering Method for Water Quality Assessment to Determine the Impact of Mining Company, Peru," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 4, pp. 557–564, 2021, doi: 10.14569/IJACSA.2021.0120471.

[25] A. Delgado and H. Flor, "Selection of the best air purifier system to urban houses using AHP," in 2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies, CHILECON 2017 - Proceedings, 2017-January, pp. 1-4, doi: 10.1109/CHILECON.2017.8229622.

[26] H. Li, Y. Fu, W. Tang, and W. Yang, "The application of grey clustering analysis on social impact assessment of natural forest protection project," in Proceedings 2010 IEEE International Conference on Information Theory and Information Security, ICITIS 2010, 2010, pp. 776–780, doi: 10.1109/ICITIS.2010.5689689.

A Hybrid Intrusion Detection Model for Identification of Threats in Internet of Things Environment

Nsikak Pius Owoh[1]
Department of Cyber Security, School of Information and Communication Technology, Federal University of Technology, Owerri, Imo State, Nigeria

Manmeet Mahinderjit Singh[2], Zarul Fitril Zaaba[3]
School of Computer Sciences
Universiti Sains Malaysia, 11800 USM
Penang, Malaysia

***Abstract*—Internet of Things (IoT) has transcended from its application in traditional sensing networks such as wireless sensing and radio frequency identification to life-changing and critical applications. However, IoT networks are still vulnerable to threats, attacks, intrusions, and other malicious activities. Intrusion Detection Systems (IDS) that employ unsupervised learning techniques are used to secure sensitive data transmitted on IoT networks and preserve privacy. This paper proposes a hybrid model for intrusion detection that relies on a dimension reduction algorithm, an unsupervised learning algorithm, and a classifier. The proposed model employs Principal Component Analysis (PCA) to reduce the number of features in a dataset. The K-means algorithm generates clusters that serve as class labels for the Support Vector Machine (SVM) classifier. Experimental results using the NSL-KDD and the UNSW-NB15 datasets justify the effectiveness of our proposed model in detecting malicious activities in IoT networks. The proposed model, when trained, identifies benign and malicious behaviours using an unlabelled dataset.**

***Keywords*—Internet of things; intrusion detection system; k-means; principal component analysis; support vector machine**

I. Introduction

Internet of Things (IoT) is a self-organizing and adaptive network that interconnects uniquely identifiable "Things" to the internet via communication protocols [1]. The "Things" (also known as devices) are capable of sensing data from humans and the environment. IoT devices collect and sometimes store information that can be accessed pervasively and at any time. The Internet of Things (IoT) is a proliferating technology that offers many advantages in many areas of life [2]. However, the IoT is faced with several information security vulnerabilities and threats. Considering the intrinsic computational limitations of IoT devices and their vulnerabilities and the increasing rate of unauthorized access to these devices [3], IoT risks increase exponentially. Threats to the IoT network are similar to a traditional network, which threatens confidentiality, integrity, and availability. Such threats, when exploited, may lead to eavesdropping, data leakage/loss, and denial-of-service attacks [4].

The connection of IoT devices to the internet through vulnerable networks such as 6LoWPAN and IPv6 makes them susceptible to various intrusions. Nevertheless, these intrusions can be detected by intrusion detection systems (IDS) [5]. Intrusion detection systems (IDS) can identify internal and external attacks [6]. Though a post-active security measure, Intrusion detection systems can identify attacks in networks using adaptive network detection algorithms and act as a multilayer security mechanism to cryptographic solutions in a network. The different types of IDS are signature-based (misuse), anomaly-based, and specification-based detection systems.

In signature-based detection systems, predefined attack patterns are modelled and stored in a database. IDSs of this type accurately detect known intrusions. Also, low false-positive rates and minimal computation overhead are experienced with signature-based IDS. However, they ignore unknown intrusions, making them ineffective in detecting network attacks [7]. On the other hand, anomaly-based detection systems employ statistical or machine learning approaches to identify unusual (possible threats) from normal behaviours in network traffic or system activities. Detection, in this case, is based on the features and labels in each data. Detection rates are higher with the anomaly-based system since they can detect new and unseen attacks. Nevertheless, increased computation overhead and false alarms are some drawbacks of anomaly-based IDSs [7]. Specification-based detection systems are like anomaly-based detection systems but require involvement of users in obtaining valid network traffic to develop a normal behaviour model [5].

A significant problem with anomaly detection systems is that they require unlabelled data. This approach is challenging because of the difficulty of acquiring large datasets that are labelled as "normal" or "malicious." Detecting anomalies in IoT becomes even more complicated when applied to high-dimensional data with large features. High-dimension datasets often reduce the accuracy of anomaly detection systems due to the presence of irrelevant features, exponential search space, and data bias [8]. To this end, there is a need for a detection system capable of detecting threats (such as anomalies and attacks) in an IoT network with high accuracy using unlabelled data. Achieving the proposed high accuracy would require the removal of irrelevant and redundant data through feature reduction.

This paper proposes a hybrid intrusion detection system for IoT, which relies on PCA for dimension reduction, K-means for threats clustering, and SVM for anomaly classification. To the best of our knowledge, this is the first paper to apply these algorithms to detect anomalies in both unlabelled and labelled datasets. The contributions of this paper are summarized as follows:

1) To develop an intrusion detection model that performs feature reduction and anomaly detection in unlabelled and labelled datasets.

2) To build a classification model using the generated cluster labels from the unsupervised learning phase.

3) To evaluate the performance of the anomaly detection model when trained with different number of clusters and features.

The rest of this paper is structured as follows: Section II presents a review of related works on attacks in the IoT and intrusion detection systems used in identifying such threats. In Section III, we present our proposed hybrid intrusion detection model. Furthermore, datasets and methods used for data clustering and classifier training are also discussed in this section. The hybrid model results, including feature reduction, data clustering, and binary and multi-class classification, are shown in Section IV. In Section V, we discuss obtained results and conclude the paper in Section VI.

II. Related Work

Akin to the desired security requirements in traditional networks, IoT networks need to ensure confidentiality, integrity, availability, non-repudiation, and privacy. It is worthy to note that, in IoT networks, a breach in any of these requirements can be life-threatening because of its applicability and peculiarity [9]. The availability of sensitive data in IoT devices makes them an attractive target for cyber-attacks. Threats on IoT networks are increasing massively, especially as IoT devices can automatically join and leave sensor networks [10]. Another reason for the increasing number of successful IoT attacks is their limited resources (power, storage, and computational capabilities). These constraints make it challenging to implement sophisticated security and privacy mechanisms [11].

A. Attacks on the Internet of Things (IoT)

There are several possible attacks on IoT networks. Among these attacks, distributed denial of service (DDoS) attack has grown to become one of the most severe. Even so, its detection and prevention have also been a security challenge. DDoS exploits compromised devices (zombie or botnet) to flood IoT devices or communication channels with bogus requests and eventually rendering their services unavailable to legitimate users. Solving this problem has brought about several proposed solutions in different applications and networks. However, detecting and preventing DDoS attacks is tasking due to the difficulty of differentiating attack packets from legitimate ones. Even more troubling is that DDoS attacks can be perpetuated over any of the four layers of the IoT [11]. In what follows, we enumerate some attacks at each layer of the IoT.

The perception layer, also referred to as the sensing layer, handles the data gathering from users and the environment. It employs technologies such as wireless sensor networks (WSNs), radio frequency identification (RFID), mobile crowdsensing (MCS), and micro-electro-mechanical (MEMS) [12]. Eavesdropping, tag cloning, spoofing, unauthorized access, and Radio Frequency jamming are some of the attacks in this layer. These attacks compromise devices by affecting vital architectural components of the IoT system. Memory corruption and misconfiguration of IP addresses are reasons for these attacks [13].

The network layer transmits sensor data between the information processing system and sensor devices using communication infrastructures such as wired and wireless connections. Attacks in the network layer include sinkhole, Man-In-The-Middle, Sybil, and DDoS attacks [14]. In the network attack, an adversary targets intercommunication among devices by causing latency or dropping sent messages. Such attacks destroy computational processes within the IoT configuration systems. The middleware layer guarantees and oversees services needed by applications or clients. Furthermore, service management and database connection are handled in this layer. DoS and unauthorized access are possible attacks in this layer [14].

The application layer consists of interaction techniques of users and applications, and it conveys application services to users. Attacks such as phishing, sniffing, code injection, and DoS are possible threats in the application layer. These attacks compromise system applications (Mobile and Web applications) [13]. Table I summarizes the different attack types at the different layers of the IoT.

B. Intrusion Detection Systems in the Internet of Things (IoT)

Predicting threats or detecting them at their initial stages effectively prevents successful attacks on IoT devices [15]. Interestingly, several cybersecurity tasks can be performed using machine learning. These tasks include anomaly detection, spam filtering, user monitoring, risk analysis, and zero-day exploit identification [16]. Machine learning algorithms have been used widely in developing intrusion detection systems for IoT networks. Its adoption in this area is justified in its ability to detect anomalies in network traffic. Based on their properties, data usage patterns, and learning style, machine learning algorithms are classified into three groups: supervised, unsupervised, and semi-supervised algorithms [17]. The algorithm is trained using training data (labelled input) in supervised learning, often called ground truth [18].

TABLE I. Attack Classification in the different Layers of the Internet of Things

	Perception layer	**Network layer**	**Middleware layer**	**Application layer**
Components	GPS, RFID tags, RFID reader/writers, Barcodes, BLE devices	WLAN, Social networks, WSNs, Cloud network	Database, Service APIs, Service management	Interface, Smart applications
Possible Attacks	Code injection, Noisy data, Unauthorized access	Routing attacks, DoS, DDoS, Network congestion	Spoofing, DoS, Malicious information, Unauthorized access, Data manipulatio	Phishing, Misconfigurations, Code injection.

On the other hand, unsupervised learning algorithms do not require labels in the training datasets as they can infer from the input data. They can reveal the hidden structure and distribution in data which provides more information about the data. A typical example of this category of algorithms is clustering (K-means). With clustering, structures or patterns in an unlabelled dataset are identified by grouping the data of interest into *k* number of clusters [18].

The work proposed by Li et al. [19] presents an approach that employs deep belief networks and Autoencoder for intrusion detection. The authors evaluated their proposed system using the KDD-CUPP 99 dataset. The authors' results from the 2000 records show that the proposed hybrid system can accurately detect anomalies in data but takes too long to pre-process data. Similarly, an unsupervised hybrid architecture for anomaly detection in large-scale high-dimensional is proposed by Erfani, Rajasegarar [8]. This work also evaluated the performance of deep belief networks against one-class SVMs when detecting anomalies in high-dimensional data. The DBN in the proposed system extracts only relevant features in the dataset, while the ISVM is trained using the extracted features. However, the datasets used for the evaluation of the proposed model do not ideally simulate real-world scenarios. In Nskh, Varma [20], a dimension reduction and classifier model relies on the KDD Cup 99 dataset is proposed. The model employs Principal Component Analysis for dimension reduction and Support Vector Machine for attack classification. However, the model is non-trivial, and the computing complexity of the model is not provided.

Meanwhile, Pajouh, Javidan [21] proposed a two-layer dimension reduction and two-tier classification model for intrusion detection in IoT. The model uses Principal Component Analysis and Linear Discriminant Analysis for feature extraction, while Naïve Bayes and K-nearest Neighbour algorithms are used for attack classification. The authors show that the model is trivial as it uses fewer computing and memory resources. Zhao, Li [22] present a model for anomaly-based intrusion detection in IoT. The model is based on PCA for dimension reduction and SoftMax Regression for classification. Low computing complexity was obtained with the reduced dimension, while accurate detection was accomplished with small training sets. Accuracy results obtained from the SoftMax regression model are 84.9%, 84.4%, and 84.4% for 3, 6, and 10 features, respectively. SVM classifier, on the other hand, produced slightly better results when tested with similar features.

A malware detection model for IoT devices that employ KNN and Random Forest classifiers were developed in Narudin, Feizollah [23]. KNN used in the proposed system allocates network traffic to a class with the most objects among its K-nearest neighbours. On the other hand, the random forest uses the labelled network traffic from the KNN classifiers to develop decision trees that identify malware in network traffic. Obtained results from the experiments performed with the MalGenome dataset show a true positive rate (TPR) of 99.7% and 99.9% for KNN and Random Forest, respectively. A Host-based Intrusion Detection and Mitigation framework for home-based IoT is proposed in Nobakht, Sivaraman [24]. The framework uses software-defined networks (SDN) and machine learning techniques to ensure security in IoT devices. The authors of this work also proposed an attack simulation model that collects data then distinguishes malicious actions from normal activities.

A machine learning framework that detects DDoS attacks in the IoT by collecting data, extracting its features, and performing binary classification is shown in Doshi, Apthorpe [25]. The proposed framework has four steps: traffic capture, packet grouping, feature extraction, and binary classification. The authors also evaluated several classifiers, including support vector machine, K-Nearest Neighbour (KNN), Decision Trees (DT), Neural Networks (NN), and Random Forests. Furthermore, Abeshu and Chilamkurti (2018) proposed an intrusion detection system that uses deep learning. The proposed IDS can detect zero-day attacks in a fog-to-things computing environment using the NSL-KDD dataset for evaluation. The IDS model uses 150 neurons in the first layer, 120 in the second, 50 in the third, and a SoftMax layer in the last layer. Also, the model was compared with shallow models, and an accuracy score of 99.20% was obtained with a FAR of 0.85% against a FAR of 6.57% in shallow models. However, detecting attack types such as probing, DoS and U2R were omitted in the presented work.

Few works have been proposed on anomaly detection with the capability of dimension reduction and attack classification. These works mostly rely on labelled data for accurate attack classification in IoT networks. Zhao, Li [22] presented an anomaly detection system that employs PCA and SoftMax regression algorithms. However, the proposed method is based on a supervised learning model and only functions as a binary classifier that detects only normal or malicious attacks, leaving out other attack vectors. Furthermore, the authors evaluated their proposed system on the KDD-CUP 99 dataset, which contains old records. Considering this, we propose an anomaly detection system that employs an unsupervised learning technique with a classifier capable of detecting up to four classes of attacks present in the NSL-KDD dataset. We also evaluate our proposed hybrid model using the UNSW-NB15 dataset, a more recent dataset with new attack activities.

III. Methodology

This section presents the architecture of the proposed model, including the datasets and techniques employed for the detection of anomalies in the IoT.

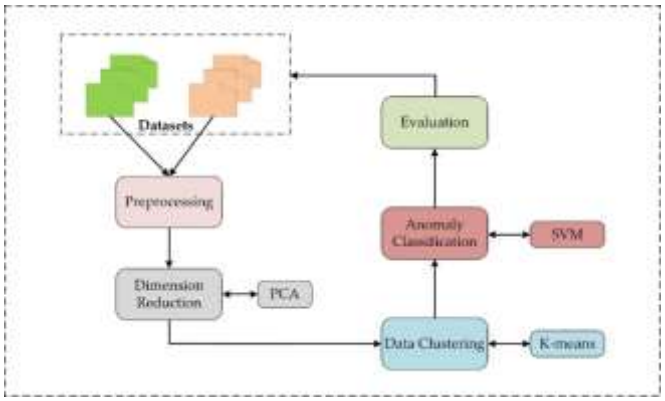


Fig. 1. Architecture of the Proposed Hybrid Detection Model.

A. Architecture

The architecture for our proposed model, as shown in Fig. 1, consists of three parts: dimension reduction, data clustering, and anomaly classification. The model is implemented in Python using available libraries such as SciKit-Learn, Pandas, Numpy [26], and Matplotlib [27]. The experiments which involved the implementation of all three components of the proposed model (i.e., PCA, K-means, and SVM) were performed on an Intel(R) Core (TM) i7500U CPU@2.70GHz laptop with a 12 GB RAM and running Windows 10 Home edition.

1) Dataset: The first dataset used in the proposed model is the NSL-KDD dataset [28]. The dataset is commonly used for the simulation of anomaly detection systems and models. Most of the inherent issues with the earlier KDD-CUP 99 dataset are resolved in the NSL-KDD dataset, and it is a preferred choice for baseline evaluation of IDSs. The dataset consists of training and testing datasets with 41 features: duration, protocol, service, flag, source bytes, destination bytes, and normal/attack labels. Furthermore, the dataset consists of 125,973 records for the training data and 22,544 records for the test data. The labels in the dataset can be categorized into four attack classes, which are Denial of Service (DoS) attack, User to Root (U2R) attack, Probing attack, and Remote to Local (R2L) attack. Table II presents the details of these attack classes.

a) Probing Attack: This attack involves scanning IoT targets and serves as a starting point for other attacks. Scanning programs are used to discover vulnerabilities in IoT applications. Tools such as mscan and saint can be used for this purpose.

b) Remote-to-Local (R2L): After a successful scan, the attacker may employ a remote-to-local ((R2L) attack to access the local system from remote ports, thereby escalating system privileges. Examples of this attack include *ftp-write*, *guest-exploit*, which either exploit poorly configured security policies or network programs.

c) User to Root (U2R) Attack: This attack originates from the R2L attacks and exploits unsecured programs running as roots. This attack-type leads to a buffer overflow caused by ffbconfig, fdformat, and eject.

d) Denial of Service (DoS) Attack: A denial-of-service (DoS) attack is successfully launched on a target machine or device by flooding such device with overloaded requests to stop legitimate requests from getting access to the device(s) [29].

Though the NSL-KDD dataset [28] solved most issues, such as data imbalance among normal and malicious records associated with the earlier KDDCUP dataset, the NSL-KDD dataset still does not depict present-day attack activities. To ascertain the effectiveness of our proposed hybrid model on recent malicious activities, we also evaluate the proposed model on the UNSW-NB15 dataset [30]. The UNSW-NB15 dataset consists of 49 features, including the class label. Table III shows the different features and categories in the dataset [30].

TABLE II. ATTACK CLASSIFICATION IN THE NSL-KDD DATASET

Probing	Remote to Local (R2L)	User to Root (U2R)	Denial of Service (DoS)
ipsweep	ftp_write	buffer_overflow	back
nmap	guess_passwd	Loadmodule	land
portsweep	imap	Perl	neptune
satan	Multihop, Phf	Rootkit	Pod, smurf
	spyware_client		teardrop
	spyware _master		

TABLE III. RECORD DISTRIBUTION OF THE UNSW-NB15 DATASET

Type	No. of Records	Description
Normal	2,218,761	The name of each attack category. In this data set, nine categories (e.g., Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms)
Fuzzers	24,246	0 for normal and 1 for attack records
Analysis	2,677	It contains different attacks of the port scan, spam, and HTML file penetrations.
Backdoors	2,329	A technique in which a system security mechanism is bypassed stealthily to access a computer or its data.
DoS	16,353	A malicious attempt to make a server or a network resource unavailable to users, usually by temporarily interrupting or suspending the services of a host connected to the internet
Exploits	44,525	The attacker knows of a security problem within an operating system or a piece of software and leverages that knowledge by exploiting the vulnerability.
Generic	215,481	A technique that works against all block-ciphers (with a given block and key size) without considering the block-cipher structure.
Reconnaissance	13,987	It contains all Strikes that can simulate attacks that gather information
Shellcode	1,511	A small piece of code is used as the payload in the exploitation of software vulnerability.
Worms	174	The attacker replicates itself to spread to other computers. Often, it uses a computer network to spread itself, relying on security failures on the target computer to access it.

2) Data pre-processing: For machine learning algorithms to perform optimally, feature scaling is necessary since the range of values may vary in the input data. The range of data of some features in the NSL-KDD and UNSW-NB15 datasets is enormous, and such dimensions determine the distance variance; hence the need for data normalization. Similar to the work proposed by Zhao, Li [22], we adopted the Min-Max normalization method to ensure that all the data values come under the range of 0 and1. This approach is presented mathematically in equation 1.

$$A_j^{(i)} = \frac{A_j^{(i)} - min}{max - min} \quad (1)$$

3) Dimension reduction: Dimension reduction was chosen in the proposed model to solve the problems faced with high dimensional data, typical with anomaly-based datasets such as the NSL-KDD and UNSW-NB15 datasets [28, 30]. The high dimensional data contain redundant and irrelevant features, which degrade the performance of the detection model. In the proposed model, the 41 features present in the NSL-KDD dataset and the 49 features in the UNSW-NB15 dataset are reduced using PCA to 3, 6, and 10 features. To reduce the dimension of the features, the covariance matrix is calculated to obtain the matrix for projection using equation 2 [22]:

$$\sum = \frac{1}{m}\sum_{i=1}^{m}\left(A^{(i)}\right)\left(A^{(i)}\right)^T \quad (2)$$

Three different components were used to evaluate the proposed model. When developing the model with three features from the dataset, 75% were retained, while 89% were kept from the original data when the features were reduced to six. When the features were reduced to10, 96% of the data were retained from the original dataset. Furthermore, categorical features were encoded into discrete features via the 1-to-n encoding method, and the class labels were dropped before clustering was performed.

4) Data clustering: Clustering algorithms search for groups of similar data vectors in a dataset. This unsupervised approach does not require labelled data to ascertain which class or cluster data inputs should be assigned. It is also a non-parametric technique requiring no prior knowledge of data parameters [31]. The K-means algorithm [32] is a clustering algorithm based on the similarity measure between data inputs. In our hybrid model, the algorithm was employed to accept both random observations N and a parameter showing the number of clusters (i.e., their centroids) $C_i \leq i \leq k$. An observation is assigned to a cluster in each iteration using the shortest distance between the observation and the centroids. The algorithm reassigns the centroids by reducing the mean distance of all observations in the cluster to its centroids after each iteration. The algorithm converges when the position of the centroids no longer changes. The aim is to find a set of k cluster centres, represented as $\{C_1, \dots, C_k\}$ such that there is minimization in the distance between data points and their nearest centre. Assigning data points to a cluster centre requires a set of binary variables $\lambda_{nk} \in \{0,1\}$, such that if cluster centre C_k contains data point a_n, then $\lambda_{nk} = 1$ as captured in the algorithm in Table IV. Two different experiments were conducted using the K-means algorithm. The first involved generating two clusters (k=2), representing normal and malicious. The second generated four clusters (k=4), representing normal data and the different attack types in the NSL-KDD dataset (Normal, DoS, Probing, U2R, and R2L). Meanwhile, for the UNSW-NB15 dataset, only two clusters are generated (i.e., normal and malicious).

TABLE IV. THE ALGORITHM FOR THE PROPOSED MODEL

Algorithm: Dimension Reduction and Data Clustering
Inputs: Unlabelled dataset $\{a_1, \dots, a_N\}$; Number of clusters N = Number of samples X, Y
Output: Principal components , cluster centres $\{C_k\}$ and assigned data points $\{\lambda_{nk}\}$

Set $P_k = \{3,6,10\}$ *(Reduction)*
Initialize
$C_k = \{2,4\}$ (Number of clusters)
for $n = 1$ to N **do**
For $K=1$ to K **do**
$if k = (C_i - a_i, \dots, C_n - a_n)$ then
$\lambda_{nk} = 0$
Else
$\lambda_{nk} = 1$
end if
end for
end for
For $n=1$ to K do
$\lambda_{nk} = 4$
end for
$\lambda_{nk} \mid C_k$ converges

5) Anomaly classification: The proposed model uses the Support Vector Machine algorithm for anomaly classification. The SVM is a supervised learning model used for data classification, regression, and outlier detection. SVM, which is most suitable for non-linear data used in this paper, can be represented formally in equation 3 [33].

$$\Theta = \sum_{i=1}^{m} \alpha_i c_i x_i \quad (3)$$

Where x is the given input, c is the Class label, α is the LaGrange multiplier, and $\Theta$is the weight vector.

In this paper, the class labels used by the SVM classifier are cluster labels generated from the K-means algorithm. The classification task incorporates both binary and multi-class classification. The binary classification trains the classifier to predict unseen data from IoT network traffic as either normal or malicious. Meanwhile, the multi-class classification implements a more detailed classification, where the classifier was trained to predict unseen data into the normal, DoS, Probing, U2R.R2L classes. The U2R.R2L class is a merged class due to its low occurrence as captured in the NSL-KDD dataset. For the UNSW-NB15 dataset, only binary classification is performed.

IV. RESULTS AND DISCUSSION

The first results obtained are from the data clustering task. Fig. 2 shows the normal and malicious clusters when k is set to 2. Fig. 3, on the other hand, displays the four different clusters when k is set to four. These clusters illustrate the similarity of data points in the same group (normal traffic data) from the malicious cluster.

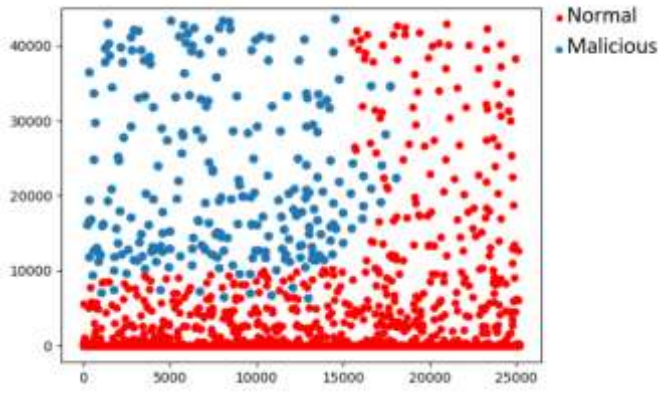


Fig. 2. Data Clusters when k=2.

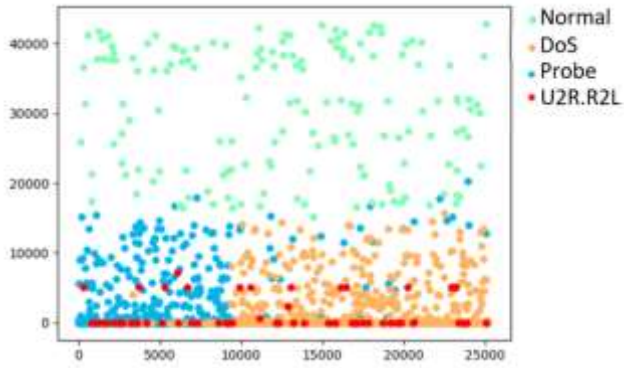


Fig. 3. Data Clusters when k=4.

As stated earlier in this paper, the generated clusters from the first phase of the detection model are used to train the SVM classifier. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are performance indicators used to evaluate the proposed anomaly detection model to ascertain its accuracy, precision, and recall, as shown in equation 4 to 6, respectively). TP shows that normal behaviours are classified correctly as normal behaviours; TN shows that malicious activities are classified correctly as malicious. FP demonstrates that malicious activities are incorrectly classified as normal behaviours, while FN shows that normal behaviours are incorrectly classified as malicious activities. In addition to the above performance metrics, the Detection Rate (DR) of the classifier in identifying malicious activities was also evaluated using equation 7. False Alarm Rate (FAR) (incorrectly detecting normal behaviour as malicious activities) was also examined using equation 8. The classification summaries for the NSL-KDD and the UNSW-NB15 datasets are presented in Table V.

$$Accuracy = \frac{TN+TP}{FN+FP+FN+TP} \quad (4)$$

$$Precision = \frac{TsP}{FP+TP} \quad (5)$$

$$Recall = \frac{TP}{FN+TP} \quad (6)$$

$$DR = \frac{TP}{(FN+TP)} \quad (7)$$

$$FAR = \frac{FP}{(FP+TN)} \quad (8)$$

When two clusters were used as class labels for the SVM classifier, accuracy scores of 97.82%, 97.58%, and 97.01% were obtained for the 3, 6, and 10 features NSL-KDD dataset as depicted in Table VI. There was no significant difference in the accuracy scores recorded across the different number of features. However, DR was remarkably higher with three features than with six features. Nevertheless, FAR was significantly lower with six features with 0.95% (less than one per cent) against 2.81% observed with three features. In this experiment, data were classified either as normal or malicious. This result proves that high-dimension features do not necessarily equal high accuracy and detection rate in datasets used for the experiment.

Furthermore, with four clusters employed as class labels (normal, DoS, Probing, U2R.R2L) for the SVM classifier, accuracy scores of 93.96%, 95.03%, and 91.79% were recorded for features reduced to 3, 6, and 10, respectively. These accuracy scores are lower compared to those observed with two class labels. The results show that the model performs better when predicting data into a binary class. However, reasonably high detection rates were recorded when detecting data as normal, DoS, Probing, U2R.R2L. The performance of the model based on accuracy, precision, recall, DR, and FAR when trained with two and four clusters is presented in Fig. 4.

TABLE V. CLASSIFICATION DISTRIBUTION FROM SVM CLASSIFIER

NSL-KDD Dataset					
3 Features		**6 Features**		**10 Features**	
TN=1730	FP=47	TN=5520	FP=50	TN=5525	FP=53
FN=87	TP=4434	FN=99	TP=629	FN=94	TP=626
UNSW-NB15 Dataset					
3 Features		**6 Features**		**10 Features**	
TN=19844	FP=2	TN=19840	FP=6	TN=23941	FP=3
FN=5	TP=23995	FN=3	TP=23997	FN=4	TP=19888

TABLE VI. THE PERFORMANCE OF SVM CLASSIFIER WITH DIFFERENT NUMBER OF CLASSES AND FEATURES ON THE NSL-KDD DATASET

K=2				**K=4**		
Metrics	**3 Features**	**6 Features**	**10 Features**	**3 Features**	**6 Features**	**10 Features**
Variance	75%	89%	96%	75%	89%	96%
Accuracy	97.82%	97.58%	97.01%	93.96%	95.03%	91.79%
Precision	98.88%	92.19%	94.55%	92.77%	94.09%	87.82%
Recall	98.07%	86.34%	90.76%	91.07%	93.30%	83.95%
Detection Rate	98.07%	86.34%	90.76%	96.13%	97.54%	94.21%
False Alarm Rate	2.81%	0.95%	1.36%	3.24%	2.93%	3.79%

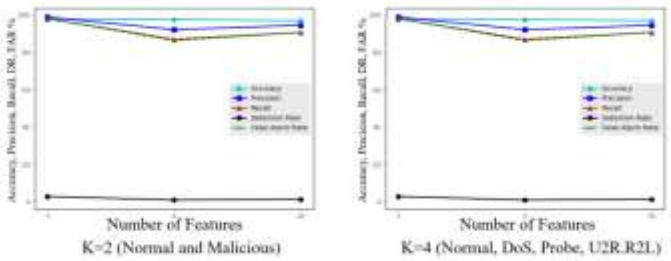


Fig. 4. Performance of Proposed Model on Two different Clusters.

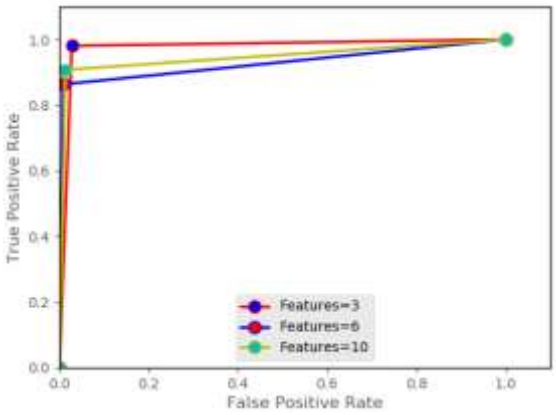


Fig. 5. Data Clusters when k=4.

Fig. 5 shows a ROC curve for the three experiments, where features were reduced to 3, 6, and 10. The trained classifier obtained from the cluster labels was applied to the NSL-KDD dataset, which contained different features. We then analysed how accurately the model detects anomalies from normal traffic data because the initial process was achieved using unsupervised learning.

TABLE VII. THE PERFORMANCE OF THE SVM CLASSIFIER ON THE UNSW-NB15 DATASET

UNSW-NB15 (K=2)			
Metrics	**3 Features**	**6 Features**	**10 Features**
Variance	**75%**	**89%**	**96%**
Accuracy	99.95%	99.98%	99.99%
Precision	99.93%	99.97%	99.98%
Recall	99.92%	99.95%	99.97%
Detection Rate	99.98%	99.98%	99.99%
False Alarm Rate	0.4%	0.5%	0.42%

Similarly, Table VII presents accuracy results from the evaluation of the proposed model on the UNSW-NB15 dataset. An accuracy of 99% was obtained when the model was tested using 3, 6, and 10 features. These results show the effectiveness of the proposed model in detecting malicious activities in recent datasets.

Apart from evaluating the classification accuracy, precision, DR, and FAR of the proposed intrusion detection model, we also identified features selected by the PCA algorithm after feature dimension reduction. These features were chosen from the 41 available features in the NSL-KDD dataset. Table VIII shows the most important features after dimension reduction to 3, 6, and 10. The features are presented in descending order of importance, and the top three features are DST_HOST_SRV_SERROR_RATE, SRV_RERROR_ RATE, and DST_HOST_SAME_SRC_PORT_RATE.

On the other hand, Table IX presents the most relevant features in the INSW-NB15 dataset after dimension reduction using the proposed model. The model also captures the associated weight of each feature. To accurately compare results obtained from our model with an earlier work presented in Zhao, Li [22], we adopted the same number of dimensions after dimension reduction (i.e., best 3, 6, and 10 dimensions of the singular vector). With a variance of 75%, dimensions were reduced to 3, 6 dimensions were obtained with a variance of 89%, while a variance that retained 96% of the data produced ten dimensions from the available 41 features. The two experiments conducted in this paper are based on the reduced features and are used to generate clusters (i.e., k-2 and k=4), which served as cluster labels for the classifier. A comparison of the results presented in Zhao, Li [22] shows that our proposed model performs better accuracy using 3 and 6 features, as demonstrated in Fig. 6.

TABLE VIII. THE MOST RELEVANT FEATURES IN THE NSL-KDD DATASET

PCs =3		
S/N	**Features**	**Weights**
1	DST_HOST_SRV_SERROR_RATE	0.508
2	SRV_RERROR_RATE	0.212
3	DST_HOST_SAME_SRC_PORT_RATE	0.068
PCs =6		
S/N	**Features**	**Weights**
1	DST_HOST_SRV_SERROR_RATE	0.508
2	SRV_RERROR_RATE	0.212
3	DST_HOST_SAME_SRC_PORT_RATE	0.068
4	DST_HOST_COUNT	0.052
5	DST_HOST_SAME_SRV_RATE	0.043
6	SRV_DIFF_HOST_RATE	0.021
PCs =10		
S/N	**Features**	**Weights**
1	DST_HOST_SRV_SERROR_RATE	0.508
2	SRV_RERROR_RATE	0.212
3	DST_HOST_SAME_SRC_PORT_RATE	0.068
4	DST_HOST_RERROR_RATE	0.052
5	IS_GUEST LOGIN	0.043
6	IS_HOST_LOGIN	0.021
7	DST_HOST_SRV_DIFF_HOST_RATE	0.018
8	DST_HOST_SRV_COUNT	0.017
9	WRONG_FRAGMENT	0.012
10	DST_HOST_SAME_SRV_RATE	0.010

TABLE IX. THE MOST RELEVANT FEATURES IN THE UNSW-NB15 DATASET

PCs =3		
S/N	**Features**	**Weights**
1	*dwin*	0.588
2	*sttl*	0.146
3	*ct_srv_dst*	0.078
PCs =6		
S/N	**Features**	**Weights**
1	*dwin*	0.588
2	*sttl*	0.146
3	*ct_srv_dst*	0.078
4	*dttl*	0.034
5	*stcpb*	0.024
6	*dtcpb*	0.023
PCs =10		
S/N	**Features**	**Weights**
1	*dwin*	0.588
2	*sttl*	0.146
3	*ct_srv_dst*	0.078
4	*dttl*	0.034
5	*stcpb*	0.024
6	*dtcpb*	0.023
7	*dmeanz*	0.019
8	*ct_srv_src*	0.014
9	*smeanz*	0.012
10	*swin*	0.011

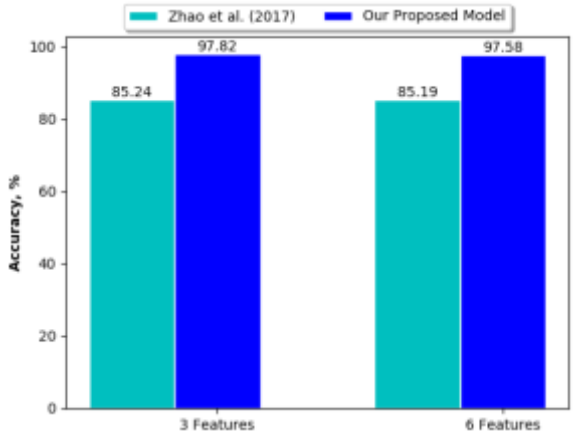


Fig. 6. Performance Comparison of the Proposed Model.

Cybersecurity has become an important research area in the Internet of Things, especially with the vast amount of sensitive data stored and transmitted by IoT devices. IoT devices have several security threats such as eavesdropping, data leakage/loss, denial-of-service attacks, etc. In tackling these issues, this paper presented a hybridized machine model that detects several anomalies. The proposed hybrid detection model detects anomalies in two most common communication models in IoT devices (i.e., direct and gateway-based communication models). One of the proposed model features is learning and detecting malicious patterns in IoT traffic data. Such functionality involves learning the benign and detecting the anomalies that do not conform to the normal patterns.

The model presented in this paper detects threats in the network layer of the IoT. The need for such a detection model in this layer of the IoT cannot be overemphasized since the network layer is most vulnerable to attacks due to the large amount of data it transmits. The proposed model accurately detects the denial of Service (DoS) attack in the IoT network layer with a low false alarm rate. Another threat in the IoT network layer detected by the model proposed in this paper is the routing attack (Probing attack). Such attacks are used to scan the network for possible vulnerabilities. Attacks used to escalate privileges (such as U2R and R2L attacks) come under this category. The model can detect normal and malicious behaviours and identify four different attack types in the IoT (using its multi-classification feature).

The uniqueness of the proposed hybrid intrusion detection model is in its ability to be trained with unlabelled data. The model ensures a quality experience for users and security experts as manual data identification and labelling are not needed. This attribute is required in detection models in IoT networks since the acquisition of labels in big data from IoT devices can be time-consuming and laborious. Furthermore, the high accuracy score of the model guarantees that malicious data (threats) in IoT traffic can be detected, thereby reducing zero-day exploits in IoT networks. The dimension reduction performed on the features ensures the low complexity of the model desired when dealing with IoT devices with limited resources such as memory and processing power. The model, when accurately deployed, can alert security experts to initiate preventive measures from the identified threats. Providing prior warnings aids administrators, stakeholders in IoT and minimizes exploitable vulnerabilities. Consequently, the security of sensitive data is enhanced, which preserves the privacy of IoT users.

V. CONCLUSION

This paper proposed a hybrid model for the detection of anomalies in the network layer of the IoT. The proposed system performs dimension reduction (using PCA algorithm), data clustering (using K-means algorithm), and a data classification based on the Support Vector Machine (SVM) algorithm. The proposed hybrid model was evaluated on both the NSL-KDD and the UNSW-NB15 datasets. Performance evaluation of the proposed model shows that dimension reduction improves the detection rate of attacks since irrelevant features that increase noise are removed from the new dataset (with reduced features). The conducted experiments also revealed that classification accuracy is higher with binary classification than with multi-class, mainly when classes are generated from cluster labels (i.e., unsupervised learning). Also, the classifier was benchmarked with the classifier presented by Zhao, Li [22]. Our proposed model outperforms the model shown by Zhao, Li [22] in terms of detection rate and accuracy. As future work, we will employ the proposed

hybrid anomaly detection model to detect different categories of IoT attacks that are not covered in this paper (i.e., from other datasets that simulate various attack activities).

REFERENCES

[1] Zhang, Z.-K., M.C.Y. Cho, and S. Shieh. Emerging security threats and countermeasures in IoT. in Proceedings of the 10th ACM symposium on information, computer and communications security. 2015. ACM.

[2] Singh, S. and N. Singh. Internet of Things (IoT): Security challenges, business opportunities & reference architecture for E-commerce. in 2015 International Conference on Green Computing and Internet of Things (ICGCIoT). 2015. IEEE.

[3] Gartner, Gartner Says 6.4 Billion Connected. (2015). Retrieved September 14, 2017 from http://www.gartner.com/newsroom/id/ 3165317. 215.

[4] Baig, Z.A., et al., Future challenges for smart cities: Cyber-security and digital forensics. Digital Investigation, 2017. 22: p. 3-13.

[5] Sheikhan, M. and H. Bostani. A hybrid intrusion detection architecture for internet of things. in 2016 8th International Symposium on Telecommunications (IST). 2016. IEEE.

[6] Desai, A.S. and D. Gaikwad. Real time hybrid intrusion detection system using signature matching algorithm and fuzzy-GA. in 2016 IEEE international conference on advances in electronics, communication and computer technology (ICAECCT). 2016. IEEE.

[7] Sedjelmaci, H., S.M. Senouci, and M. Al-Bahri. A lightweight anomaly detection technique for low-resource IoT devices: A game-theoretic methodology. in 2016 IEEE International Conference on Communications (ICC). 2016. IEEE.

[8] Erfani, S.M., et al., High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recognition, 2016. 58: p. 121-134.

[9] Cherian, M. and M. Chatterjee. Survey of Security Threats in IoT and Emerging Countermeasures. in International Symposium on Security in Computing and Communication. 2018. Springer.

[10] Adat, V. and B. Gupta, Security in Internet of Things: issues, challenges, taxonomy, and architecture. Telecommunication Systems, 2018. 67(3): p. 423-441.

[11] Lohachab, A. and B. Karambir, Critical Analysis of DDoS—An Emerging Security Threat over IoT Networks. Journal of Communications and Information Networks, 2018. 3(3): p. 57-78.

[12] Khan, R., et al. Future internet: the internet of things architecture, possible applications and key challenges. in 2012 10th international conference on frontiers of information technology. 2012. IEEE.

[13] Tweneboah-Koduah, S., K.E. Skouby, and R. Tadayoni, Cyber security threats to IoT applications and service domains. Wireless Personal Communications, 2017. 95(1): p. 169-185.

[14] Choraś, M., R. Kozik, and I. Maciejewska, Emerging cyber security: Bio-inspired techniques and MITM detection in IoT, in Combatting Cybercrime and Cyberterrorism. 2016, Springer. p. 193-207.

[15] Sapienza, A., et al. Discover: Mining online chatter for emerging cyber threats. in Companion Proceedings of the The Web Conference 2018. 2018. International World Wide Web Conferences Steering Committee.

[16] Harel, Y., I.B. Gal, and Y. Elovici, Cyber security and the role of intelligent systems in addressing its challenges. ACM Transactions on Intelligent Systems and Technology (TIST), 2017. 8(4): p. 49.

[17] Berral-García, J.L. A quick view on current techniques and machine learning algorithms for big data analytics. in 2016 18th international conference on transparent optical networks (ICTON). 2016. IEEE.

[18] Shanthamallu, U.S., et al. A brief survey of machine learning methods and their sensor and IoT applications. in 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA). 2017. IEEE.

[19] Li, Y., R. Ma, and R. Jiao, A hybrid malicious code detection method based on deep learning. International Journal of Security and Its Applications, 2015. 9(5): p. 205-216.

[20] Nskh, P., M.N. Varma, and R.R. Naik. Principle component analysis based intrusion detection system using support vector machine. in 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). 2016. IEEE.

[21] Pajouh, H.H., et al., A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in IoT backbone networks. IEEE Transactions on Emerging Topics in Computing, 2016.

[22] Zhao, S., et al. A dimension reduction model and classifier for anomaly-based intrusion detection in internet of things. in 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech). 2017. IEEE.

[23] Narudin, F.A., et al., Evaluation of machine learning classifiers for mobile malware detection. Soft Computing, 2016. 20(1): p. 343-357.

[24] Nobakht, M., V. Sivaraman, and R. Boreli. A host-based intrusion detection and mitigation framework for smart home IoT using OpenFlow. in 2016 11th International conference on availability, reliability and security (ARES). 2016. IEEE.

[25] Doshi, R., N. Apthorpe, and N. Feamster. Machine learning ddos detection for consumer internet of things devices. in 2018 IEEE Security and Privacy Workshops (SPW). 2018. IEEE.

[26] McKinney, W., Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. 2012: " O'Reilly Media, Inc.".

[27] Hackeling, G., Mastering Machine Learning with scikit-learn. 2017: Packt Publishing Ltd.

[28] Tavallaee, M., et al. A detailed analysis of the KDD CUP 99 data set. in 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. 2009. IEEE.

[29] Rehim, R., Python Penetration Testing Cookbook: Practical recipes on implementing information gathering, network security, intrusion detection, and post-exploitation. 2017: Packt Publishing Ltd.

[30] Moustafa, N. and J. Slay. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). in 2015 military communications and information systems conference (MilCIS). 2015. IEEE.

[31] Zheng, Y., et al. Smart car parking: temporal clustering and anomaly detection in urban car parking. in 2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP). 2014. IEEE.

[32] MacQueen, J. Some methods for classification and analysis of multivariate observations. in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967. Oakland, CA, USA.

[33] Hasan, M., et al., Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. Internet of Things, 2019. 7: p. 100059.

Data Dissemination for Bioinformatics Application using Agent Migration

Shakir Ullah Shah[1]
Iqra University, Islamabad, Pakistan
National University of Computer and Emerging Sciences
Peshawar, Pakistan

Jamil Ahmad[3]
Senior IEEE Member Hazara University
Mansehra KP
Pakistan

Abdul Hameed[2]
Department of Computer Science
Iqra University, Islamabad, Pakistan

Hafeez Ur Rehman Safia Fatima[4], Muhammad Amin[5]
National University of Computer and Emerging Sciences
Peshawar, Pakistan

***Abstract*—Bioinformatics is research intensive field where agents operate in highly dynamic environment. Due to extensive research in this domain leads to basic but important problems for the researchers that are (1) Bandwidth (2) storage and (3) computation. We are using agent migration approach to reduce the network load and resolve the resource problem for the client by using server side resources for the computations on large data. The proposed approach does not demand extra storage and extensive computational resources on clients slide fsage. It solves the problem of bandwidth, storage, computation. Our results show that this approach saves the time of the user up to 12.5 % approximately, depending on the size of the data. Similarly the agent can work like a mashup to get heterogeneous data from different service providers and presents in homogeneous shape to its owner.**

***Keywords*—Data dissemination; protein-protein interactions; agent migration; inter-platform mobility; multi-agent systems**

I. Introduction

Bioinformatics [1], [2] is one of the applications of computer science for managing the biological information. It is interdisciplinary field of sciences which combines mathematics, computer science, statistics, and engineering to understand bio- logical data. It is a vast field and with the number of researcher doing research in it makes it more important. Researchers have done a lot of work to find and understand the nature as well as dynamics of proteins [3]. The protein-protein interactions [4], [5] involve structural information [4] of proteins as well as the non-structural information [6], [7]. A variety of computational approaches [8] have also been developed. Yet comparative studies suggest that the data involve in it is large, complex and not in one standard format; with Next Generation Sequences (NGS) [9], [10], [11] on which researcher are doing work to extract new interactions, using this (NGS) leads to another problem that is storage [12], [13].

There are a lot of experimental methods [14] which are noisy, costly in term of computation and storage and time consuming to predict the protein-protein interactions. Due to high computation, high storage need and data heterogeneity it is hard for researchers to carry out their research, so here we recommend an agent [15], [16], [17] based approach which will reduce the bandwidth need, transfer computations to the machine who has high computational power and will give data in homogeneous format to increase the researchers' productivity. Although agent itself has many characteristics but the characteristics we will be using throughout this study is agent mobility. The big picture is that an agent gets requested from its owner and visits service provide as per the list to fetch required data, manipulate the data at service side and return back to its originated platform with the results. In this study, we are not targeting on communication problems and assume that environment is up and running.

The rest of the paper is organized as follows. In section II, literature about the domain is given. Proposed solution along with main steps is given Section III. Detail about reference implementation is given in Section IV for the proof of concept. The importance of the proposed solution is given in the form of results are discussed in Section V. The implication of the research work is in Section VI.

II. Literature Review

Proteins are large molecules, which are the collection of amino acids, are essential to our bodies to function properly [18]. Proteins are very important component in the proper functioning and maintaining of our body structures, its normal functions and the regulation of the body's different parts. Enzymes which are responsible to speed up a chemical reaction are also proteins. Oxygen is an essential element for all living being for their survival and proteins play utmost important role as a carrier in the form of hemoglobin. Proteins help us fight infection as well as DNA the building blocks to life. It's too required to create up muscle tissue, which in turn makes a difference to keep our bodies dynamic, solid, and healthy. Most protein is put away within the body as muscle, by and large bookkeeping for around 40-45% of our bodies add up to pool [6].

Researchers have done a huge amount of work to find and understand the nature as well as changing aspects of proteins. The protein-protein interactions [4], [5], [19] involve either structural information of proteins like Domains, 3-D shape of

proteins, structural neighbors as well as the non-structural information that includes protein homology, sequence similarity, functional similarity etc. Extended form of computational approaches, for illustration, on arrangement homology, quality co-expression and phylogenetic profiles, have moreover been created for the genome-wide deduction of proteinprotein interactions (PPIs).

Prediction of PPIs at the structural level is essential as it allows predication of protein functions, helps in the discovery of drug and so play vital role in so many other areas [6], [20]. Protein Interactions by Structural Matching (PRISM) [21], [22] protocol is provided huge scale forecast of protein-protein intuitive and gathering of protein complex structure. PRISM method consists of two parts:

- Firm body basic structural comparison of the intended protein to know the template PPIs.
- Adaptable refinement through the use of docking energy function.

PRISM predicts binding residue by using structural like-ness and developmental conversation of putative binding residue but require high computational power and high band-width to stay active. Huge number of tools and models have been developed in recent years for the interpretation of biological data, but not all of these are publicly available or permit bulk submission via web [23]. While few tools and models require proper training and background knowledge but the proposed solution is very simple.

There is a huge growth in the biological sequence where a tremendous sum of information is being created and uploaded on the web sites/servers. Now to get the data we would need to interact with the interface using web based queries [24]. This means that the researcher has to do a query each time he/she needs the data source. Above all these resources would be in different formats, entries, query options etc. Moreover, this process requires the researchers to remain online and wait for the required results. Secondly this approach needs high bandwidth. This is also very important retrieve the results on low computation powered resources like mobile phones. This study propose migration feature of the agent to overcome aforementioned problems by transforming the computation at the required host of resources [17].

Knowledge administration is a repetitive, complex and time-consuming task. It requires high computational resources. In specific, the kinds of assets accessible within the bioinformatics space are various databases and investigation instruments. These resources can be autonomously managed in topographically unmistakable areas, using Multi-agent approach [25]. Researchers, in the field of bioinformatics, consistently propose various techniques to resolve such issues.

There are various approaches to retrieve data from server like remote procedure call (RPC), java Remote Method Invocation (RMI), etc. [26], [27]. This study focus only on using Multi-Agent Systems (MAS) [16]. MAS, like Jason [28] is a system comprised of multiple agents, a peace of computational logical unit to perform different tasks on behalf of its creator.

III. Proposed Solution

We propose Agent migration approach to solve the problems of low Bandwidth, low storage extensive resources and dynamic environment. The purpose of using agent migration approach is to minimize the network load, increase flexibility and enhance parallel processing. Mobile agents are actually autonomous (act independently) programs that do travel form one system to another in a network. Mobile agents are proactive, reactive, flexible and social. They are trained to the task assigned the users. Due to their functionalities as mentioned above , they proved highly effective many scenarios where everyone is busy and have less time and more task to do[10, 11]. They have the capability that they suspend their execution in one system and migrate to other system to resume their computation. To consume less time suspension strategy has proven its worth, by suspending when one system migrating to another system to resume the work is highly effective when previous system doesn't have required sources to complete the task, agent move to another system with the task, complete the task over their on second system and get to fist system with results only. This way all the users don't have to buy a high specifications system, they can use other high power system to complete their work.

A. Agent Mobility

In agents migration, agent mobility has further two types: Inter-platform mobility and Intra-platform mobility. In case of intra-platform mobility agent moves between different containers within the same platform on the other hand in inter-platform mobility agent leave one system lets say client and move or migrate to other system so called server, means agent is moving between different platforms. The main focus of this is intra-platform mobility.

B. Main Steps of Proposed Solution

This study proposes a step by step solution based on agent mobility. The pre and post details of each step provide great insight of the requirements. These steps are visualized in Fig. 1 and its sequence diagram is given in Fig. 2.

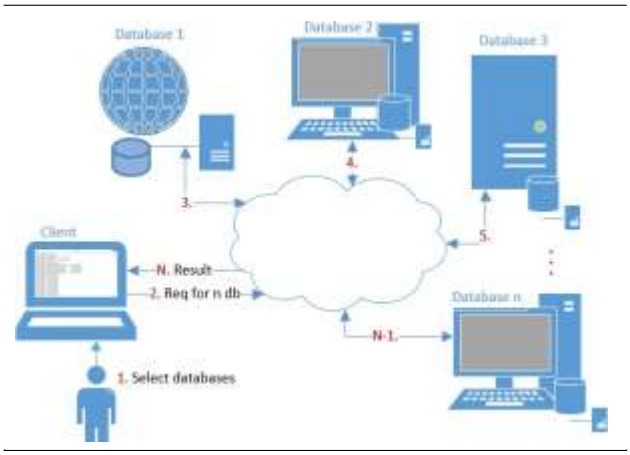


Fig. 1. Main Steps of Proposed Solution.

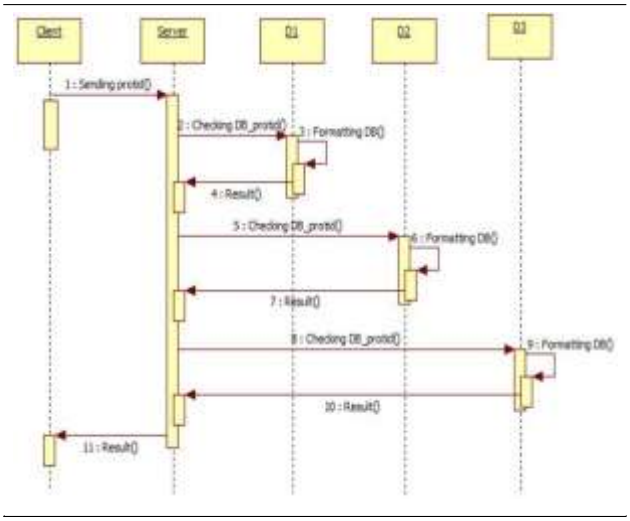


Fig. 2. Sequence Diagram.

The detail of each step is given below:

1) Enter protein ID: Here the user will enter the protein id of which he/she wants to check the protein to protein Interactions.

2) Select dbs: On the bases of given id, all possible protein to protein interactions will be shown to user. User will select databases according to its need.

3) Fetching the data: Agent will leave the client and will fetch the data from each service provider.

4) Merge all in one file: As the user can select multiple databases where the interactions would be in different formates. Agent will convert heterogeneous formates into homoge- neous formate, like xml.

IV. Reference Implementation

There are various agent framework to deploy agents [29],[30] In this study JADE [31] is used as an agent framework as it is fully complaint to Foundation for Intelligent Physical Agents (FIPA) [32], open source and is used in the state of art agent technology. JADE does not support intra-platform mobility, so for that JIPMS [33] is used for this purpose. It is also based on JAVA, so it becomes fare to compare it with JAVA RMI. For the sack of comparison, protein id and selection of other parameter like number of interactions remained same for both. The main GUI can be seen in Fig. 3. When the user enters protein id and number, then the agent fetches the record of possible protein to protein interaction from all available databases and gives a table view. The user will select all those dbs that are required for her. When the user complete its selection, then the agent will fetch all the data from selected databases, Agent will combine all required data in a predefined format. All the files or interactions would be downloaded in XML format with different tags i.e. Source db, author etc. Then agent comes back to its originator platform. When agent notifies the user about the work done, then the user can view the fetched data on client side.

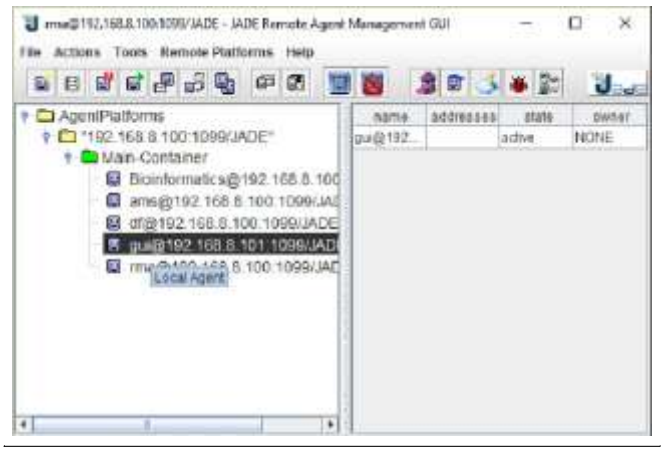


Fig. 3. Main GUI.

V. Result and Discussion

This section describes the results achieved so far and concludes with some discussion. There are different studies carried to compare performance between JAVA RMI and agent based system against response time and network load [?]. The study focused on implementing proposed method using JAVA RMI and mobile agent framework. The Fig. 4 shows that agent based approach causes less load over the network as compared to JAVA RMI.

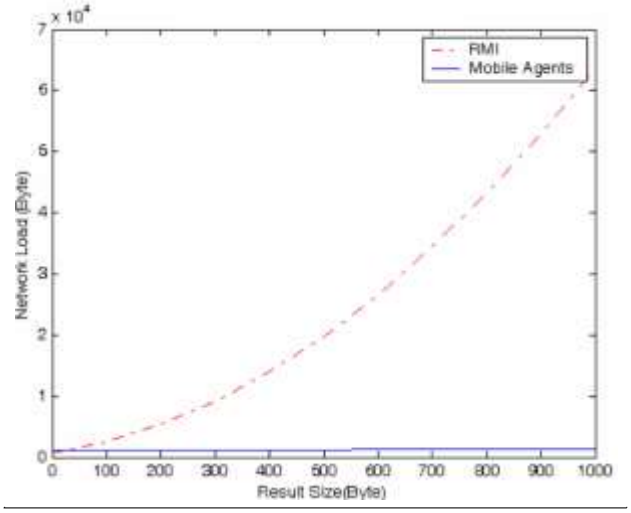


Fig. 4. Client Side Network Load and Result Size.

Fig. 5 shows relation between response time and result size. It shows that agent based approach produce better results as compared to JAVA RMI when the result size becomes greater than 2KB.

In the Fig. 6 given above the blue line shows agents graph and red line shows the graph of java RMI we can clearly see from the graph that if we decrease the bandwidth our agent is computing faster as compare to JAVA RMI. As agent can move at low bandwidth and don't require high bandwidth for migration and so on that's why agent is showing good results even at low bandwidth unlike JAVA RMI. Moreover JAVA RMI won't even work when bandwidth is 5Mbs unlike agent;

agent can migrate at 5 Mbs. If let's say the bandwidth is 15Mbs then our agent takes almost 41.66% (time/60*100) to serve the clients request while for the same bandwidth java is taking more time than agent. Using agents has solved the bandwidth problem, Storage problem and the formatting too.

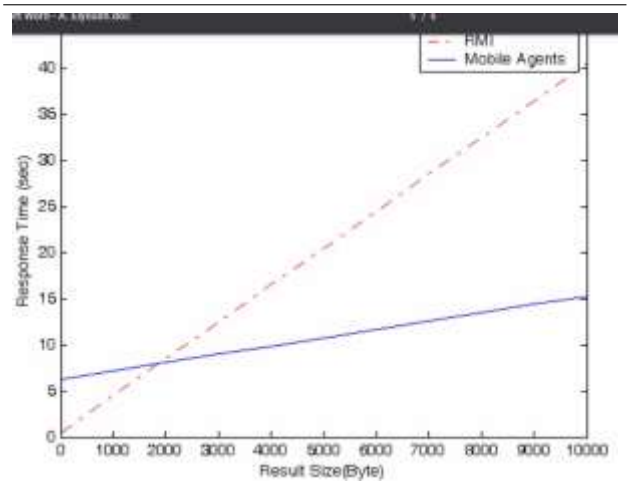


Fig. 5. Response Time vs Result Size.

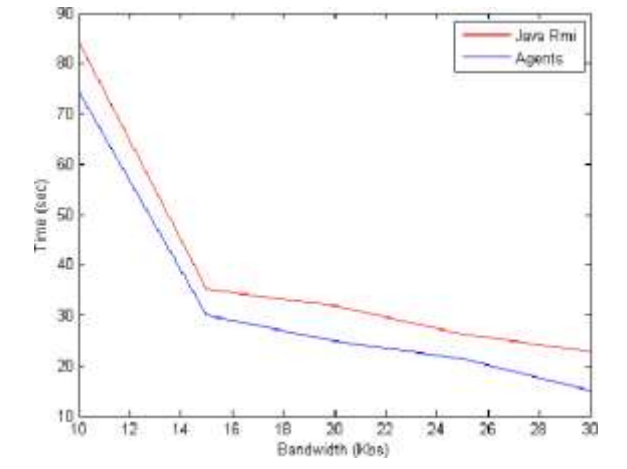


Fig. 6. Comparison between Agents and Java RMI.

VI. Conclusion and Future Work

Mobile Agent technology can be used in many areas, merging this technique with other areas can also be useful because it does not require high bandwidth or strong Internet connection. Agents are intelligent and can even work well even in low network areas. If we see the future of agents' migration approach in bioinformatics it can be highly beneficial for huge amount of data and more computations. It can be used for many generic purposes as well. This study found the interactions between proteins using agent migration protocol. This approach found that client with limited resources can also be used for finding protein to protein interaction. The finding of this study is that mobile agent technology leverage network load and storage on client side and heterogeneous data can be converted into homogeneous format. Furthermore this approach does not demand the availability of the user online for full time. Our research can be modified to make it work on different bioinformatics problem like viewing the interaction of sequences.

Acknowledgment

We are grateful to Waqas Haider Khan Bangyal for reading the manuscript and for improving proofs of this manuscript several times.

References

[1] A. D. Baxevanis, G. D. Bader, and D. S. Wishart, Bioinformatics. John Wiley & Sons, 2020.

[2] R. Stevens, C. A. Goble, and S. Bechhofer, "Ontology-based knowledge representation for bioinformatics," Briefings in bioinformatics, vol. 1, no. 4, pp. 398–414, 2000.

[3] A. Amadei, A. B. Linssen, and H. J. Berendsen, "Essential dynamics of proteins," Proteins: Structure, Function, and Bioinformatics, vol. 17, no. 4, pp. 412–425, 1993.

[4] D. F. Waugh, "Protein-protein interactions," Advances in protein chemistry, vol. 9, pp. 325–437, 1954.

[5] T. Bergg a˚rd, S. Linse, and P. James, "Methods for the detection and analysis of protein–protein interactions," Proteomics, vol. 7, no. 16, pp. 2833–2842, 2007.

[6] Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter et al., "Structure-based prediction of protein–protein interactions on a genome-wide scale," Nature, vol. 490, no. 7421, p. 556, 2012.

[7] S. Das and S. Chakrabarti, "Classification and prediction of protein–protein interaction interface using machine learning algorithm," Scientific Reports, vol. 11, no. 1, pp. 1–12, 2021.

[8] K. A Theofilatos, C. M Dimitrakopoulos, A. K Tsakalidis, S. D Likothanassis, S. T Papadimitriou, and S. P Mavroudi, "Com- putational approaches for the prediction of protein-protein interactions: a survey," Current Bioinformatics, vol. 6, no. 4, pp. 398–414, 2011.

[9] R. B. Russell, F. Alber, P. Aloy, F. P. Davis, D. Korkin, M. Pichaud, M. Topf, and A. Sali, "A structural perspective on protein–protein interactions," Current opinion in structural biology, vol. 14, no. 3, pp. 313–324, 2004.

[10] B. Suter, X. Zhang, C. G. Pesce, A. R. Mendelsohn, S. P. Dinesh-Kumar, and J.-H. Mao, "Next-generation sequencing for binary protein–protein interactions," Frontiers in genetics, vol. 6, p. 346, 2015.

[11] R. Carter, A. Luchini, L. Liotta, and A. Haymond, "Next-generation techniques for determination of protein-protein interactions: Beyond the crystal structure," Current pathobiology reports, vol. 7, no. 3, pp. 61–71, 2019.

[12] D. S. Horner, G. Pavesi, T. Castrignano, P. D. De Meo, S. Liuni, M. Sammeth, E. Picardi, and G. Pesole, "Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing," Briefings in bioinformatics, vol. 11, no. 2, pp. 181–197, 2010.

[13] J. K. Kulski, "Next-generation sequencingan overview of the history, tools, and omic applications," Next generation sequencing-advances, applications and challenges, pp. 3–60, 2016.

[14] N. Safari-Alighiarloo, M. Taghizadeh, M. Rezaei-Tavirani, B. Goliaei, and A. A. Peyvandi, "Protein-protein interaction networks (ppi) and complex diseases," Gastroenterology and Hepatology from bed to bench, vol. 7, no. 1, p. 17, 2014.

[15] M. J. Wooldridge and N. R. Jennings, "Intelligent agents: Theory and practice," The knowledge engineering review, vol. 10, no. 2, pp. 115–152, 1995.

[16] A. Omicini, A. Ricci, and M. Viroli, "Artifacts in the a&a meta-model for multi-agent systems," Autonomous agents and multi-agent systems, vol. 17, no. 3, pp. 432–456, 2008.

[17] L. Gao, H. Dai, T.-L. Zhang, and K.-C. Chou, "Remote data retrieval for bioinformatics applications: An agent migration approach," PloS one, vol. 6, no. 6, p. e20949, 2011.

[18] T. Simonson, "Electrostatics and dynamics of proteins," Reports on Progress in Physics, vol. 66, no. 5, p. 737, 2003.

[19] M. Shatsky, R. Nussinov, and H. J. Wolfson, "A method for simultaneous alignment of multiple protein structures," Proteins: Structure, Function, and Bioinformatics, vol. 56, no. 1, pp. 143–156, 2004.

[20] J. Zahiri, J. Hannon Bozorgmehr, and A. Masoudi-Nejad, "Computational prediction of protein–protein interaction networks: algorithms and resources," Current genomics, vol. 14, no. 6, pp. 397–414, 2013.

[21] U. Ogmen, O. Keskin, A. S. Aytuna, R. Nussinov, and A. Gursoy, "Prism: protein interactions by structural matching," Nucleic acids research, vol. 33, no. suppl 2, pp. W331–W336, 2005.

[22] N. Tuncbag, A. Gursoy, R. Nussinov, and O. Keskin, "Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using prism," Nature protocols, vol. 6, no. 9, p. 1341, 2011.

[23] Y. Ding and L. Gao, "Macrodynamics analysis of migration behaviors in large-scale mobile agent systems for the future internet," IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 41, no. 5, pp. 1032–1036, 2011.

[24] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," Journal of theoretical biology, vol. 273, no. 1, pp. 236–247, 2011.

[25] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," Proceedings of the IEEE, vol. 95, no. 1, pp. 215–233, 2007.

[26] G. A. Aderounmu, B. Oyatokun, and M. Adigun, "Remote method invocation and mobil agent: A comparative analysis." Issues in Informing Science & Information Technology, vol. 3, 2006.

[27] K. Miller and G. Mansingh, "Comparing the use of mobile intelligent agents vs client server approach in a distributed mobile health application." JCP, vol. 10, no. 6, pp. 365–373, 2015.

[28] R. H. Bordini and J. F. H u¨bner, "Bdi agent programming in agentspeak using jason," in International workshop on computational logic in multi-agent systems. Springer, 2005, pp. 143–164.

[29] R. H. Bordini, L. Braubach, M. Dastani, A. E. F. Seghrouchni, J. J. Gomez-Sanz, J. Leite, G. O'Hare, A. Pokahr, and A. Ricci, "A survey of programming languages and platforms for multi-agent systems," Informatica, vol. 30, no. 1, 2006.

[30] K. Kravari and N. Bassiliades, "A survey of agent platforms," Journal of Artificial Societies and Social Simulation, vol. 18, no. 1, p. 11, 2015.

[31] F. Bellifemine, "Jade-a white paper," exp, vol. 3, no. 3, 2003.

[32] T. SpA. (2001) The foundation for intelligent physical agents. Https://www.fipa.org/. [Online]. Available: https://www.fipa.org/.

[33] C. J. (2006) Inter-platform mobility project. Senda research group of Autonomous University of Barcelona. [Online]. Available: Availableat: {https://tao.uab.cat/ipmp/}.

Hybrid Metaheuristic Aided Energy Efficient Cluster Head Selection in Wireless Sensor Network

Turki Ali Alghamdi
Department of Computer Science, College of Computer and
Information Systems, Umm Al-Qura University, Makkah, Saudi Arabia

***Abstract*—Clustering is one of the significant techniques for expanding the lifetime of networks in wireless sensor networks (WSNs). It entails combining of sensor nodes (SNs) into clusters and electing cluster heads (CHs) for each and every cluster. CH collects the information from particular cluster nodes and passes the cumulative data to the base station (BS). However, the most important requirement in WSN is to choose a suitable CH with an increased network life span. This work introduces a new CHS model in WSN. The optimal CH is elected by a new hybridized model termed as "Lion Updated Dragonfly Algorithm (LU-DA) that hybrid the concepts of Dragonfly Algorithm (DA) and Lion Algorithm (LA)". Moreover, the optimal selection of CH is done depending upon constraints like "energy, delay, distance, security (risk) and trust (direct and indirect trust)". This optimal CH ensures the network lifetime enhancement. At last, the superiority of the developed approach is proved on varied measures like energy and alive node analysis. Accordingly, the proposed model has accomplished higher energy of 0.55 at 1st round, whereas at the 2000th round, the normalized energy value has been dropped to 0.1.**

Keywords*—Cluster head; security; trust; dragonfly algorithm; LU-DA model

Nomenclature

Abbreviation	Description
ACO	Ant Colony Optimization
APTEEN	Adaptive Threshold sensitive Energy Efficient sensor Network
BS	Base Station
BOA	Butterfly Optimization Algorithm
CH	Cluster Head
CHS	Cluster Head Selection
DA	Dragonfly Algorithm
EE	Energy Efficiency
FF	Firefly
FCM	Fuzzy C-Means
FPU-DA	fire fly replaced position update in dragonfly
GECR	Genetic Algorithm-based Energy-Efficient Clustering and Routing
GWO	grey wolf optimizer
GA	Genetic Algorithm
HSO	Harmony Search Optimization
HML	Hierarchical Maximum Likelihood
KFCM	kernel fuzzy C-means
LEACH	Low-Energy Adaptive Clustering Hierarchy
LU-DA	Lion Updated Dragonfly Algorithm
LA	Lion Algorithm
MOFPL	Multi-objective fractional particle lion algorithm
NAN	Number of Alive Node
OWSN	Optical WSN
PSO	Particle Swarm optimization
PEGASIS	Power-Efficient Gathering in Sensor Information Systems
QoS	Quality of Services
QCM2R	QoS-aware cross-layered multichannel multisink routing
SNs	Sensor Nodes
TEEN	Threshold sensitive Energy Efficient sensor Network
WSN	Wireless Sensor Network

I. Introduction

WSN [1] [2] involves various sensors connected to the wireless medium. The sensed data from SNs is typically forward to BS, where the data is collected, analyzed and performed certain actions accordingly [3]. The WSN is deployed in various applications like weather monitoring [4], meteorological data collection [5], and field surveillance, transportation, and health-care] [6] [7]. However, the nodes in WSN don't have any storage devices and facilities of researchable batteries [8] [9]. Though, it should support any system with effective power consumption [10] [11].

Clustering is a renowned procedure for effective data transmission with respect to energy and power utilization. Clustering involves dividing of SNs into different clusters [12] [13] [14]. All clusters in networks have distinctive CHs [15] [16], which is responsible to transfer information to other SNs in its cluster. Moreover, the communication to BS is carried out only through this CH. In this scenario, key role is to opt the optimal CH by concerning on lesser delay and low consumption of energy [17] [18]. Thereby, creating a cluster with aggregation and data fusion models, there is energy in network based on the data transmitted to BS [19] [20].

Thereby, the cluster-oriented models also engaged in facilitating the extension of network lifetime [21]. The frequently deployed algorithms include APTEEN, TEEN, LEACH, PEGASIS, and FCM. Further, "LEACH is the cluster-based algorithm that operates in the distributed

manner, which elects the CH depending on the predetermined probability" [22].

Various cluster-oriented models have been introduced so far, which is based on meta-heuristic algorithms. However, the algorithms possess some common challenges such as high convergence, local search issues in FF, and high cost. Moreover, there is a prerequisite of standard optimizations and need consideration on constraints, namely security and trust [23]. So in order to solve the above mentioned issues, this paper introduces a new CHS model in WSN.

The foremost contribution is listed here:

- The selection of optimal CH depends on certain constraints such as energy consumption, trust, security, delay and distance.
- Proposes a Lion Updated Dragonfly Algorithm for optimal CH selection, which integrates the concepts of DA as well as LA models.

The organization of the paper is as follows: Section II reviews CHS models. Section III elaborates the adopted energy aware clustering model in WSN. Section IV depicts the optimal CHS: objective model. Section V explains Lion Updated Dragonfly Algorithm for optimal CHS. The resultants and conclusions are briefed in Section VI and Section VII.

II. Literature Survey

A. Related Work

In 2020, Turki et al. [24] suggested a new clustering model with optimal CHS that considered 4 most important criterions such as security, delay, energy, and distance. Furthermore, for electing the optimum CHs, this work proposed a novel algorithm named as FF-PUD. At last, the performances of developed scheme were performed by evaluating it over other schemes regarding risk, alive nodes, energy and delay.

In 2020, Prachi et al. [25] have employed BOA for choosing an optimal CH from nodes. Accordingly, the developed work aims on lessening the energy usage and for maximizing the life span of network. The path among the CH and BS was determined using ACO and it selected optimal routes depending on node degree, residual energy and distance. At the end, the supremacy of adopted work was proved regarding energy consumption, alive and dead nodes.

In 2019, Reeta and Dinesh [26] have designed a multi-objective model that dependents on distance, traffic rates, energy, cluster densities and delay. Here, energy based routing was carried out based on MOFPL scheme. The implemented model determined the optimum CH from several nodes in WSN. Consequently, the optimal routes were introduced depending on the adopted multi-objective function. Furthermore, effective CHS with high network energy was accomplished by the designed model.

In 2020, Augustine and Ananth [27] have presented an enhanced framework for CHS based on Taylor KFCM that was modified from the KFCM approach in the Taylor series. The introduced model has chosen the CH by means of "acceptability factor" that was evaluated by the trust, distance, and energy. Further, the advantage of the proposed system was proved in terms of highest energy and high trust.

In 2019, Goswami *et al.* [28] introduced a cluster-based model by deploying HML and FF model in OWSN for improving the EE and minimizing the costs. Here, the issues in FF model were prevailed over by integrating the theory of HML with it. Furthermore, the distribution of power in nodes was carried out precisely via maximum likelihood property of HML. Finally, the resultants have shown the betterment of presented scheme regarding EE and cost function.

In 2019, Jain and Toor [29] offered a new framework for diverse WSN by considering "MEACBM routing protocol". Accordingly, optimal election of CHs takes place; particularly, the SNs with higher energy were preferred as CH. This model has minimized the energy utilization of SNs while conveying data to BS. The analysis resultants have revealed the enhancement regarding the CH count, network lifetime, throughput and dead node count.

In 2019, Daneshvar et al. [30] have offered a new clustering scheme, which selected CHs by means of GWO. For selecting CHs, the solutions were optimized depending on remaining energy of every node and predicted energy utilization. In addition, for improving the EE, the presented model deployed the similar clustering in numerous successive rounds. This allowed the framework to accumulate the energy, which was necessary for reforming the clustering. Eventually, the outcomes demonstrated that the designed model has ensured effective network lifetime.

In 2018, Tianshu *et al.* [31] suggested a routing scheme depending upon GECR and GA for expanding the lifetime of networks and improving EE. In addition, while modelling the objective function, the "load balancing factor" was taken into account that balanced the energy usage among SNs. The simulated results have exposed the supremacy of the adopted method with lower variance and improved EE.

B. Problem Formulation

Table I makes a review of existing cluster-based energy-aware CHS models in WSN. Numerous methods have been focused on energy-aware CHS models in WSN. But still, the existing models like FF-PUD[24], BOA + ACO [25], MOFPL [26], Taylor KFCM model [27], FF [28] have some common problems like high convergence, local search issues in FF, high-cost efficiency, there is a need of standard optimizations and need consideration on constraints like security and trust.

C. Objectives

The main objectives of this paper are:

- To select an optimal CH depends on certain constraints such as energy consumption, trust, security, delay and distance.
- To propose an improved Algorithm for optimal CH selection for solving the optimization issues.
- And to improve the better convergence rate.

TABLE I. REVIEWS ON TRADITIONAL ENERGY AWARE CHS MODELS IN WSN

Authors	Techniques	Feature	Challenge
Turki *et al.* [24]	FF-PUD	❖ Minimal delay ❖ High network energy	❖ Coverage issues are not deliberated.
Prachi *et al.* [25]	BOA + ACO	❖ Higher count of alive nodes ❖ Minimal energy consumption	❖ Should consider fault tolerance.
Reeta and Dinesh [26]	MOFPL	❖ Less simulation time ❖ Offers high network energy	❖ Resource management is not taken into account in this work. ❖ Cost efficiency is not considered.
Augustine and Ananth [27]	Taylor KFCM model	❖ High throughput and energy. ❖ Minimal delay	❖ No consideration on real time experiments. ❖ Standard optimizations are required for enhancing the CHS performance.
Goswami *et al.* [28]	FF	❖ Minimal cost function. ❖ Improved EE	❖ FF suffers from local search issues.
Toor and Jain [29]	MEACBM	❖ Minimized the consumption of energy ❖ Raises throughput and lifetime	❖ Needs consideration on scalability of SNs
Daneshvar *et al.* [30]	GWO	❖ Balanced energy consumption ❖ Offers high life span for network	❖ Fault tolerance is not considered.
Tianshu *et al.* [31]	GECR	❖ Better life span ❖ Optimal energy utilization	❖ More appropriate metaheuristic algorithms should be used.

III. PROPOSED ENERGY AWARE CLUSTERING MODEL IN WSN

A. Network Model

Assume M_n sensor nodes that are randomly deployed in appliance area. Consequently, the clustering process is done by merging the SNs. During clustering, the nodes forms clusters, wherein a CH is elected and the total count of CH is delineated by CH_n. Thus, the distances amongst nodes and CHs have to be reduced.

The most important task of WSN is to transfer the information among nodes. Here, the identification of shorter paths is required to enhance the data transmission. Moreover, the energy consumption of node also acts as the most role while transmitting the data. Particularly, a node requires more energy for transmitting massive data. In the clustering based strategy, the CH is responsible for transmitting more data with less energy consumption. However, the security is more crucial for minimizing the overhead and attacks. The architectural depiction of adopted model with varied SNs is illustrated in Fig. 1.

B. Distance Model

In the network, a CH is chosen only if the distance between CH and nodes is minimal. If distance among CH and nodes are higher than distances amid node and BS, the data are transmitted directly to BS by node. By deploying distance matrix $Di(g*w)$, the SNs gets clustered with selected CH as exposed in Eq. (1), wherein, $e_{M_{CH}}$ signifies Euclidean distance amid M_{CH} and normal node position, and $z_1, z_2, \ldots z_n$ signifies SNs. Assume 2 SNs q and d, and positions be x and y. The Euclidean distances amongst 2 nodes are revealed in Eq. (2). In Eq. (1), element e_{M_{CH2},z_1} occupies initial column matrix with minimal distance [24].

$$Di(g*w)=\begin{bmatrix} e_{M_{CH1},z_1}\, e_{M_{CH1},z_2} \ldots\ldots e_{M_{CH1},z_n} \\ e_{M_{CH2},z_1}\, e_{M_{CH2},z_2} \ldots\ldots e_{M_{CH2},z_n} \\ : \\ : \\ e_{M_{CHm},z_1}\, e_{M_{CHm},z_2} \ldots\ldots e_{M_{CHm},z_n} \end{bmatrix} \quad (1)$$

$$e_{q,d}=\sqrt{(q_x-d_x)^2+(q_y-d_y)^2} \quad (2)$$

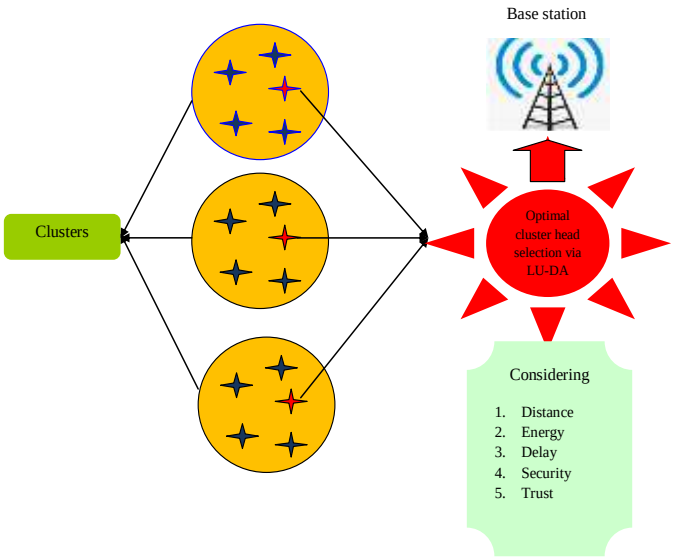


Fig. 1. Architecture of Proposed CHS Model.

Further, the time slots are assigned by M_{CH} to every node during data transmission. Here, M_{CH} collects data from all SNs in clusters. After data gathering, M_{CH} passes the specified data to BS.

C. Energy Model

Energy utilization is a foremost characteristic in WSNs. Actually, additional energy is crucial for conveying data to BS from every SNs. Thereby, the energy model for transmitting data is exposed in Eq. (3), wherein, "E_{ete} symbolizes the electronic energy as given in Eq. (4), wherein E_{agg} refers to the energy utilization during data collection and $E_{TX}(M:e)$ signifies the energy necessary for transferring M bytes of packets at distance e". Eq. (5) shows the essential energy for passing M bytes of packets. Eq. (6) shows the "amplification energy and E_{pr} refers to power amplifier energy and E_{fr} refers to energy required for deploying free space technique" [24].

$$E_{TX}(M:e)=\begin{cases}E_{ete}*M+E_{fr}*M*e^2, if\ e<e_0\\ E_{ete}*M+E_{pr}*M*e^2, if\ e\geq e_0\end{cases} \quad (3)$$

$$E_{ete}=E_{TX}+E_{agg} \quad (4)$$

$$E_{RX}(M:e)=E_{ete}M \quad (5)$$

$$E_{agg}=E_{fr}e^2 \quad (6)$$

$$e_0=\sqrt{\frac{E_{fr}}{E_{pr}}} \quad (7)$$

The whole energy of network is given in Eq. (8), wherein E_1 symbolizes the energy at idle state and E_{ST} symbolizes energy at sensing time.

$$E_{total}=E_{ST}+E_1+E_{RX}+E_{TX} \quad (8)$$

D. Security Model

The risky mode, γ-risky mode and security mode are the factors in security model that are explained below.

"Risky mode: This mode selects an existing CH for facilitating an optimal CHS, for which it takes all the risks. Thus, this mode is considered as an insistent mode while choosing CH [24].

γ-risky mode: The CH which could tolerate the utmost γ-risk are elected based upon γ-risky mode. Accordingly, γ signify the probability measure with values, $\gamma=0$ and $\gamma=1$ (i.e., 100%) similar to security and risky mode.

Security mode: This mode prefers the CH that fulfills the needs of security. In Eq. (1), s_r and s_d denotes the security rank and security needs associated with CHS. If $s_d\leq s_r$, the node are considered as CH".

The probability of security constraints is shown in Eq. (9). Further, "if the chosen CH achieves the state $s_d>s_r$ the risk should be less than 50%. If the condition is $0<s_d-s_r\leq 1$, the selection process would be implemented, and if the state is $1<s_d-s_r\leq 2$, there will be a delay in the selection process. Still, the CHS process could not be completed, and the corresponding function should be continued for the state $2<s_d-s_r\leq 5$".

$$g_{risk}=\begin{cases}0 & if\quad s_d-s_r\leq 0\\ 1-e^{\frac{(s_d-s_r)}{2}} & if\quad 0<s_d-s_r\leq 1\\ 1-e^{\frac{3(s_d-s_r)}{2}} & if\quad 1<s_d-s_r\leq 2\\ 1 & if\quad 2<s_d-s_r\leq 5\end{cases} \quad (9)$$

E. Trust

"Trust is the degree of reliability about other node for performing certain action by keeping track of all past transaction or interactions with nodes by direct or indirect observation. Trust can also be defined as the level of confidence that one node about other node to get assigned work done within some time". Trust includes direct as well as indirect trust. The final trust is computed by combining both indirect and direct trust values [32].

Direct trust: It is computed depending upon interaction of nodes. The distance and energy are regarded as trust measures and is evaluated as per Eq. (10), where $DT_{(A-G)}$ denotes the value of direct trust computed by A and G, Er denotes residual energy of node G, $d(nodeA, nodeG)$ refers to differentiation distance of node A and G.

$$DT_{(A-G)}=\frac{Er}{d(nodeA, nodeG)} \quad (10)$$

Indirect trust: It is computed depending upon recommendation of nodes. It is the summation of trust values computed by other nodes and specified as in Eq. (11), wherein, $IDT_{(A-G)}$ refers to value of direct trust computed by A and G, $DT(A-P)$ and $DT(P-G)$ refers to value of direct trust computed by A, P as well as P and G in that order.

$$IDT_{(A-G)}=\sum DT(A-P)\times DT(P-G) \quad (11)$$

Further, the final trust is computed as in Eq. (12), wherein, $T_{(A-G)}$ refers to final trust A on G and w refers to weight related with indirect and direct trusts.

$$T_{(A-G)}=wDT_{(A-G)}+(1-w)IDT_{(A-G)} \quad (12)$$

IV. Optimal Cluster Head Selection: Objective Model

This work aims to diminish the distance amid the chosen CH and SN and it aims to lessen the delay and risk while transferring the information. On the other hand, the energy, and trust have to be high for better transmission of data. The objective of developed model is delineated in Eq. (13), in which η relies amid $0<\eta<1$, o_m and o_n are calculated as revealed in Eq. (14) and Eq. (15), respectively. The delay, energy, distance, security and trust are explained by ϖ_1, ϖ_2, ϖ_3, ϖ_4, ϖ_5 and are represented as $\varpi_1+\varpi_2+\varpi_3+\varpi_4+\varpi_5=1$. In Eq. (15), $Z_z - A_s$ depicts distance amid normal node and sink.

$$K_n = \eta o_n + (1-\eta) o_m \tag{13}$$

$$o_m = \varpi_1 * o_i^{del} + \varpi_2 * o_i^{ene} + \varpi_3 * o_i^{dis} + \varpi_4 * o_i^{Sec} + \varpi_5 * o_i^{T} \tag{14}$$

$$o_n = \frac{1}{b}\sum_{z=1}^{b} \left\| Z_z - A_s \right\| \tag{15}$$

The fitness function for distance is specified by Eq. (16), wherein, $o_{(m)}^{dis}$ signify packets passed between SN to CH and between CH to BS. o_i^{dis} lies amongst [0, 1].

$$o_i^{dis} = \frac{o_{(m)}^{dis}}{o_{(n)}^{dis}} \tag{16}$$

$$o_{(m)}^{dis} = \sum_{z=1}^{M_z}\left[\left\| CH_z - A_s \right\| + \sum_{x=1}^{M_x} \left\| CH_z - Z_z \right\| \right] \tag{17}$$

$$o_{(n)}^{dis} = \sum_{z=1}^{M_z}\sum_{x=1}^{M_x} \left\| Z_s - Z_x \right\| \tag{18}$$

$o_{(m)}^{dis}$ and $o_{(n)}^{dis}$ are modelled as in Eq. (17) and (18), here Z_z symbolizes SN in z^{th} cluster, CH_z symbolizes CH of z^{th} cluster, the distance amid BS and CH is indicated as $CH_z - A_s$, $CH_z - Z_z$ symbolizes distance among CH and SN and $Z_z - Z_x$ symbolizes distance among 2 SNs, M_z and M_x symbolizes node count devoid of considering x^{th} and z^{th} clusters.

The fitness function for energy $\left(o_i^{ene}\right)$ is revealed in Eq. (19), here, $o_{(m)}^{ene}$ and $o_{(n)}^{ene}$ symbolizes high value of energy and larger CH count.

$$o_i^{ene} = \frac{o_{(m)}^{ene}}{o_{(n)}^{ene}} \tag{19}$$

The fitness function for delay $\left(o_i^{del}\right)$ is revealed in Eq. (20) and it is measured for every SN in cluster. Thus, delay gets lessened if the count of SN in CH is minimal. In Eq. (20), M_n symbolizes total node count, and numerator depicts the higher CH count.

$$o_i^{del} = \frac{\max\left(\left\| CH_z - Z_z \right\|\right)_{z=1}^{M_{CH}}}{M_n} \tag{20}$$

V. Proposed Lion Updated Dragonfly Algorithm for Optimal Cluster Head Selection

A. Solution Encoding

The proposed work focuses on introducing an efficient CHS in WSN. In WSN, the optimum selection of CH is a versatile task and it is performed based upon criteria like, trust, security energy, delay and distance. The input provided to LU-DA algorithm is shown by Fig. 2.

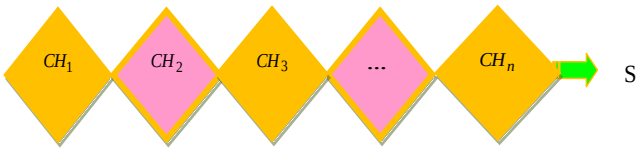


Fig. 2. Solution Encoding.

B. Proposed LU-DA Algorithm

Although the conventional DA [33] model encompasses a variety of enhancements; it suffers from specific limitations like, complexity in solving binary problems etc. Therefore, the theory of LA [34] is mingled with it to introduce a new model named as LU-DA. Hybridized optimization schemes are said to be more appropriate for specific search issues [23] [35] [36] [37]. The steps followed in the proposed LU-DA are as follows.

LU-DA model involve two stages such as: "(i) Exploration and (ii) Exploitation". The separation formula is modelled as exposed in Eq. (21), where, S_l signifies l^{th} closer individual's position, S and U symbolizes current position of individual and nearby individual's count.

$$C_i = -\sum_{l=1}^{U}\left(S - S_l\right) \tag{21}$$

Consequently, the alignment is evaluated as specified in Eq. (22), wherein, Q_l denotes velocity of l^{th} closer individual. The cohesion is modelled as in Eq. (23) and attraction to food is evaluated as in Eq. (24), wherein S^+ symbolizes position of food source and S symbolizes present individual position.

$$B_i = \frac{\sum_{l=1}^{U} Q_l}{U} \tag{22}$$

$$O_i = \frac{\sum_{l=1}^{U_b} S_l}{U_b} - S \tag{23}$$

$$F_i = S^+ - S \tag{24}$$

Distraction to enemies is indicated by Eq. (25), wherein, the enemy position is denoted as S^-.

$$En_i = S^- + S \tag{25}$$

Eq. (26) shows the modelling of step vector (ΔS), wherein, "v signifies separation weight, C_i denotes the separation of i^{th} individual, O denotes the i^{th} individual cohesion, c points out cohesion weight, B_i and F_i refers to the alignment and food resources of i^{th} individual, a signifies the alignment weight, f corresponds to food factor, b symbolizes enemy factor, h signifies the inertia weight, En_i refers to enemy's position of i^{th} individual and it points out iteration counter".

$$\Delta S(it+1) = (vC_i + aB_i + cO_i + fF_i + bEn_i) + h\Delta S(it) \tag{26}$$

As per LU-DA model, if random integer $ra \le \Delta S_{t+1}$ and if $\Delta S_{t+1} \le 0$, the position vector (S) is computed as in Eq. (27), wherein, it signifies current iteration.

$$S(it+1) = S(it) + \Delta S(it+1) \tag{27}$$

On the other hand, if $ra \le \Delta S_{t+1}$ and if $\Delta S_{t+1} > 0$,the position gets updated based on proposed female lion update as shown in Eq. (28), where, ∇_d is evaluated as in Eq. (29), $levy(\beta)$ denotes levy flight. S_d^{fem+} signifies k^{th} vector elements of S^{fem+}, ∇ refers to update function of female, k denotes arbitrary integer and $\breve{r}_2, \breve{r}_1$ signifies arbitrary integer among [0, 1].

$$S_d^{fem+} = \min[S_d^{\max}, \max(S_d^{\min}, \nabla_d)] + levy(\beta) \tag{28}$$

$$\nabla_d = \left| S_d^{fem} + (0.1\breve{r}_2 - 0.05)\left(S_d^{mal} - \breve{r}_1 S_d^{fem}\right) \right| \tag{29}$$

Else if, $ra > \Delta S_{t+1}$, the position gets updated as in Eq. (30).

$$S_{t+1} = S_t \tag{30}$$

Algorithm 1 shows the pseudo code of presented LU-DA scheme.

Algorithm 1 : Proposed LU-DA algorithm

```
Initializing population
    While end condition is not attained
    Evaluate objective as in Eq. (14)
    Update h , v , a , c , f and b
    Compute C B , O , En and F as in Eq. (21-25)
    Update close by radius
        If random integer ra ≤ ΔS_{t+1} and if ΔS_{t+1} ≤ 0
            Update position as in Eq. (27)
        else if ra ≤ ΔS_{t+1} and if ΔS_{t+1} > 0
            Update position based on proposed female lion update as
            shown in Eq. (28)
        else if ra > ΔS_{t+1}
            Update position as in Eq. (30)
        end If
end while
    New positions are verified on the basis of variable boundaries
end While
```

VI. Result and Discussion

A. Simulation Setup

The adopted LU-DA based CHS in WSN was simulated in MATLAB. The analysis was held by evaluating the alive node count for varied number of round that ranges from 0 t0 2000. Further, log of alive node count was analysed for varied distance that range from 0, 20, 40, 60 and 80. In addition, cost analysis was done for varied iterations that range from 0, 2, 4, 6, 8 and 10. Also, the proposed model was computed over extant approaches such as FF [8], GWO [9], LA [34], DA [23] and FPU-DA [24] and the outcomes were examined in terms of statistical analysis. The simulation parameters in this work are summarized in Table II.

TABLE II. Simulation Parameters

Parameters	Values
"Initial nodal energy	0.5J
Fraction of super nodes amidst advanced nodes	0.6
Network area	100×100
Energy factor of super node	3
Fraction of advanced sensor nodes amidst normal nodes	0.4
Total node count	100
Energy dispersed per bit	100nJ/bit
Data packet aggregation energy	5nJ/bit/message"

B. Analysis on Alive Nodes

The Analysis on NAN of suggested LU-DA scheme over FF, GWO, LA, DA and FPU-DA models is specified in Fig. 3(a), whereas, the log of NAN analysis is shown in Fig. 3(b). The Analysis on NAN is done for 2000 rounds, while, log of NAN is performed for varied distance that range from 0, 20, 40, 60 and 80. In fact, the transmission of data and clustering operations continues for several rounds up to the death of every node. Thereby, the NAN in cluster gets reduced in every cluster. Thus, the NAN reduces with raise in rounds. Till 700th round, the NAN for conventional and proposed models is 100, and it is gradually reduced as the round gets improved. Nevertheless, at 2000th round, the adopted scheme reveals higher NAN than extant models, thus guarantying the enhanced performance of adopted scheme. Particularly, within 75% variation in rounds (i.e. from 500th round to 2000th round), the NAN using presented technique has dropped from 100 to 40. On the other hand, for similar variation (75%) in rounds, the NAN using conventional GWO has dropped from 100 to 19. Thus, the analysis established the enhanced efficacy of LU-DA method with the subsistence of more NAN.

C. Analysis on Normalized Energy

Fig. 4 describes the examination on normalized energy attained using suggested LU-DA model over traditional models for varied number of rounds that ranges from 0 to 2000. The normalized energy is portrayed depending upon the residual network energy and it have to be high for better system performance. In Fig. 4, the network energy seems to be higher at initial rounds; however, with increase in rounds, the energy starts lessening steadily for both adopted as well as compared extant schemes. Especially, from Fig. 4, the presented model has attained higher energy of 0.55 at 1st round, while at 2000th round; the normalized energy value has been dropped to 0.1. However, the adopted model has accomplished a higher energy even at 2000th round, when distinguished over FF, GWO, LA, DA and FPU-DA models. Thus, the capable performance of developed model is confirmed.

D. Convergence Analysis

Fig. 5 describes the convergence analysis of the adopted model over conventional approaches regarding cost. Here, Analysis is performed for a varied number of iterations that ranges from 0, 2, 4, 6, 8 and 10. On noticing the analysis resultants, the developed LU-DA has a negligible cost for all iterations over conventional approaches. Predominantly, on noticing cost values from Fig. 5, the adopted scheme has attained reduced cost value (0.06) from iteration 3 to 10. At the initial iterations (from 1 to 3), the cost of developed model has accomplished a comparatively higher value, while at further iterations; the developed model has converged to a minimal cost value. Especially, at iteration 2, the adopted model is only 60% enhanced than extant FF model, while at iteration 10, the adopted model is 62.5% enhanced than FF model. Thus, the overall assessment shows the impact of the developed LU-DA on better convergence results with increase in iterations.

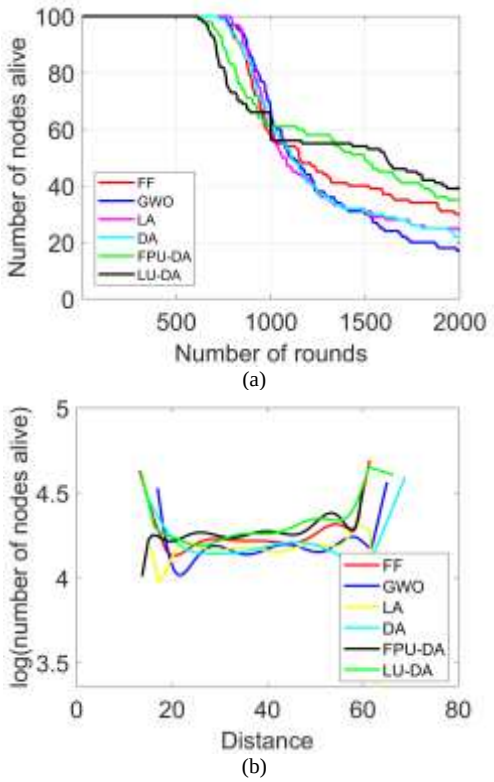


Fig. 3. NAN Analysis and Log of NAN Analysis for Adopted Scheme Over Extant Schemes in Terms of (a) Count of Rounds and (b) Distance.

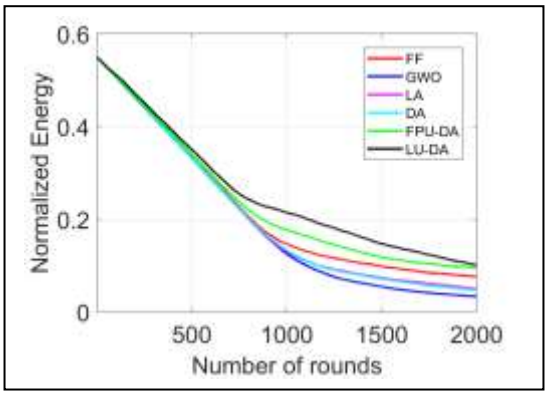


Fig. 4. Analysis on Normalized Energy Attained using Presented Work Over Existing Works.

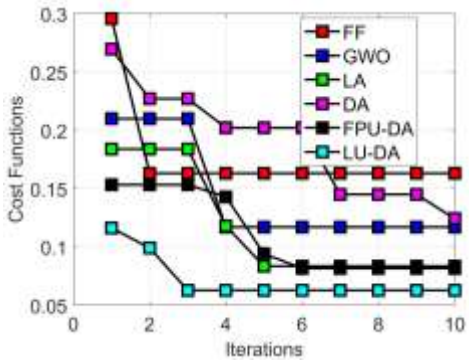


Fig. 5. Convergence Analysis of Developed Scheme Over Traditional Models in Terms of Cost.

E. Statistical Analysis

Table III describes the statistical analysis of the presented LU-DA model over prevailing approaches regarding alive nodes, energy, cost and time. "As meta-heuristic schemes are stochastic in nature, every algorithm is executed for the number of times to attain the statistic of objective function". On noticing the resultants, the adopted LU-DA model has obtained high NAN and energy values and minimal cost and time values for all scenarios, when compared over the existing schemes. Especially, on noticing the NAN from Table III, the best case scenario using proposed LU-DA model has attained superior values over distinguished schemes. Moreover, the cost and time values of conventional schemes is superior to the presented model. Nevertheless, at specific scenarios, the conventional schemes have exposed its enhanced performance regarding time. Though, the entire evaluation of objectives reveals the impact of improving the adopted scheme, thus ensuring a secured transmission.

F. Analysis on Delay, Security and Trust

The analysis on delay, security and trust attained using implemented model over existing models is tabulated in Table IV. Here, examination is done by altering rounds from 0 to 2000. On analysing the delay, the presented LU-DA model has obtained high values for trust and minimal values for delay and security (risk) for all rounds. Initially, at round 1, the delay of adopted scheme seems to be 0.98, while, as the number of rounds increases, the developed method has acquired a minimal delay value of 0.89 at 2000th iteration. Simultaneously, while analysing the security (risk), the developed approach at 1st round has accomplished a minimal risk value of 0.06, whereas, at 2000th iteration, a comparatively higher risk value of 0.094 has been acquired by adopted model. However, for all rounds, the developed approach has acquired fine outcomes than the compared models as per the desired objectives. Also, while examining the trust values, the adopted method has accomplished a higher trust value of 0.32 at 2000th round, whereas, the compared models like FF, GWO, LA, DA and FPU-DA has acquired relatively minimal values of 0.24443, 0.21992, 0.17411, 0.23624 and 0.23446. Therefore, the improvement of LU-DA scheme is confirmed from the outcomes.

TABLE III. STATISTICAL ANALYSIS OF DEVELOPED SCHEME OVER TRADITIONAL SCHEMES REGARDING VARIED METRICS

Alive nodes						
Measures	FF	GWO	LA	DA	FPU-DA [24]	LU-DA
Mean	67.887	64.314	64.819	64.82	70.774	70.693
Best	30	17	24	22	35	39
Median	58	65	61	62	61	56
Worst	100	100	100	100	100	100
STD	28.791	34.082	32.67	32.304	24.449	23
Normalized energy						
	FF	GWO	LA	DA	FPU-DA [24]	LU-DA
Mean	0.21887	0.19672	0.20528	0.20481	0.23513	0.25519
Best	0.076316	0.033484	0.049668	0.046585	0.095484	0.10092
Median	0.14681	0.12504	0.13032	0.13293	0.17645	0.21429
Worst	0.54958	0.54958	0.54958	0.54958	0.54958	0.54958
STD	0.14569	0.1624	0.15523	0.15647	0.13754	0.12921
Cost Function						
	FF	GWO	LA	DA	FPU-DA [24]	LU-DA
Median	0	0.14251	0.13267	0.12611	0.11535	0.017313
Worst	0.38671	0.25673	0.21372	0.23847	0.19976	0.15659
Best	0.023504	0.017313	0.029904	0.017313	0.023528	0
Mean	0	0.14035	0.1317	0.12599	0.11456	0.023129
STD	0	0.032131	0.028361	0.031633	0.027952	0.016704
Time						
	FF	GWO	LA	DA	FPU-DA [24]	LU-DA
Mean	1.7643	1.5862	1.4678	1.6114	4.549	2.0993
Best	1.4428	1.4751	1.4255	1.4872	2.6577	1.1549
Median	1.5791	1.575	1.4588	1.5999	4.4943	1.6068
Worst	6.16	2.0474	2.2782	6.2107	6.888	8.0005
STD	0.58403	0.060094	0.036912	0.12653	0.36098	1.1177

TABLE IV. ANALYSIS ON DELAY, SECURITY AND TRUST: DEVELOPED SCHEME OVER TRADITIONAL SCHEMES

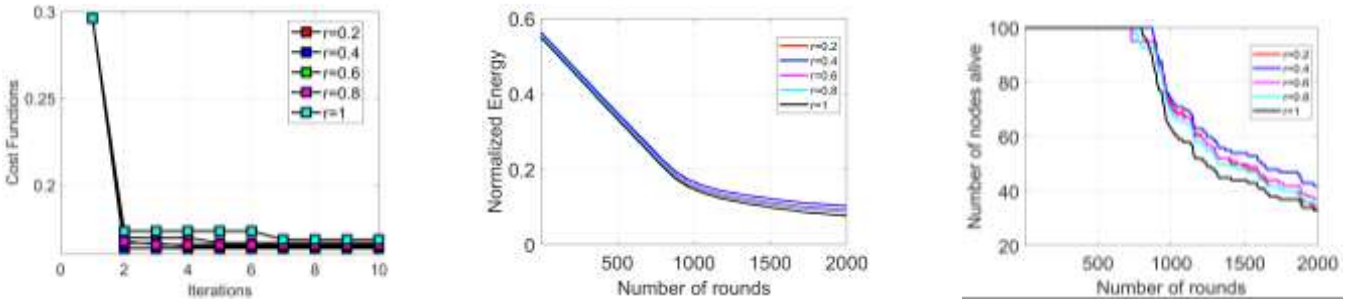
Delay						
Rounds	FF	GWO	LA	DA	FPU-DA [24]	LU-DA
0	0.96697	0.93975	1.0545	1.0547	0.84699	0.98305
100	1.0892	1.1839	1.1871	0.96307	1.1825	0.92846
225	1.1839	0.97482	0.94676	1.1367	1.1019	0.92955
500	1.0665	1.015	1.02	1.015	1.0442	0.93306
725	0.94676	0.89849	1.1963	0.94247	1.0012	0.89247
1000	1.1112	1.0665	0.99593	1.0291	0.98194	1.1112
1225	0.89849	1.0218	1.0545	1.0044	0.81323	1.2899
1500	0.89658	0.93068	0.9879	1.1963	0.72301	0.94676
1726	1.0119	1.015	1.02	0.97447	0.81323	1.0218
2000	0.94767	1.1644	1.0116	1.0044	0.90933	0.893
Security (risk)						
	FF	GWO	LA	DA	FPU-DA [24]	LU-DA
0	0.14793	0.11718	0.2206	0.10757	0.077825	0.065633
100	0.1713	0.18119	0.26347	0.071713	0.12692	0.097313
225	0.16429	0.15659	0.19628	0.1246	0.10803	0.086563
500	0.11718	0.10757	0.34161	0.13545	0.090417	0.085969
725	0.053169	0.15371	0.34767	0.063057	0.018887	0.018656
1000	0.089979	0.13928	0.3868	0.14793	0.072902	0.06253
1225	0.089979	0.12584	0.22542	0.13477	0.067928	0.065784
1500	0.19505	0.13449	0.46562	0.10825	0.077875	0.077313
1726	0.090257	0.11595	0.12679	0.14698	0.14311	0.1272
2000	0.1441	0.14451	0.19878	0.11628	0.11925	0.094513
Trust						
	FF	GWO	LA	DA	FPU-DA [24]	LU-DA
0	0.25457	0.32173	0.17335	0.34906	0.21982	0.25919
100	0.28324	0.2001	0.13852	0.25616	0.22706	0.39341
225	0.21758	0.23228	0.18336	0.21304	0.22499	0.24245
500	0.32646	0.20875	0.48212	0.33051	0.35113	0.3607
725	0.21877	0.25422	0.23838	0.25893	0.31657	0.46922
1000	0.24332	0.25835	0.19616	0.21464	0.30112	0.4071
1225	0.31622	0.20314	0.30029	0.23106	0.24709	0.17652
1500	0.26154	0.14357	0.13765	0.38285	0.25689	0.2413
1726	0.17468	0.16548	0.29837	0.26808	0.2644	0.51033
2000	0.24443	0.21992	0.17411	0.23624	0.23446	0.3193

G. Parametric Analysis

Fig. 6 shows the parametric analysis of normalized energy, convergence, and alive nodes. In the LU-DA scheme algorithm, there is a random number that varies from 0-1. Here, the analysis has been done by varying the random variable r=0.2, r=0.4, r=0.6, r=0.8, and r=1. On observing the results, it can be noticed that when r=0.4, the proposed LU-DA attains the best results. By setting the random variable r=0.4, the above analysis like normalized energy, convergence, and alive nodes has been portrayed by comparing with other existing models.

H. Discussion

This paper presents a new LU-DA model for optimal CHS. Here, the optimal CH selection is carried out by considering the constraints like “energy, delay, distance, security and trust. Here the analysis is performed for alive nodes, normalized energy, convergence analysis, analysis based on delay, security and trust. The betterment of the proposed LU-DA model is proved on various measures like energy and alive node analysis. From the above analysis, it is evident that the proposed model attains better results when compared with the existing models like FF, GWO, LA, DA and FPU-DA. Thereby attaining improved the network lifetime.



(a) (b) (c)

Fig. 6. Parametric Analysis on (a) Convergence Analysis (b) Normalized Energy (c) Alive Nodes.

VII. CONCLUSION

This paper introduced a new LU-DA model for optimal CHS. For optimization, a novel approach termed as LU-DA was developed. Eventually, the primacy of offered method was confirmed over conventional models. The presented model has attained higher energy of 0.55 at 1st round, while at the 2000th round; the normalized energy value has been dropped to 0.1. However, the adopted model has accomplished higher energy even at the 2000th round, when distinguished over FF, GWO, LA, DA and FPU-DA models. At the initial iterations (from 1 to 3), the cost of the developed model has accomplished a comparatively higher value, while at further iterations, the developed model has converged to a minimal cost value. Especially, at iteration 2, the adopted model was only 60% enhanced than the extant FF model, while at iteration 10, the adopted model was 62.5% enhanced than FF model. Therefore, the development of the suggested LU-DA scheme was authenticated over other techniques.

REFERENCES

[1] Alagumuthukrishnan S, Dr. Geetha K, "A Locality Based Clustering and M-Ant Routing protocol for QoS in Wireless Semsor Networks", Department of science and engineering, vol. 6, no. 10, 14 october 2016.

[2] M. Yuvaraja, Sabrigiriraj M, "Lifetime Enhancement of WSN using Energy-Balanced Distributed Clustering Algorithm with Honey Bee Optimization",Asian Journal of research in social sciences and humanities, vol. 6, no. 11, 2016.

[3] Q. Ni, Q. Pan, H. Du, C. Cao and Y. Zhai, "A Novel Cluster Head Selection Algorithm Based on Fuzzy Clustering and Particle Swarm Optimization," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 14, no. 1, pp. 76-84, 1 Jan.-Feb. 2017.

[4] S. H. Kang and T. Nguyen, "Distance Based Thresholds for Cluster Head Selection in Wireless Sensor Networks," IEEE Communications Letters, vol. 16, no. 9, pp. 1396-1399, September 2012.

[5] J. Leu, T. Chiang, M. Yu and K. Su, "Energy Efficient Clustering Scheme for Prolonging the Lifetime of Wireless Sensor Network With Isolated Nodes," IEEE Communications Letters, vol. 19, no. 2, pp. 259-262, Feb. 2015.

[6] Jin Wang, Yiquan Cao, Bin Li, Hye-jin Kim, and Sungyoung Lee, "Particle swarm optimization based clustering algorithm with mobile sink for WSNs, Future Generation Computer Systems, vol. 76, pp 452-457, 2017.

[7] Shishupal Kumar, Nidhi Lal, Vijay Kumar Chaurasiya, "A forwarding strategy based on ANFIS in internet-of-things-oriented wireless sensor network (WSN) using a novel fuzzy-based cluster head protocol", Annals of Telecommunications, vol. 73, no. 9-10, pp 627–638 , 2018.

[8] Turki Ali Alghamdi, "Parametric analysis on optimized energy-efficient protocol in wireless sensor network", Soft Computing > Issue 6/2021, 20-11-2020.

[9] Turki Ali Alghamdi, "Enhanced QoS routing protocol using maximum flow technique", Computers & Electrical Engineering, Volume 89, 2021, 106950, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2020.106950.

[10] Turki Ali Alghamdi, "Underwater Wireless Sensor Network Route Optimization using BIHH Technique" International Journal of Advanced Computer Science and Applications(IJACSA), 11(6), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0110645.

[11] Alghamdi, T.A. Route optimization to improve QoS in multi-hop wireless sensor networks. Wireless Netw (2020). https://doi.org/10.1007/s11276-020-02388-y.

[12] D. Jia, H. Zhu, S. Zou and P. Hu, "Dynamic Cluster Head Selection Method for Wireless Sensor Network," IEEE Sensors Journal, vol. 16, no. 8, pp. 2746-2754, April15, 2016.

[13] Pawan Singh Mehra, Mohammad Najmud Doja, and Bashir Alam, "Fuzzy based enhanced cluster head selection (FBECS) for WSN", Journal of King Saud University - Science, Available online, 2018.

[14] Awais, Muhammad & Ali, Ishtiaq & Alghamdi, Turki & Ramzan, Muhammad & Tahir, Muhammad & Akbar, Mariam & Javaid, Nadeem. (2020). Towards Void Hole Alleviation: Enhanced GEographic and Opportunistic Routing Protocols in Harsh Underwater WSNs. IEEE Access. 10.1109/ACCESS.2020.2996367.

[15] Sara Al-Sodairi, and Ridha Ouni, "Reliable and energy-efficient multi-hop LEACH-based clustering protocol for wireless sensor networks", Sustainable Computing: Informatics and Systems, vol. 20, pp 1-13, 2018.

[16] Gaurav Kumar Nigam, and Chetna Dabas, "ESO-LEACH: PSO based energy efficient clustering in LEACH", Journal of King Saud University - Computer and Information Sciences, Available online, 2018.

[17] R. Raj Priyadarshini, and N. Sivakumar, "Cluster head selection based on Minimum Connected Dominating Set and Bi-Partite inspired methodology for energy conservation in WSNs", Journal of King Saud University - Computer and Information Sciences, Available online, 2018.

[18] Khalid A. Darabkh, Saja M. Odetallah, Zouhair Al-qudah, Ala' F. Khalifeh, and Mohammad M. Shurman, "Energy-Aware and Density-Based Clustering and Relaying Protocol (EA-DB-CRP) for gathering data in wireless sensor networks", Applied Soft Computing, vol. 80, pp 154-166, 2019.

[19] Shilpa Mahajan, Jyoteesh Malhotra, and Sandeep Sharma, "An energy balanced QoS based cluster head selection strategy for WSN", Egyptian Informatics Journal, vol. 15, no. 3, pp 189-199, 2014.

[20] Muthukumaran K, Chitra K, and Selvakumar C, "An energy efficient clustering scheme using multilevel routing for wireless sensor network", Computers & Electrical Engineering, vol. 69, pp 642-652, 2018.

[21] G. Kannan, and T. Sree Renga Raja, "Energy efficient distributed cluster head scheduling scheme for two tiered wireless sensor network", Egyptian Informatics Journal, vol. 16, no. 2, pp 167-174, 2015.

[22] Palvinder Singh Mann, and Satvir Singh, "Improved metaheuristic based energy-efficient clustering protocol for wireless sensor networks", Engineering Applications of Artificial Intelligence, vol. 57, pp 142-152, 2017.

[23] M. Marsaline Beno, Valarmathi I. R, Swamy S. M and B. R. Rajakumar, "Threshold prediction for segmenting tumour from brain MRI scans", International Journal of Imaging Systems and Technology, Vol. 24, No. 2, pages 129-137, 2014, DOI: https://doi.org/10.1002/ima.22087.

[24] Alghamdi, T.A. Energy efficient protocol in wireless sensor network: optimized cluster head selection model. Telecommun Syst 74, 331–345 (2020). https://doi.org/10.1007/s11235-020-00659-9.

[25] Prachi MaheshwariAjay K. SharmaKaran Verma, "Energy efficient cluster based routing protocol for WSN using butterfly optimization algorithm and ant colony optimization", Ad Hoc Networks, vol.110 (Cover date: 1 January 2021), Article 102317, 6 October 2020.

[26] Reeta BhardwajDinesh Kumar, "MOFPL: Multi-objective fractional particle lion algorithm for the energy aware routing in the WSN", Pervasive and Mobile Computing, vol. 58, Article 101029, August 2019.

[27] Augustine, S., Ananth, J.P. "Taylor kernel fuzzy C-means clustering algorithm for trust and energy-aware cluster head selection in wireless sensor networks", Wireless Netw 26, pp. 5113–5132, 2020. https://doi.org/10.1007/s11276-020-02352-w.

[28] Pratik Goswami, Ziwei Yan, Amrit Mukherjee, Lixia Yang, Sidheswar Routray, and G. Palai, "An energy efficient clustering using firefly and HML for optical wireless sensor network", Optik, vol. 182, pp 181-185, 2019.

[29] AmanjotSinghToor, A.K.Jain," Energy Aware Cluster Based Multi-hop Energy Efficient Routing Protocol using Multiple Mobile Nodes (MEACBM) in Wireless Sensor Networks", AEU - International Journal of Electronics and Communications, vol.102, pp.41-53, April 2019.

[30] S. M. M. H. Daneshvar, P. Alikhah Ahari Mohajer and S. M. Mazinani, "Energy-Efficient Routing in WSN: A Centralized Cluster-Based Approach via Grey Wolf Optimizer," in IEEE Access, vol. 7, pp. 170019-170031, 2019, doi: 10.1109/ACCESS.2019.2955993.

[31] WangTianshu, Zhang Gongxuan, YangXichen, VajdiAhmadreza," Genetic algorithm for energy-efficient clustering and routing in wireless sensor networks", Journal of Systems and Software, vol.146, pp.196-214, December 2018.

[32] Nejla RouissiHamza GharsellaouiSadok Bouamama, "Improvement of Watermarking-LEACH Algorithm Based on Trust for Wireless Sensor Networks", Procedia Computer Science, vol. 159 (Cover date: 2019), pp. 803-813, 14 October 2019.

[33] Mohammad Jafari and Mohammad Hossein Bayati Chaleshtari, " Using dragonfly algorithm for optimization of orthotropic infinite plates with a quasi-triangular cut-out", European Journal of Mechanics A/Solids, vol. 66, pp.1-14, 2017.

[34] Rajakumar Boothalingam, "Optimization using lion algorithm: a biological inspiration from lion's social behavior", Evolutionary Intelligence, 7 September 2018.

[35] Renjith Thomas and MJS. Rangachar, "Hybrid Optimization based DBN for Face Recognition using Low-Resolution Images", Multimedia Research, Vol.1,No.1, pp.33-43,2018.

[36] Moresh Madhukar Mukhedkar,Uttam Kolekar, "Hybrid PSGWO Algorithm for Trust-Based Secure Routing in MANET", Journal of Networking and Communication Systems, Vol.2,No.3, pp.1-10,2019.

[37] Rupam Gupta Roy, "Rescheduling Based Congestion Management Method Using Hybrid Grey Wolf Optimization - Grasshopper Optimization Algorithm in Power System", Journal of Computational Mechanics, Power System and Control, Vol.2,No.1, pp.9-18,2019.

Risk Assessment Methods for Cybersecurity in Nuclear Facilities: Compliance to Regulatory Requirements

Lilis Susanti Setianingsih[1], Reza Pulungan[2], Agfianto Eko Putra[3], Moh Edi Wibowo[4], Syarip[5]
Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia[1,2,3,4]
BAPETEN, Nuclear Energy Regulatory Agency, Jakarta, Indonesia[1]
BATAN, National Nuclear Energy Agency, Yogyakarta, Indonesia[5]

***Abstract*—As strategic infrastructures, nuclear facilities are considered attractive targets for attackers to commit their malicious intention. At the same time, for efficiency, those infrastructures are increasingly implemented, equipped with, and managed by digitally computerized systems. Attackers, therefore, try to realign their attack scenarios through such cyber systems. It is crucial to understand various existing risk assessment methods for cybersecurity in nuclear facilities to prevent such attacks. Risk assessment is designed to study the nature of the originated attack threats and the consequences implied. This paper studies a series of risk assessment methods implemented for security related to cybersecurity of strategic infrastructures, including nuclear facilities. Extended from cybersecurity, the required concepts in nuclear security cover defense-in-depth, synergy of safety and security, and probabilistic safety/risk assessment. Selecting cybersecurity risk assessment methods should integrate these three essential concepts in their evaluation. This paper highlights the suitable and appropriate risk assessment methods that meet security requirements in the nuclear industry as specified in the national and international regulations.**

***Keywords*—Risk assessment; cybersecurity; nuclear facilities; security requirements; regulatory requirements**

I. Introduction

As critical infrastructures containing extremely hazardous materials, nuclear facilities have to operate flawlessly to avoid predicaments and unwanted catastrophes. Even a small failure cannot happen in such facilities, let alone mistakes or oversights committed by their operators. To reach this high level of standards, the International Atomic Energy Agency (IAEA) has provided guidance and recommendations for the management of nuclear facilities. At the same time, IAEA member states are obliged to follow the IAEA regulatory frameworks in providing in-line regulation, authorization, licensing, and inspection to ensure compliance to nuclear energy implementation to national and international standards. Nowadays, concerns regarding nuclear facilities include not only nuclear safety and nuclear safeguard but also nuclear security. Cybersecurity has emerged as a crucial issue among the different aspects of nuclear security as more hardware and software are composed of cyber-physical systems in nuclear facilities.

Cybersecurity is a fundamental issue not only in nuclear facilities but also in any computer-based systems in general. IAEA has provided a guideline for computer security [1] to cope with cyberattacks that can potentially penetrate information systems at nuclear facilities. Cybersecurity focuses on the protection of assets, including humans, data, systems, and organizations, using recent developments of digital technology and follows technical guidance stated by government regulations [2], [3], [4], [5]. Cybersecurity also analyzes risk information [6], [7], including threats, vulnerabilities, and adversaries, to anticipate various cyberattack scenarios. A cybersecurity plan is constructed to ensure information preservation in terms of confidentiality, integrity, availability, authenticity, and validity [2], [8], system's robustness and (fault) tolerance, and system's protection from any potential attacks [9], [10], [11]. Such a plan, according to the US Nuclear Regulatory Control (USNRC) [12], has to cover: (1) components supporting safety-related functions, (2) elements that contribute as important-to-safety functions, (3) competence to maintain security functions, (4) capability to perform emergency preparedness functions, covering off-site communications, and (5) sustainability of support systems and equipment. Considering the critical role of nuclear facilities, these facilities are expected to implement the best model and the best security practice. Achieving this purpose is not trivial, and more studies must be dedicated to understanding this issue.

The development of precautionary and preventive measures has become an important approach in cybersecurity. In nuclear facilities, such measures are far more preferred than the detection and mitigation of cyberattacks. Cybersecurity risk assessment comprises the continuous process of identifying, analyzing, and evaluating any possible risks in a cybersecurity system. Conforming to the cybersecurity requirements, it considers potential risks, consequences of emerging threats, and acquiring cost due to the consequences [13]. Risk assessment, in this context, plays an important role that can assist the understanding, analysis, and evaluation of risks [14] exposed by critical infrastructures. Therefore, various impacts caused by undesired events and attacks can be calculated, anticipated, and minimized [15].

This paper presents a study on risk assessment methods of cybersecurity that have been recently proposed for nuclear facilities, with the purpose of providing a comprehensive understanding of how risk assessments have typically been conducted. It is expected to serve as helpful information for nuclear regulatory agencies in performing their task for regulatory control. Its scope is to provide a unified, collective overview of the current state of knowledge and highlight an inclusive

foundation on risk assessment methods for cybersecurity in nuclear facilities. This study is limited to the compliance and conformity of selected suitable risk assessment methods to the nuclear energy regulatory system.

The rest of the paper is organized as follows: Section II discusses related work on cybersecurity and cyberattacks in general and specifically in nuclear facilities. Section III highlights cybersecurity risk assessments in critical infrastructures, comprising the need for synergy of safety-security and the application of the defense-in-depth concept in security aspects. Section IV focuses on assorted methods of cybersecurity risk assessment methods recently proposed for nuclear facilities. The section also discusses the conformity of the selected risk assessment methods to the regulatory aspects, specifically those concerning nuclear facilities. Section V concludes the paper and offers future works.

II. Related Work

A. Cybersecurity and Cyberattacks

Cybersecurity deals with risks and is expected to meet its objective in protecting assets, including humans, data, systems, and organizations. Cybersecurity, to this end, involves the management of five aspects: data/information, software/hardware, procedures, human resources, and communication, each of which plays an essential role in establishing system integrity.

Table I provides summarized descriptions of the five aspects of cybersecurity management. Data encryption serves to enhance cybersecurity by eliminating the possibility of eavesdropping, data falsification, and data tempering [16], [17], [18]. Software and hardware must follow a strict quality assurance, for example, to avoid getting infected by viruses or worms from wider networks [2], [19]. Cybersecurity procedures include strategies, security principles, security guarantees, guidelines, and risk management approaches [20]. Human resources relate to personnel that has to be equipped with skills and knowledge that are up-to-date and are continuously refreshed through training [21] (otherwise, the personnel can also turn into a threat [22]). Communication integrates all of the aspects mentioned earlier in cybersecurity systems.

TABLE I. Components of Cybersecurity

Component	Description	Precautions to cyberattacks
Data	Information being transferred within the network or internetwork.	Use encrypted data [23], [24].
Software or hardware	The software provides packages of programs, operating systems, including platforms to control the instrumentation and control, which are vulnerable to attacks [25].	Obtain software and hardware from various providers [2]. Secure updating software in accordance with quality assurance (QA) [26].
	Hardware comprises smaller parts as in I&C components and larger devices as the assembly of microprocessors or other peripherals.	Periodic maintenance and regular inspection of hardware [2].
Procedures	Regulations both national and international to follow.	Meet all the requirements as regulated [26], [27].
Human resources	Personnel in charge for the system to run smoothly.	Providing adequate training and regular refresh training [2].
Communication	Interaction inter-device, network, and human-machine interface.	Providing authenticity for communication [28], [23].

All measures implemented in cybersecurity are mainly prepared to anticipate cyberattacks appearing in various forms. Cyberattacks can be defined as attacks on information systems [19] through intrusion conducted by internal [4] or external malicious attackers [16] that may compromise the confidentiality, integrity, and availability [29] of the system or may result in failure and property loss [30], [31] leading to jeopardized safety functions [16], [4], [6]. Cyberattacks have been classified into three categories: active attacks, passive attacks, and cyberwars. Active attacks refer to activities that compromise information systems, including reconnaissance attacks, access attacks, cybercrime, cyber espionage, cyber terrorism, malicious and non-malicious attacks on mobile ad hoc networks and wireless sensor networks. On the other hand, passive attacks do not involve compromising systems but are more on retaining critical information for further use. Most cyberattacks start from cyber scanning followed by enumeration, intrusion attempts, the elevation of privilege, performing malicious tasks, deploying malware/backdoor, deleting forensic evidence, and exiting. Cyber-criminals apply attacks in the forms of cyber vandalism, hacking, denial of service, hijacking the domain name, and even spreading infectious viruses.

In nuclear facilities, cybersecurity issues mainly come from the connectivity between the cyber and the physical systems in the facilities [2]. Current instrumentation and control (I&C) devices are mostly connected to cyber-physical systems. They are distributed control systems (DCS) comprising digitalized automated controllers distributed within the systems, implementing a geographically distributed control loop, and having four main components: controllers, distributed controllers, human-machine interfaces, and communication channels. Nowadays, treating cyber connectivity and physical systems separately is no longer in favor due to more sophisticated and varied cyberattacks [3], [5]. Even though most nuclear facilities are not directly connected to the cyber networks, cyberattacks may still violate system protection [28] in varied forms [32], including denial of services.

B. Cyberattacks at Nuclear Facilities

History has recorded cybersecurity attacks and attempts of attacks at nuclear facilities around the world. Reports have been submitted to IAEA as the international atomic regulatory agency to oversee nuclear energy utilization and ensure the concept of safety, security, and safeguard in nuclear energy implementation.

The David Besse nuclear power plant in the United States was attacked through cyber activity in 2003 [19], [33], leading to the loss of displayed data related to safety and non-safety system for up to five hours. The Slammer worm infected the enterprise workstation through a consultant's network, causing a clogged data traffic connection. The control room personnel could not get the vision of the safety parameter display for four hours fifty minutes.

In 2006, a cyberattack targeted the Browns Ferry nuclear facility in Alabama, United States. The attack resulted in a reactor shutdown because the pump regulating the circulation of demineralized condensate water failed to perform its functions [33]. A programmable logic controller (PLC) controlled the demineralized condensate water, and a variable frequency drive

modulated the circulating pump's speed. Both utilized devices that communicate through a local area network, neglecting the high data traffic that the I&C system could not handle, forcing it to cause malfunction to the PLC and the variable frequency drive. As the recirculating pump was critical to supplying coolant to the reactor, its malfunction could lead to a reactor core meltdown if it failed to support the cooling process. The operator had to manually shut down the operating reactor to avoid further undesired events.

In 2008, the Edwin I. Hatch nuclear facility in Georgia, United States, suffered from a serious incident during a software updating process by an employee [33] on a personal computer in an enterprise network while it was occupied for data input collection to the I&C systems. It led to reset data that should appear on the I&C network. Due to the loss of data display, the systems then considered it an emergency by immediately shutting down to protect the nuclear reactor.

Iran's Natanz uranium enrichment facility suffered from the Stuxnet attack in 2010 [16], [28], [19], [33]. A combination code triggered a PLC in its process control system to send a list of commands to its frequency converter that changed the maximum rotation frequency of the centrifuge, causing the centrifuge to rotate out of its designed range and the rotation speed frequently changed. Due to its operation exceeding the originally designed operational range, the affected centrifuges wore out significantly, reducing the operation period's life and eventually damaging the physical system. The Stuxnet covered its ability by not disrupting the control system's sensor output [51]. It faked the output display and did not interfere with the PLC but gave instructions and commands. It required no connection to the cyber system but used a memory stick instead, plugged into the internal network.

III. Recent Development of Cybersecurity Risk Assessment at Critical Infrastructure

Table II depicts a list of surveyed papers about security, cybersecurity, and risk assessment. Some of them are related to critical infrastructures and, in some cases, to nuclear facilities. As part of the energy sector, nuclear facilities and the electrical supply are classified as critical infrastructures along with finance, transportation, oil and gas industries, water distribution, health care, government services, and emergency installation [52]. However, unlike other critical infrastructures, nuclear facilities are often more attractive to become targets of malicious attacks by different attackers.

A. Synergy of Safety and Security

One major difference distinguishing critical infrastructures such as nuclear facilities from other infrastructures is the requirement to maintain the synergy of security and safety [36], [43]. Any measures taken for security, in this case, should consider their impacts on safety while safety systems in all phases of nuclear energy implementation should not be disturbed. Even though security and safety may oppose one another—for example, security provisions should be kept confidential while safety procedures are to be published as widely as possible—the primary objectives are similar, namely protecting human beings. In any case of the procedures to keep up the synergy of safety and security, safety should be prioritized.

As most safety and security systems are now performed in digital equipment, cybersecurity risks can present in any interface between the two systems. Cybersecurity risk assessment should focus on the people, processes, and equipment related to safety and security. Safety and security digital systems should at least consist of operational technology, as in I&C and information technology. I&C relates to both safety and security systems, while information technology concerns the security system [53].

B. Defense-in-depth Concept

Defense-in-depth is a popular term used in the safety and security field in the nuclear industry. Kim *et al.* [16] initiated a defense-in-depth strategy to strengthen system information and event management (SIEM) in detecting cyberattacks. While dealing with security issues, defense-in-depth SIEM (DID-SIEM) considers all constraints and requirements of nuclear facilities to maintain its safety aspects. DID-SIEM has managed to alleviate technical constraints that can become barriers to security measures. One of the most important technical constraints is that safety function becomes the top priority over security.

Within the DID-SIEM framework, the network is separated into safety, non-safety, and security control levels. The industrial control system network is distantly isolated from the office/enterprise network to eliminate external attack options. Under this security level, no data transfer can be shifted from a lower level to a higher one. It only allows one-way data transmission from safety to non-safety networks. Higher-level DID can deliver command or information to a lower level, but not the other way around. A safety system is isolated within the level where it can share or relay information to a non-safety-related network assigned at a lower level. The monitoring visualization system can only receive data transmission for display from both non-safety and safety log collection and analysis systems.

The defense-in-depth concept can be elaborated into layers of leveled obstructions to prevent any attacks targeting the facilities. The obtained barriers create delays for attackers in accomplishing their missions. These delays can give those in charge of the facilities time to anticipate and prevent the attacks from escalating.

IV. Results and Discussion

A. Cybersecurity Risk Assessment in Nuclear Facilities

Risk can be defined as a combined frequency or probability and consequences of an event or incident that may compromise a system [20], [14]. Another way of expressing risk is the likelihood that a certain vulnerability of a particularly attractive object as the target will be manipulated by a certain threat leading to undesired consequences [54]. Risk can be formulated as a function of (1) the threat for any attack to occur, (2) the vulnerability of the targeted object to endure the attack, and eventually, (3) the damage caused by the threat attack [20], [55], [49]. In a cybersecurity concept, cyber risk is associated with the risk of operational activities in cyberspace, in which the impact can threaten the information systems and assets, the information and communication technology, devices, and peripheral technology resources, and can create

TABLE II. SURVEYED PAPERS ON CYBERSECURITY

Reference	Method	Risk assessment	Cyber-attacks	Critical infrastructure	Nuclear facility
(Kim *et al.*, 2021) [16]	DID-SIEM	Yes	No	Yes	Yes
(Ali, 2021) [2]	Cybersecurity management for DCS	Yes	Yes	Yes	No
(Wang *et al.*, 2021) [34]	Social engineering in cybersecurity	Yes	No	No	No
(Lee *et al.*, 2021) [35]	Estimating security state	Yes	Yes	Yes	Yes
(Yang *et al.*, 2021) [36]	Safety-security risk analysis	Yes	Yes	Yes	No
(El-Genk *et al.*, 2021) [37]	Cybersecurity investigation	Yes	Yes	Yes	Yes
(Yousefnezhad *et al.*, 2020) [38]	Security product lifecycle in IoT	Yes	No	No	No
(Park & Lee, 2020) [6]	Quantitative risk assessment based on difficulty and consequences	Yes	Yes	Yes	Yes
(Lee *et al.*, 2020) [30]	Identifying fault-prone cybersecurity control	Yes	Yes	Yes	Yes
(Syed, 2020) [39]	Cybersecurity vulnerability management	Yes	No	No	No
(Peterson *et al.*, 2019) [40]	Cybersecurity vulnerability assessment	Yes	Yes	Yes	Yes
(Oughton *et al.*, 2019) [41]	Stochastic counterfactual risk analysis	Yes	Yes	Yes	No
(Vessels *et al.*, 2019) [42]	Cybersecurity risk assessment for space systems	Yes	Yes	Yes	No
(Hahndel & Draeger, 2019) [43]	Risk assessment for safety and security	Yes	Yes	Yes	No
(Hashim *et al.*, 2018) [44]	Risk assessment for insider threat	Yes	Yes	No	No
(Lallie *et al.*, 2018) [45]	Evaluating cybersecurity attack graph	Yes	Yes	No	No
(Leszczyna, 2018) [46]	Cybersecurity standards for smart grid	No	No	Yes	No
(Ivanchenko *et al.*, 2018) [47]	Risk assessment of critical infrastructures	Yes	Yes	Yes	No
(Porcedda, 2018) [48]	Regulatory framework on cybersecurity	No	No	No	No
(Zio, 2018) [49]	Future risk assessment	Yes	Yes	Yes	No
(Daria & Massel, 2018) [20]	Risk identification of cybersecurity violation	Yes	Yes	Yes	No
(Lee *et al.*, 2018) [50]	Evaluating the efficacy of cybersecurity control	Yes	Yes	Yes	Yes
(Jana & Ghosh, 2018) [7]	IT2FLC risk assessment model for cybersecurity	Yes	Yes	Yes	No

damage to the tangible and intangible materials [56]. By managing information security risks, good information security practices in cybersecurity are expected to maintain reliable services by the system [57].

Risk assessment plays an important role in understanding and evaluating risks [14] to ensure cybersecurity and to calculate impacts caused by undesired events [15]. Therefore, risk assessment for cybersecurity comprises identifying threats, vulnerabilities, and property assets available within the attack targets [58] and is intended to minimize the negative impacts of potential threats. As the demands for cybersecurity increase to secure data, peripherals, and systems, the need for risk assessment on cybersecurity, especially those implemented at critical infrastructures, including nuclear facilities, also increases.

Table III lists selected risk assessment methods used in nuclear facilities that will be discussed further in this paper. The selected methods relate to security, particularly cybersecurity, in nuclear facilities and critical infrastructures, such as space systems [42] and distributed control systems [2].

TABLE III. CYBERSECURITY RISK ASSESSMENT METHODS FOR NUCLEAR SECURITY

Reference	Methods applied
(Kim *et al.*, 2021) [16]	Method of risk assessment to detect cyberattack; then manage, collect, analyze security information and events, called system information and event management (SIEM) defense in depth. Defense-in-depth is a strategy implemented for nuclear security.
(Lee *et al.*, 2021) [35]	Estimating security state for an integrated response to cyberattacks in nuclear power plants.
(El-Genk *et al.*, 2021) [37]	Cybersecurity investigation.
(Park & Lee, 2020) [6]	Risk assessment is performed by evaluating the risk based on the difficulty and consequences caused by cyberattacks through implementing a Bayesian belief network and probabilistic safety assessment.
(Lee *et al.*, 2020) [30]	Development of a quantitative method for estimating fault-proneness of cybersecurity controls.
(Peterson *et al.*, 2019) [40]	Cybersecurity vulnerabilities in nuclear power plants.
(Lee *et al.*, 2018) [50]	Mean Time To Compromise (MTTC) quantifies the abstract concept of efforts linked to Common Vulnerability Scoring System (CVSS). MTTC value is the estimated figure of the time required for the valid attack. CVSS is one method assigning numerical scores to vulnerabilities based on severity and difficulty to exploit.

B. Estimating Security State

Lee *et al.* [35] developed a probabilistic safety assessment to assist operators in conducting safety-related security evaluations. This method allows operators to conduct cause investigation and security impact analysis. The developed security state evaluation can calculate security failure probability, accounting for the damage probability of critical data assets. Quantification of cyberattack-induced impact is typically carried out by fault tree and event tree analysis. An initiating or basic event relates to the response function failure in the event tree analysis. In contrast, the top event represents the control's functional failure linked to the basic events using logical gates. The top event's probability can be calculated provided that all basic events' probabilities are available, typically obtained from fault tree analysis.

Security state evaluation can lead to the estimation of functional performance impact due to cyberattacks progress. Security state estimation proceeds by observing the target system, developing a hidden Markov model based on the observation, and inputting the model to the evaluation module. The attack level is determined for its minimum and maximum values after being estimated by the evaluation module. The decoding module then uses the selected model from the evaluation module to provide a state-transition path to estimate the current security state.

C. Risk Assessment for Difficulty and Consequences of Cyberattacks

Each cyberattack scenario has its difficulty and produces different impacts that depend on the existing protection systems. This observation gives rise to methods that assess the difficulty and consequences of cyberattacks and consider various aspects from the adversary's point of view. Park and Lee [6]

demonstrated a risk evaluation method based on difficulties and consequences of cyberattacks that combines Bayesian belief network (BBN) and probabilistic safety assessment (PSA) [4].

The BBN method offers quantitative measurements for the attack difficulties level by considering the number of targets and cyberattack scenarios for calculating the conditional cyberattack probability. The number of targets indicates the vulnerability points of the system. Conditional cyberattack probabilities can incorporate the vulnerabilities and failure modes due to the cyberattacks. PSA, on the other hand, evaluates the consequences of the assessment. Using the selected basic events constructed as event trees or fault trees, PSA uses the Boolean logic in analyzing the sequence of basic events to a system failure.

This research also identified several types of cyberattacks based on previous reports of incidents and accidents in nuclear history. These include attacks on man-machine interface systems, attacks related to errors on omission or errors on commission, and attacks that lead to blocked information as well as incorrect displayed information.

D. Fault-proneness in Cybersecurity Control

Lee *et al.* [30] developed a quantitative method to estimate fault-proneness in cybersecurity control by implementing: (1) an analysis of fault prediction models, (2) adoption of the software change entropy model, and (3) development of the security control entropy model. Achievement of high-level quality assurance through consistent attempts, fault proneness, and software complexity correlates to the focus of study for this particular method. Fault proneness can be predicted by software modules exploiting software complexity and fault data recorded in history. Cybersecurity control in the nuclear industry can use the change entropy model to identify the fault-prone in planning and preparing for future issues. To do so, it needs a definition of the amount and complexity of information regarding cybersecurity control.

Nuclear facilities, including nuclear power plants, must provide cybersecurity control in their I&C systems covering intrusion detection systems, surveillance, and access control [19], [35], [37], [40]. For that purpose, the I&C system needs to be modified and extended to comply with the security requirements. The more robust and complex the system coupling is, the more vulnerable are the operating system and application programs to cyberattacks. A proper software development life cycle is expected to increase the security level, although it requires efforts to provide quality assurance for keeping the fault-proneness at a minimum. Information on the software development life cycle can be used to estimate the software failure probability from the number of hidden faults and fault activation probabilities. Software failure probability is then used to anticipate the damage in critical data assets that can jeopardize the safety functions of the nuclear facility.

E. Cybersecurity Vulnerability Assessment

Peterson *et al.* [40] suggested that a risk assessment on a nuclear facility should cover [12]: (1) digital system inventory, (2) penetration testing, (3) vulnerability database and software, (4) modernized discussion of risk, and (5) ongoing risk assessment. They also noted that the assessment should also include the vulnerability assessment. Assessment of cybersecurity vulnerability at nuclear facilities assumes that most incidents occur due to insufficient cybersecurity procedures or unintentional avoidance of the facilities' security measures. History of successful cyberattacks in nuclear facilities, in some manners, involved ingenuous insider participation and the digitalization of I&C in nuclear facilities. Thus, it is vital to pay special attention to such particular attack vectors involving humans, insiders, or technological changes.

F. Cybersecurity Investigation

El-Genk *et al.* [37] noted that the main concern arising in the digitalization of I&C systems in nuclear facilities is the vulnerability of being targeted by cyberattacks. PLCs for I&C in pressurized water reactors face the threat of disturbance caused by cyberattacks, which can manipulate data display collected from data sensors used for safety monitoring. Such an attack can be executed through a false data injection attack. They strongly advised that nuclear facilities need to provide high fidelity analyses in investigating the response and identifying the vulnerabilities to the threat of cyberattacks.

The method of this cybersecurity investigation is applied for program emulated for PLC in the physics-based transient prototype of pressurizer. The pressurizer adapts and controls both system pressure and required water level in a pressurized water reactor type. Setpoints of pressure magnitude and accustomed water level within the pressurizer are preprogrammed. The PLC performs any opening water spray nozzle changes while controlling the charging and adjusting speed levels of letdown water. The on/off position is based on the command of electrical power changes controlled by a PLC.

Such cybersecurity investigation can also be implemented for cases other than pressurizers within the PLC and I&C systems. History showed that several cyberattacks in nuclear industries interfered with the sensor display, leading to immediate shutdown compromising the safety-supporting system.

G. Intrusion-tolerance-based Cybersecurity Index

Another risk assessment method was proposed by Lee *et al.* [50], namely the intrusion-tolerance-based cybersecurity index (InTo-CSI). This method is performed through the reduction of the ratio probability that a cyberattack can damage the target. As safety is the primary concern in nuclear facilities, the intrusion tolerant concept can be a popular option in the evaluation method. Attack difficulty is used to determine the failure probability of intrusion-tolerant strategy in terms of resistance strategy [7], [47]. Attack difficulty depends strongly on unexpected and abstract factors covering attackers' skills and ability to access target system information. Quantifying abstract attempts to attack can be modeled by mean time to compromise (MTTC) based on the assumption of time required for an attack to proceed. Later, MTTC is linked to a common vulnerability scoring system (CVSS) to examine the scores of vulnerabilities based on their severity and level of difficulty to exploit.

The InTo-CSI can be calculated based on the failure probability of each state of the security system by taking into account the failure probability of the existing system from cyberattacks and the failure probability of the upgraded

system. In the upgraded system, the failure probability of securing the system tends to be smaller than that of securing the system before upgrading. There are five intrusion tolerant strategies that should be considered in using the InTo-CSI: (1) a resistance strategy to see the vulnerability difficulties, (2) a detection strategy in detecting a valid attack during the exploitation phase, (3) a backup strategy to provide redundancy in case of error service, (4) an elimination strategy to reduce the risk sources, and (5) a graceful-degradation strategy to keep the essential system functions despite the degraded less important functions.

H. Discussion

From the papers discussed in Sections III and IV, several cybersecurity aspects can be highlighted:

1. The defense-in-depth concept has been recognized as a comprehensive approach for keeping cyberattacks' impacts at a minimum level.
2. The synergy of safety and security has become one of the most important basic requirements in the nuclear industry. Protection of human beings, in this case, comes before protection of properties and assets.
3. The probabilistic approach has emerged as a prominent method of cybersecurity risk assessment. The probabilistic approach can help evaluate the consequences based on the likelihood or probability of the fault propagation as composed in attack scenarios [59].

Draeger & Hahndel [43] proposed a unified risk assessment for both aspects to accommodate the need to maintain a balance between safety and security. The proposed framework provides a simulation that generates paths of sequenced state events. Risk measurement based on the paths is then conducted to predict the criticality and probability of each branch leading to its successor state. The risk assessment itself is modeled to be dynamic and time-dependent [60]. It covers both sides of time response for the defender as the target and the time frame available for the attacker to perform their action in threatening the system [61].

Probabilistic risk assessment (PRA) is an assessment method inspired by the probabilistic safety assessment (PSA) and commonly implemented in nuclear facilities. It provides a combination of quantitative and graphical analysis (fault trees or event trees) to ensure system protection. It can express the basic manifestation of possible attacks scenarios, indicate vulnerabilities, and assist with preparing anticipation before the occurrence of attacks. Since cyberspace is subjected to severe attacks with drastic consequences at a very high speed and complete anonymity [62], modeling threats and vulnerabilities using a probabilistic approach in risk assessment might deliver a better prospect for the entire security system. PRA also includes analysis of adversaries that can be performed based on their capability, opportunity, and intent to build behavioral characteristics [18].

Past experiences in the nuclear industry and several other critical infrastructures, cyberattacks variedly depending on the initial intention of the attackers [17], [63]. In attempting attacks, the adversaries always try to find vulnerabilities in the protected systems. The capability of the adversaries determines the severity level of the impact after attacks. The adversaries can consist of terrorists, criminals, extremists or demonstrators, outsider agents, and insider agents [18], [64], [44], [65], [66], with their respective capacity and capability based on their financial and technical assets. The better the capacity of the adversaries to execute the attack, the more significant is the probability of the attack succeeding.

The selected candidates of cybersecurity risk assessment for nuclear facilities are compared to examine their compliance to the nuclear security aspects required in all facilities. Table IV highlights the conformity of each risk assessment candidate to the three aspects: defense-in-depth, the synergy of safety and security, and implementation of PSA/PRA. All three aspects are highly recommended to be considered in developing or selecting cybersecurity risk assessment. As can be seen from the table, all the listed cybersecurity risk assessment methods have considered the synergy of safety and security during the assessment process. In most methods, the procedures employ distinct parts of assessment for each safety and security-related section.

TABLE IV. ASPECTS OF NUCLEAR SECURITY CONFORMITY

Method	Ref.	DID	Synergy Safety & Security	PSA/PRA
DID-SIEM	[16]	√	√	
Security state estimation	[35]		√	√
Cybersecurity investigation	[37]		√	
BBN PSA/PRA	[6]		√	√
Fault proneness estimation	[30]		√	
Cybersecurity vulnerability assessment	[40]	√	√	√
MTTC-CVSS InTo-CSI	[50]	√	√	√

In analyzing the balance between cybersecurity measures and safety systems simultaneously, the fault-prone estimation method requires that all security controls be evaluated before implementation to ensure that none of the safety and emergency preparedness systems is negatively affected by security measures. This method, however, is not equipped with probabilistic analysis tools nor has instruments to evaluate the implementation of the defense-in-depth concept. The BBN-PSA method does not evaluate the defense-in-depth concept either. However, this method is still more complete because it can conduct probabilistic analysis for the risk assessment. Mirroring this situation, both the cybersecurity investigation and the DID-SIEM methods have the instruments to evaluate the defense-in-depth concept and the synergy between security and safety. However, they do not support probabilistic analysis of risks. The DID-SIEM method, in particular, conducts steps of risk assessment based on levels of priority to put safety concerns as main objectives and separate the safety-network section and the non-safety-network section.

The remaining risk assessment methods in Table IV support the consideration of all three cybersecurity aspects. For the implementation of the defense-in-depth concept, the security state estimation method evaluates whether safety-critical components are secured from any cyberattack. This technique also examines system vulnerabilities while it analyses the progress of the estimated cyberattacks, which meets the requirements in probabilistic risk assessment. The security state estimation method helps the security personnel keep the safety state during the operational period and conduct cause analysis while

establishing cyberattack responses at the right moment.

The InTo-CSI method implements the defense-in-depth strategy comprising the capability to protect, detect, respond, and recover during cyberattacks. The method utilizes event trees to evaluate existing protection functions and supports vulnerability analysis should successful attempts penetrate the protected systems. The combination of the defense-in-depth concept and the probability risk assessment is well implemented in the InTo-CSI method. The event trees generated in the method can show vulnerability areas and identify possible intrusion and attacks to the systems.

The cybersecurity vulnerability assessment method also supports the defense-in-depth concept and the probabilistic risk assessment. This method focuses on the lack of cybersecurity procedures that may lead to cybersecurity incidents originated from unintentional actions. Meanwhile, the possibility of cyberattacks involving insider threats also exists [2], [44]. A typical case of applying this method is the modernization of digital control, including I&C, which is unavoidable anymore in the nuclear industry. The assessment and analysis, in this case, require a dynamic process as the data involved should be reliable and updated [22]. They also need the engagement of all relevant stakeholders with better access to intelligent information.

The conformity to nuclear aspects and compliance with nuclear energy regulations are essential in finding suitable cybersecurity risk assessment methods. Such methods can deliver the best performance in securing the protected system and still fulfill the requirements specified in regulations.

Risk assessment methods for cybersecurity discussed in this paper have been implemented at nuclear facilities in various countries. The methods were designed to align with regulations released both by national and international regimes, such as IAEA and USNRC. The Korean Hydro & Nuclear Power, for instance, has implemented a cybersecurity risk assessment within the scope of nuclear facilities of nuclear power plants. The assessment has been conducted for Generation III and Generation III+ reactors, which are more digitized, like AP 1000. A similar assessment will be used in the United Arab Emirates for the APR 1400 units installed by South Korea.

In Indonesia, BAPETEN, as the Nuclear Energy Regulatory Agency, has been determined to include risk assessment in supervising nuclear energy utilization [67]. However, the tools used by the regulator to perform risk assessment are not always available. In 2012, BAPETEN released Regulation no. 6 to regulate the computer systems in nuclear facilities to comply with its requirements to support safety and security aspects [68]. The regulation emphasizes that immediate attention to any impending threats through early detection can revive the sustainability of proper cybersecurity in the nuclear industry.

V. Conclusion

Nuclear energy implementation should cover pillars of safety, security, and safeguard. They should be considered in selecting the most suitable risk assessment. The safeguard aspect has been declared earlier at the national level by ratifying the non-proliferation treaty to implement nuclear energy for peaceful purposes. Thus, we need to ensure that both safety and security aspects are well-maintained for operational during commissioning in nuclear facilities. Proper risk assessment can enhance cybersecurity in nuclear facilities. Risk assessment in cybersecurity for nuclear facilities is expected to keep the aspects of security and safety simultaneously.

Basic concepts that should be examined in nuclear security implementations include the synergy of safety-security and the defense-in-depth aspect. Moreover, security risk assessments in nuclear facilities should also implement commonly used PSA/PRA to integrate the attack scenario graphs. Cybersecurity vulnerability assessment is a viable method based on the selection process we have carried out in the previous section. The method puts together the defense-in-depth, the synergy of safety and security, and the application of PSA/PRA along with the scenario graph analysis. We believe that cybersecurity vulnerability assessment conforms to the stated requirements in regulations for nuclear energy implementation.

It is important to note that cybersecurity risk assessment is a must in nuclear energy utilization for peaceful use. Existing technologies of the I&C system, including PLCs, are vulnerable as they are attractive targets for the cyberattack threats. Appropriate risk assessment can enhance strong cybersecurity for providing preventive measures in avoiding potential cyberattacks.

This research will continue the in-depth study of the topics by focusing on cybersecurity and vulnerability aspects and include them in the PSA/PRA analysis generally implemented in nuclear facilities.

Authorship Contribution Statement

Lilis Susanti Setianingsih: methodology, writing, and original draft. Reza Pulungan: conceptualization, validation, supervision, writing-review, and editing. Agfianto Eko Putra: validation, supervision, review, and editing. Moh. Edi Wibowo: validation, supervision, review, and editing. Syarip: validation, supervision, review, and editing.

Acknowledgment

The authors would like to thank BAPETEN (Indonesian Nuclear Energy Regulatory Agency), and Universitas Gadjah Mada. This research is partially funded by Universitas Gadjah Mada's Rekognisi Tugas Akhir program in 2020.

References

[1] "Computer security at nuclear facilities: Technical guidance reference manual," *IAEA Nuclear Security Series No. 17*, pp. 1–69, 2011.

[2] S. Ali, "Cybersecurity management for distributed control system: Systematic approach," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2021.

[3] M. Alali, A. Almogren, M. M. Hassan, I. A. Rassan, and M. Z. A. Bhuiyan, "Improving risk assessment model of cyber security using fuzzy logic inference system," *Computers & Security*, vol. 74, pp. 323–339, 2018.

[4] J. Shin, H. Son, and G. Heo, "Cyber security risk evaluation of a nuclear I&C using BN and ET," *Nuclear Engineering and Technology*, vol. 49, no. 3, pp. 517–524, 2017.

[5] N. Ben-Asher and C. Gonzalez, "Effects of cyber security knowledge on attack detection," *Computers in Human Behavior*, vol. 48, pp. 51–61, 2015.

[6] J. W. Park and S. J. Lee, "A quantitative assessment framework for cyber-attack scenarios on nuclear power plants using relative difficulty and consequence," *Annals of Nuclear Energy*, vol. 142, p. 107432, 2020.

[7] D. K. Jana and R. Ghosh, "Novel interval type-2 fuzzy logic controller for improving risk assessment model of cyber security," *Journal of Information Security and Applications*, vol. 40, pp. 173–182, 2018.

[8] C. O'Halloran, T. G. Robinson, and N. Brock, "Verifying cyber attack properties," *Science of Computer Programming*, vol. 148, pp. 3–25, 2017.

[9] X. Fan, K. Fan, Y. Wang, and R. Zhou, "Overview of cyber-security of industrial control system," in *2015 International Conference on Cyber Security of Smart Cities, Industrial Control System and Communications (SSIC)*, 2015, pp. 1–7.

[10] D. Young, J. Lopez, M. Rice, B. Ramsey, and R. McTasney, "A framework for incorporating insurance in critical infrastructure cyber risk strategies," *International Journal of Critical Infrastructure Protection*, vol. 14, pp. 43–57, 2016.

[11] C. Alcaraz and S. Zeadally, "Critical infrastructure protection: Requirements and challenges for the 21st century," *International Journal of Critical Infrastructure Protection*, vol. 8, pp. 53–66, 2015.

[12] "Cyber security programs for nuclear facilities," *Regulatory Guide 5.71, Office of Nuclear Regulatory Research, U.S. Nuclear Regulatory Commission*, pp. 1–40, 2010.

[13] R. Leszczyna, "Review of cybersecurity assessment methods: Applicability perspective," *Computers & Security*, vol. 108, p. 102376, 2021.

[14] M. Touhiduzzaman, S. N. G. Gourisetti, C. Eppinger, and A. Somani, "A review of cybersecurity risk and consequences for critical infrastructure," in *2019 Resilience Week (RWS)*, vol. 1, 2019, pp. 7–13.

[15] M. R. Asghar, Q. Hu, and S. Zeadally, "Cybersecurity in industrial control systems: Issues, technologies, and challenges," *Computer Networks*, vol. 165, p. 106946, 2019.

[16] S. Kim, S.-m. Kim, K.-h. Nam, S. Kim, and K.-h. Kwon, "Security information and event management model based on defense-in-depth strategy for vital digital assets in nuclear facilities," in *Advances in Computer Science and Ubiquitous Computing*, J. J. Park, S. J. Fong, Y. Pan, and Y. Sung, Eds. Springer Singapore, 2021, pp. 331–339.

[17] Y. Ashibani and Q. H. Mahmoud, "Cyber physical systems security: Analysis, challenges and solutions," *Computers & Security*, vol. 68, pp. 81–97, 2017.

[18] S. Moskal, S. J. Yang, and M. E. Kuhl, "Cyber threat assessment via attack scenario simulation using an integrated adversary and network modeling approach," *The Journal of Defense Modeling and Simulation*, vol. 15, no. 1, pp. 13–29, 2018.

[19] H. E. Kim, H. S. Son, J. Kim, and H. G. Kang, "Systematic development of scenarios caused by cyber-attack-induced human errors in nuclear power plants," *Reliability Engineering & System Safety*, vol. 167, pp. 290–301, 2017.

[20] G. Daria and A. Massel, "Intelligent system for risk identification of cybersecurity violations in energy facility," in *3rd Russian-Pacific Conference on Computer Technology and Applications*, 2018, pp. 1–5.

[21] A. Reeves, P. Delfabbro, and D. Calic, "Encouraging employee engagement with cybersecurity: How to tackle cyber fatigue," *SAGE Open*, vol. 11, no. 1, pp. 1–18, 2021.

[22] "Risk informed approach for nuclear security measures for nuclear and other radioactive material out of regulatory control: Implementing guide," *IAEA Nuclear Security Series No. 24-G*, pp. 1–69, 2015.

[23] X. Yao, F. Farha, R. Li, I. Psychoula, L. Chen, and H. Ning, "Security and privacy issues of physical objects in the IoT: Challenges and opportunities," *Digital Communications and Networks*, 2020.

[24] E. H. Riyadi, T. K. Priyambodo, and A. E. Putra, "The dynamic symmetric four-key-generators system for securing data transmission in the industrial control system," *International Journal of Intelligent Engineering and Systems*, vol. 14, pp. 376–386, 2021.

[25] S. Eggers, "A novel approach for analyzing the nuclear supply chain cyber-attack surface," *Nuclear Engineering and Technology*, vol. 53, no. 3, pp. 879–887, 2021.

[26] "IEEE standard criteria for digital computers in safety systems of nuclear power generating stations," *IEEE Std 7-4.3.2-2003 (Revision of IEEE Std 7-4.3.2-1993)*, pp. 1–65, 2003.

[27] "Information technology - Security techniques - Evaluation criteria for IT security - Part 1: Introduction and general model," *ISO/IEC 15408-1:2009*, pp. 1–64, 2009.

[28] A. Khalid, P. Kirisci, Z. H. Khan, Z. Ghrairi, K.-D. Thoben, and J. Pannek, "Security framework for industrial collaborative robotic cyber-physical systems," *Computers in Industry*, vol. 97, pp. 132–145, 2018.

[29] S. Abraham and S. Nair, "Comparative analysis and patch optimization using the cyber security analytics framework," *The Journal of Defense Modeling and Simulation*, vol. 15, no. 2, pp. 161–180, 2018.

[30] C. Lee, S. M. Han, and P. H. Seong, "Development of a quantitative method for identifying fault-prone cyber security controls in NPP digital I&C systems," *Annals of Nuclear Energy*, vol. 142, p. 107398, 2020.

[31] J. Shin, H. Son, R. Khalil ur, and G. Heo, "Development of a cyber security risk model using Bayesian networks," *Reliability Engineering & System Safety*, vol. 134, pp. 208–217, 2015.

[32] H. El-Sofany, "A new cybersecurity approach for protecting cloud services against DDoS attacks," *International Journal of Intelligent Engineering and Systems*, vol. 13, pp. 205–215, 2020.

[33] W. Ahn, M. Chung, B.-G. Min, and J. Seo, "Development of cyber-attack scenarios for nuclear power plants using scenario graphs," *International Journal of Distributed Sensor Networks*, vol. 11, no. 9, p. 836258, 2015.

[34] Z. Wang, H. Zhu, and L. Sun, "Social engineering in cybersecurity: Effect mechanisms, human vulnerabilities and attack methods," *IEEE Access*, vol. 9, pp. 11 895–11 910, 2021.

[35] C. Lee, Y. Ho Chae, and P. Hyun Seong, "Development of a method for estimating security state: Supporting integrated response to cyber-attacks in NPPs," *Annals of Nuclear Energy*, vol. 158, p. 108287, 2021.

[36] Z. Ji, S.-H. Yang, Y. Cao, Y. Wang, C. Zhou, L. Yue, and Y. Zhang, "Harmonizing safety and security risk analysis and prevention in cyber-physical systems," *Process Safety and Environmental Protection*, vol. 148, pp. 1279–1291, 2021.

[37] M. S. El-Genk, R. Altamimi, and T. M. Schriener, "Pressurizer dynamic model and emulated programmable logic controllers for nuclear power plants cybersecurity investigations," *Annals of Nuclear Energy*, vol. 154, p. 108121, 2021.

[38] N. Yousefnezhad, A. Malhi, and K. Främling, "Security in product lifecycle of IoT devices: A survey," *Journal of Network and Computer Applications*, vol. 171, p. 102779, 2020.

[39] R. Syed, "Cybersecurity vulnerability management: A conceptual ontology and cyber intelligence alert system," *Information & Management*, vol. 57, no. 6, p. 103334, 2020.

[40] J. Peterson, M. Haney, and R. Borrelli, "An overview of methodologies for cybersecurity vulnerability assessments conducted in nuclear power plants," *Nuclear Engineering and Design*, vol. 346, pp. 75–84, 2019.

[41] E. J. Oughton, D. Ralph, R. Pant, E. Leverett, J. Copic, S. Thacker, R. Dada, S. Ruffle, M. Tuveson, and J. W. Hall, "Stochastic counterfactual risk analysis for the vulnerability assessment of cyber-physical attacks on electricity distribution infrastructure networks," *Risk Analysis*, vol. 39, no. 9, pp. 2012–2031, 2019.

[42] L. Vessels, K. Heffner, and D. Johnson, "Cybersecurity risk assessment for space systems," in *2019 IEEE Space Computing Conference (SCC)*, 2019, pp. 11–19.

[43] J. Draeger and S. Hahndel, "Simulation-based unified risk assessment for safety and security," *arXiv*, vol. abs/1709.00567v2, pp. 1–24, 2019.

[44] N. A. Hashim, Z. Z. Abidin, A. Puvanasvaran, N. A. Zakaria, and R. Ahmad, "Risk assessment method for insider threats in cyber security: A review," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 11, pp. 126–130, 2018.

[45] H. S. Lallie, K. Debattista, and J. Bal, "Evaluating practitioner cybersecurity attack graph configuration preferences," *Computers & Security*, vol. 79, pp. 117–131, 2018.

[46] R. Leszczyna, "Cybersecurity and privacy in standards for smart grids - a comprehensive survey," *Computer Standards & Interfaces*, vol. 56, pp. 62–73, 2018.

[47] O. Ivanchenko, V. Kharchenko, B. Moroz, L. Kabak, and S. Konovalenko, "Risk assessment of critical energy infrastructure considering physical and cyber assets: Methodology and models," in *2018 IEEE 4th International Symposium on Wireless Systems within the International*

Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS), 2018, pp. 225–228.

[48] M. G. Porcedda, "Patching the patchwork: appraising the EU regulatory framework on cyber security breaches," *Computer Law & Security Review*, vol. 34, no. 5, pp. 1077–1098, 2018.

[49] E. Zio, "The future of risk assessment," *Reliability Engineering & System Safety*, vol. 177, pp. 176–190, 2018.

[50] C. Lee, H. B. Yim, and P. H. Seong, "Development of a quantitative method for evaluating the efficacy of cyber security controls in NPPs based on intrusion tolerant concept," *Annals of Nuclear Energy*, vol. 112, pp. 646–654, 2018.

[51] M. Yampolskiy, P. Horváth, X. D. Koutsoukos, Y. Xue, and J. Sztipanovits, "A language for describing attacks on cyber-physical systems," *International Journal of Critical Infrastructure Protection*, vol. 8, pp. 40–52, 2015.

[52] M. Ficco, M. Choraś, and R. Kozik, "Simulation platform for cybersecurity and vulnerability analysis of critical infrastructures," *Journal of Computational Science*, vol. 22, pp. 179–186, 2017.

[53] R. Busquim e Silva, J. Piqueira, J. Cruz, and R. Marques, "Cybersecurity assessment framework for digital interface between safety and security at nuclear power plants," *International Journal of Critical Infrastructure Protection*, vol. 34, p. 100453, 2021.

[54] V. Casson Moreno, G. Reniers, E. Salzano, and V. Cozzani, "Analysis of physical and cyber security-related events in the chemical and process industry," *Process Safety and Environmental Protection*, vol. 116, pp. 621–631, 2018.

[55] J. Neeli and S. Patil, "Insight to security paradigm, research trend & statistics in internet of things (IoT)," *1st International Conference on Advances in Information, Computing and Trends in Data Engineering (AICDE)*, vol. 2, no. 1, pp. 84–90, 2021.

[56] G. Strupczewski, "Defining cyber risk," *Safety Science*, vol. 135, p. 105143, 2021.

[57] M. Alohali, N. Clarke, and S. Furnell, "The design and evaluation of a user-centric information security risk assessment and response framework," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 10, pp. 148–163, 2018.

[58] J. Hu, S. Guo, X. Kuang, F. Meng, D. Hu, and Z. Shi, "I-HMM-based multidimensional network security risk assessment," *IEEE Access*, vol. 8, pp. 1431–1442, 2020.

[59] A. Nakai and K. Suzuki, "Risk assessment system for verifying the safeguards based on the HAZOP analysis," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 10, pp. 48–53, 2014.

[60] F. Arnold, H. Hermanns, R. Pulungan, and M. Stoelinga, "Time-dependent analysis of attacks," in *Principles of Security and Trust*, M. Abadi and S. Kremer, Eds. Springer Berlin Heidelberg, 2014, pp. 285–305.

[61] Y. Zhao, L. Huang, C. Smidts, and Q. Zhu, "Finite-horizon semi-Markov game for time-sensitive attack response and probabilistic risk assessment in nuclear power plants," *Reliability Engineering & System Safety*, vol. 201, p. 106878, 2020.

[62] E. Bou-Harb, M. Debbabi, and C. Assi, "Cyber scanning: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 16, no. 3, pp. 1496–1519, 2014.

[63] P. H. Nguyen, S. Ali, and T. Yue, "Model-based security engineering for cyber-physical systems: A systematic mapping study," *Information and Software Technology*, vol. 83, pp. 116–135, 2017.

[64] K. Geers, "The challenge of cyber attack deterrence," *Computer Law & Security Review*, vol. 26, no. 3, pp. 298–303, 2010.

[65] W. Tounsi and H. Rais, "A survey on technical threat intelligence in the age of sophisticated cyber attacks," *Computers & Security*, vol. 72, pp. 212–233, 2018.

[66] M. A. Hawila and S. S. Chirayath, "Combined nuclear safety-security risk analysis methodology development and demonstration through a case study," *Progress in Nuclear Energy*, vol. 105, pp. 153–159, 2018.

[67] "Undang-Undang Republik Indonesia no. 10/1997 (Indonesia Nuclear Energy Act no. 10/1997)," 1997, (In Bahasa Indonesia).

[68] "Peraturan Kepala BAPETEN no. 6/2012 (BAPETEN Chairman Regulation No. 6/2012)," 2012, (In Bahasa Indonesia).

Comparative Analysis of Spark and Ignite for Big Spatial Data Processing

Samah Abuayeid, Louai Alarabi
Department of Computer Science
Umm Al-Qura University
Mecca, Saudi Arabia

***Abstract*—Recently, spatial data became one of the most interesting fields related to big data studies, in which the spatial data have been generated and consumed from different resources. However, the increasing numbers of location-based services and applications such as Google Maps, vehicle navigation, recommendation systems are the main foundation of the idea of spatial data. On the other hand, several researchers started to discover and compared spatial frameworks to understand the requirements for spatial database processing, manipulating, and analysis systems. Apache Spark, Apache Ignite, and Hadoop are the most widely known frameworks for large data processing. However, Apache Spark, Apache Ignite have integrated different spatial data operations and analysis queries, but each system has its advantages and disadvantages when dealing with spatial data. Dealing with a new framework or system that needs to integrate new functionality sometimes becomes a risky decision if we did not examine it well The main aim of this research is to conduct a comprehensive evaluation of big spatial data computing on two well-known data management systems Apache Ignite and Apache Spark. The comparative has been done on four different domains, experimental environment setup, supported features, supported functions and queries, and performance and execution time. The results show that GeoSpark has recorded more flexibility to use than SpatialIgnite. We thoroughly investigated and discovered that multiple factors affect the performance of both frameworks, such as CPU, Main memory, data set size the complexity of data type, and programming environment. spark is more advanced and equipped with several functionalities that made it well suitable with spatial data queries and indexing. such as kNN queries; in which these functionalities are not supported in SpatialIgnite.**

***Keywords*—Big spatial data; GeoSpark; SpatialIgnite; Apache Ignite; Apache Spark**

I. Introduction

Big data processing has always been a critical research area in both academia and industry. Several big tech organizations invested billions of dollars to build Big data Eco-system, For example, Facebook [1], LinkedIn [2], Microsoft [3], ESRI [4] to name a few. Meanwhile, several non-tech companies have integrated one or more available platforms to scale out and perform their big data analytic tasks. One important domain of this market is building Eco-systems for spatial data due to the plethora of applications and services that create them. For instance, earth observation has continuously provided a significant volume of geospatial data over the last few years, resource tracking [5], environmental protection, and disaster predictions [6]. Thus, big data spatial computing has become extremely valuable with the widespread use of these services and applications.

The growth market and data size of location-based services contributed to the advancement and complexity of computing spatial data [7]. Several efforts from both industry [8] and academia [9], [10], [11], [12] introduced a specialized spatial data processing engine to process this complex data. The developments of specialized geospatial processing engines are driven by adopting new technologies for processing big data. An essential aspect of any spatial system is how the system deals with the Big V's such as Volume, Variety, veracity, and Velocity. For instance, several research studies extended these frameworks to specific domain applications, such as Transportation [13], [14], [15], [16]

Although many systems have been established to leverage the processing of spatial data. Yet, there is no single source of a comprehensive benchmark that distinguishes between these systems. The lack of possessing this kind of benchmark is due to the complexity and efforts associated with building them up. On the other hand, the variety of spatial data types in different domains (e.g., location, routing, navigation) makes it even harder to benchmark. Thus, we extensively spend a decent effort designing and assisting the performance of well-known big in-memory spatial platforms built on Apache Spark and Apache Ignite.

Apache has founded and managed several big data processing projects [17], such as Hadoop [18], Spark [19] and Ignite [20] among others. This paper, investigated Spark and Ignite which are commonly known as distributed in-memory big data processing platforms. Researchers investigated these two platforms in processing big spatial data, by introducing and building spatial properties, operations, and queries in these two platforms. For instance, Apache Spark [21] was introduced as the GeoSpark system, in which users can interacting with the system by either: Spatial SQL API or a Scala/Java RDD API. spark provides for the users an operational programming language for writing a custom spatial analytic application. On the other hand, Apache Ignite [22] is another open-source distributed database system that includes an in-memory data grid (IMDG) that was established to store and compute big data across a cluster of nodes. However, to integrate the spatial data processing and quires in Apache Ignite, users must add the dependency of an ignite-geospatial library that is included in the JTS Topology Suite.

Dealing with a new framework or system that needs to integrate new functionality sometimes becomes a risky decision if we did not examine it, and In the literature up to now, there are few comparative studies between Apache Spark and Apache Ignite systems in spatial data management domain.

This paper, motivated by a prior study comprehensively investigated and evaluated a native version of Apache Hadoop and Spark, respectively [23]. In this research the study took the leverage of the advancement of the field of spatial data computing and comprehensively evaluate and compare two distributed in-memory computing frameworks and well-known data management frameworks of Apache Ignite and Apache Spark. In particular, the study evaluated the spatial extension of Apache Spark that is well-known as GeoSpark with it's competitor SpatialIgnite on Apache ignite.

The rest of this paper has organized as follows. Section 3 research background, which has divided into two subsections. Section 4 consists of materials and methods, which has divided into two subsections. Section 5 included the results and discussion. Concluding remarks are contained in the last section.

II. Related Work

New research by Cristiana-Stefania Stan [24] has compared Apache Spark and Apache Ignite from different aspects such as implementation, features, architecture, and performance metrics using word count and k-means clustering to specify the best framework for Big data processing. The experiment was conducted on two nodes to offer a small cluster, which made the testing step accurate because it can measure the communication between the nodes. On the other hand, they also test the computation duration and multiple systems resources utilization such as CPU, memory, and network, as a result the author diseased that Apache Spark has addressed a higher performance in processing Big data than Apache Ignite.

Md Mahbub Alam [12] has studied the performance of Big spatial data on Apache Spark, Spatial-Ignite benchmark developed based on Apache Ignite and HADOOP. The author has realized that not all the systems have supported all Spatial Data features. The work has inspired form [33], which proposed the d Jackpine spatial database benchmark to support as many as possible spatial databases with minimal effort. However, the experiment in [12] has been conducted using a real-world spatial dataset from TIGER on a cluster of 8 machines with the two frameworks (Apache Spark and Hadoop) and the proposed benchmark (Spatial-Ignite). The result has addressed three categories of the operations for examination: spatial join, spatial analysis queries, and range query. At the end of the experiment, Spatial-Ignite performs better than Apache Spark and Hadoop on the spatial queries.

Recently, social media platforms provided a huge amount of spatial and temporal datasets, and several spatial data processing frameworks were developed to dealing with this type of massive data. To illustrate this, the study by Zhibo Sun [25], investigated Apache Spark to perform some geospatial analyses operations, such as K-Nearest Neighbors (KNN), the distribution of the median points, and the geographic mean and median points. The study applied three different sizes of Twitter data: 180GB, 120GB, and 60GB. Additionally, they compared the execution results and performance of Apache Spark with Hadoop, in which Spark addressed a high performance than Hadoop when applied to the Twitter dataset. Furthermore, the experiment was conducted on the Amazon EC2 cluster, that consists off 11 m3.xlarge nodes. However, each instance in the m3.xlarge has "Intel Xeon E5-2670 v2, 4 vCPU, 15 GB memory, and 100 GB magnetic storage", where the study was dependent on Apache Hadoop 2.4, Spark 1.1.1. By comparing Hadoop and Spark in computing the distribution of the median points, and geographic mean, Spark addressed better performance than Hadoop, in which writing and reading data in and out of the disk take more time with the Hadoop solution. On the other hand, Spark also records a high performance faster than Hadoop, equals 2.3x, 1.6x, and 1.8x on the three diffident sizes of Twitter datasets. As a result Spark solution was outperformed the Hadoop solution, but the author was found some disadvantages and limitations in the Spark-based solution:

- Spark uses some coarse-grained mode, that enhance their performance, but on the other hand it wastes lots of resources, which cause jobs delay.
- Spark uses a lazy operation mechanism, that requires a run of some actions before debugging.
- Spark can cost on physical nodes, in which it needs large memory in compared to Hadoop.

Moreover, Randall T. Whitman [26] was studied the spatial join operations on Apache Spark, and proposed a framework that uses the spatiotemporal join algorithms using two datasets. The framework is expected to runs in commercial off-the-shelf (COTS) applications. The study has employed two approaches to perform the spatial join operation. The first approach was the broadcast spatial join, that built to join a big spatial dataset with another small dataset, where, the second approach was the Bin spatial join, which is a technique for joining two large datasets. However, the broadcast join is similar to the map-side join in MapReduce programming, and it is suitable when one of two datasets can fit into memory on the Spark executors, so the study used this join operation to determine which one of the dataset is smaller and can fit be into the Spark memory. On the other hand, Bin Join is similar to the reduce-side join in MapReduce programming, and it is usually convenient when there is a need for a partitioned approach because the datasets are too large to fit into memory. The study was started by defining the size of the two datasets, and they were distributed into Spark executors, then the datasets partitions in each executor were joining using the broadcast join. The next step was using the bin join, and the same binning operation was applied on the two datasets, to ensure that the features on each side of the spatial grid are the same. In the end, the de-duplication step was performed to remove the duplicate matches when the features become in multiple bins. However, to estimate the proposed framework performance the study was conducted on New York City Taxi and Limousine Commission's taxi dataset, and they concluded, that join algorithms depend on the characteristics of the two datasets. On the other hand, they observed that the broadcast join is faster when one of the datasets is in a modest size, and the operation performance decreased when the two datasets are large. Moreover, the fastest binning technique is the regular grid mesh.

On the other hand, Jia Yu [27] was presented the GeoSpark framework, which is an in-memory cluster computing framework for manipulating large spatial data. The proposed framework consists of three main layers: Apache Spark layer,

TABLE I. LITERATURE REVIEW SUMMARY

Paper	Author	Year	Framework	Result
[24]	Cristiana-Stefania	2019	Spark and Ignite	Spark addressed higher performance
[12]	Md Mahbub	2018	Spark, Spatial-Ignite, and Hadoop	Spatial-Ignite addressed higher performance than Spark and Hadoop
[25]	Zhibo Sun	2016	Spark and Hadoop	Spark outperformed Hadoop
[26]	Randall T. Whitman	2017	Spark spacial join algorithms	broadcast join address best performance
[27]	Jia Yu	2019	GeoSpark and Hadoop	GeoSpark outperform Hadoop
[28]	Jayati Gandhi	2020	GeoSpark and Spatial Hadoop	GeoSpark more efficient than Spatial Hadoop
[29]	Mingjie Tang	2016	Location-Spark performance	addressed good performance in spatial queries processing
[30]	Francisco Garcıa	2017	spatialHadoop and LocationSpark	LocationSpark outperformed spatialHadoop
[31]	Zhou Huang	2017	GeoSpark SQL, PostGIS, and Hadoop ESRI	GeoSpark SQL was more user friendly
[32]	Panagiotis Nikitopoulos	2018	DiStRDF under SPARK	process RDF spatial temporal queries in minimum time
[10]	Louai Alarabi	2018	ST-Hadoop	ST-Hadoop outperformed Hadoop and SpaitalHadoop

spatial RDD layer, and the spatial query layer, the framework allows the user to built a spatial index. However, each layer is responsible for specific functionality in the framework: the Apache Spark layer responsible for providing the basic Spark functionality as loading and sorting data in the disk, where the second layer the spatial RDD layer can support the geometrical, and spatial objects. On the other hand, the spatial query layer's main functionality is to execute the main spatial query processing algorithms such as KNN, Join, and spatial range. Moreover, the study was compared the performance of the proposed framework with Hadoop spatial computation. The study was conducted on three different datasets from TIGER files: Zcta510 1.5 GB dataset, Areawater 6.5 GB dataset, and Edges 62 GB dataset. The experiment was started by configuration setup: one CPU for each worker, two Memory per worker each one consist of 61 GB, 50 GB registered memory in Spark and Hadoop and three storage per worker. After that they studied the time performance of some large spatial data processing systems: GeoSpark-NoIndex, Quad-Tree, RTree, GeoSpark without spatial index, and with spatial Quad-Tree or R-Tree index, then the performance was compared with spatial Hadoop. The experiment concludes with that GeoSpark record a better performance than Spatial Hadoop, and as a future work GeoSpark can be used in multiple fields and different users such as space scientists, geographers, politicians, commercial institutions to support the spacial data analyzing process.

On the other hand, a study by Jayati Gandhi [28] was discussed the importance of geospatial data processing and analysis, which lead to the discovery of various spatial data frameworks. The study also presented a comparative study between GeoSpark and Spatial Hadoop, in which they are two of the most known open source geospatial big data analytic frameworks. The study was compared between the two frameworks according to multiple characteristics such as the architecture View, spatial data processing approach, and real-time performance of each framework. As a result, the author was found, that both frameworks: GeoSpark and Spatial Hadoop are suitable and flexible to dealing with geospatial data, but the experiment was recorded, that GeoSpark is more efficient and fast than Spatial Hadoop. In future work, the author was mentioned that they tend to use the study result to enhance the way of disaster management, and geospatial health infrastructure.

Another study in the research area was presented by Mingjie Tang [29] discussed the efficiency of MapReduce-based systems, in which it allows performing spatial queries using predefined spatial operations without the need to worry about fault tolerance and computation distribution problems. However, MapReduce-based systems had addressed some disadvantages such as, the lack of leverage distributed memory, and they do not allow immediate data reuse, in which data reusing is very common in spatial data preprocessing. Consequently, the study has introduced a solution that can fix these problems by built a spatial data processing system on Apache Spark and known as the Location-Spark. The system consists of 6 different layers: memory management, spatial index, query executor, query scheduler, spatial operator, and spatial analysis. Moreover, Location-Spark allows users to use a set of spatial query operators, such as range search, spatial-join, KNN, spatial-textual operation, and kNN-join. Moreover, the proposed system was offered multiple in-memory data spatial indexes, and guaranteed fixed spatial indexes decreased the overhead of fault tolerance. As a result, Location-Spark was outperformed in speed and performance of other spatial systems.

A study by Francisco Garcıa [30] was compared between spatialHadoop and LocationSpark by estimating the performance of each spatial data system in the way that they

performed two spatial join queries: K Closest Pair Query (KCPQ) and the ϵ Distance Join Query. The study was conducted on Linux operating system using three 2d point spatial datasets from Open-Street-Map data: building dataset includes 115M records, lakes dataset contains 8.4M points, and parks consist of 10M records. The experiment was implemented the KCPQ and ϵ Distance, Join Query, on the three datasets on spatialHadoop and LocationSpark. The study has adopted the idea of plane sweep techniques. On the other hand, The performance was evaluated depending on n the total execution time. As a result, the author found that LocationSpark was addressed the best performance comparing to spatialHadoop, according to the in-memory processing and the query plan scheduler offered by Spark. Additionally, the advantage of SpatialHadoop appeared in it is had more mature and robust DSDMS than Spark. However, as a future work, the paper tended to apply a new study on Spark-based DSDMS, such as Simba, and apply some spatial partitioning techniques on LocationSpark.

Moreover, Zhou Huang [31] was discussed the requirements of spatial data and spatial data query processing in the area of big data. Additionally, the study was introduced the GeoSpark SQL, which is a framework that allows the operations of spatial data query processing on Apache Spark. Furthermore, the paper was addressed the main issues related to introducing spatial big data query processing: storage management methods, spatial operations implementation approach, and spatial query optimization approaches. The author was chosen Apache spark in this study rather than Hadoop, in which that spark had the ability of efficient memory management, that enhance the computations of MapReduce. The experiment was covered multiple spatial queries such as KNN query, point query, window query, range query, directional query, topological query, and multi-table spatial join query. However, for experiment evaluation 10 different test cases were created, and the average time was taken as a performance evaluation metric, for comparing the performance of GeoSpark SQL with PostGIS, and Hadoop ESRI Spatial Framework. As a result, GeoSpark SQL was user-friendly, and it only need to add Java dependencies for spatial data query processing in Spark, and then start spark terminal. As future work, the author tended to improve the environment of GeoSpark SQL to integrate complex spatial data indexes.

Moreover, Panagiotis Nikitopoulos in his research[32] proposed the DiStRDF system using the resource description framework, which is a framework for web data modeling and interchanging. Additionally, the study has used Spark as the processing framework. Moreover, the experiment introduced a challenging solution for interlinking and interchanging the spatial-temporal data obtained from mobile devices. Moreover, the proposed system has addressed a good performance in the process of the resource description framework spatial-temporal queries with minimum execution time.

On the other hand, Alarabi in his research [10] proposed an open-source MapReduce framework, which supports the spatial-temporal data known as ST-Hadoop. However, the proposed framework contained one primary node, which helps in breaking map-reduce jobs into multiple tasks. Additionally, three different users are allowed by the system: casual users, developers, and administrators. According to the author, the framework is divided into four layers: language layer, indexing layer, MapReduce layer, and operations layer. The experiment was conducted on a dataset with over 1 Billion spatial-temporal data of size 10 TB. As a result, ST-Hadoop outperformed Hadoop and SpaitalHadoop in processing spatial-temporal data. However, Table I shows a summary for this section.

III. Materials and Method

This section introduce the research methodology in more detail with a clear explanation for the dataset, tools, and experiment environment that have been used to conduct the study. However, the experiment was conducted on one machine a HP Pavilion laptop with 8GB RAM, 1 TB HDD storage with windows 10 as an operating system. The research methodology as it shown in Fig. 1 started by defined the comparative study four main domains, which are experimental environment setup, supported features, supported functions and queries, and performance and execution time. Additionally, in the last step of comparison we have selected multiple datasets and calculated the execution time. However, this research is inspired by Md Mahbub Alam [12], in which it was depend on some assumptions and results obtained by Md Mahbub Alam.

A. Data

In this section we have explained the dataset used for study the performance and execution time for GeoSpark on Apache Spark, to understand the assumption that has been proposed by Md Mahbub Alam [12], in which that GeoSpark recorded the worst performance and execution time in comparing to SpatialIgnite and SaptialHadoop. However, we have utilizes real-world spatial datasets which were obtained from the Spatial Hadoop framework website[1]. However, the Spatial Hadoop website provided multiple TIGER 2015 spatial datasets and Open Street Map datasets. TIGER datasets were extracted from the US Census Bureau TIGER files. On the other hand, Open Street Map is a collaborative project to create a free editable map of the world and offered an API for data extraction. Two different TIGER datasets were chosen in this research: AREALM dataset 140MB, and PRIMARY ROADS dataset 52 MB. However, we have used two points CSV datasets 487 KB and 325 KB, which was been obtained from SpiderWeb[2]. The SpiderWeb is a website that generates spatial data, in which users can easily choose the number of layers, Cardinality, file extensions, and other proprieties. In the end, we have applied four different types of datasets on Apache Spark, Table II shows some information about our experiment datasets. However, all datasets have one attribute as the coordinate attribute.

B. The Proposed Study among Apache Spark and Apache Ignite

The study in this research was divided into four parts as follows:

1) A comparison among experimental environment setup: explained the main requirements and steps

[1]http://spatialhadoop.cs.umn.edu/datasets.html
[2]https://spider.cs.ucr.edu/

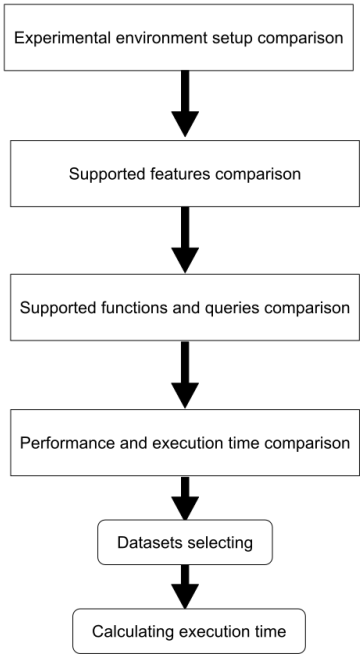


Fig. 1. The Research Methodology.

TABLE II. DATASETS DESCRIPTION

Dataset	Geometry	Cardinality	Size	Extension	Description
AREALM	Polygon	129K	140MB	CSV	Area Landmark
PRIMARYROADS	Linestrings	12,396	52 MB	CSV	Primary Roads
data1	points	12,000	487 KB	CSV	points data from SpiderWeb
data2	points	8000	325 KB	CSV	points data from SpiderWeb

need it for each framework to activate the big spatial data computing.

2) A Comparison among main supported features in each framework: explained the main supported input format, geometry type, query language, and more.
3) A comparison among the existing supported functions and queries.
4) Discussing the performance and execution time in each framework, and we have depended on the assumption proposed by Md Mahbub Alam [12], in which SpatialIgnite have the best performance, for that we have focused on calculating the performance time for Geospark using multiple datasets to understand the performance rate of GeoSpark.

1) Experimental Environment Setup: To activate the spatial computing option. The experiment was conducted using two Apache software: Apache Ignite 2.8.0[3], Apache Spark 3.0.1[4]. However, Apache NetBeans IDE 11.3 as an IDE environment was installed for working with Apache Ignite, in which the only way to support the spatial data operations and quires is by adding the Spatial Ignite dependencies on Maven in any Java IDE. Additionally, Spatial Ignite is an API of spatial predicates in the JTS Topology Suite[5]. On the other hand, Apache Spark was installed to add the GeoSpark dependency for working with GeoSpark on Apache Spark. Additionally, the study was conducted on Geospark rather than Spacial Spark because GeoSpark offered more file extensions and allow spatial indexing. However, the main programming environment that was used was Scala on the Spark shell, and the main experimental environment setup for the framework shows in Fig. 2.

2) Supported Features: This part presented a comparison between GeoSpark and SpatialIgnite among their main features. According to Table III, GeoSpark and SpatialIgnite supported the same geometry type: point, line, and polygon. Additionally, spatial analysis is supported only by SpatialIgnite. However, GeoSpark has the advantage of supporting multiple file formats: CSV, TSV, WKT, and GeoJSON, where SpatialIgnite only supports WKT file format, which decreases the flexibility of SpatialIgnite. On the other hand, Table III, shows that SpatialIgnite supports the distributed SQL query language, where GeoSpark supports SQL 2017, which may make GeoSpark easier than SpatialIgnite in the way of handling SQL queries. Another advantage for GeoSpark is that it supports two types of indexing, which are the R-tree and Quadtree, where SpatialIgnite only supports the R-tree

[3]https://ignite.apache.org/
[4]https://spark.apache.org/
[5]http://www.tsusiatsoftware.net/jts/main.html

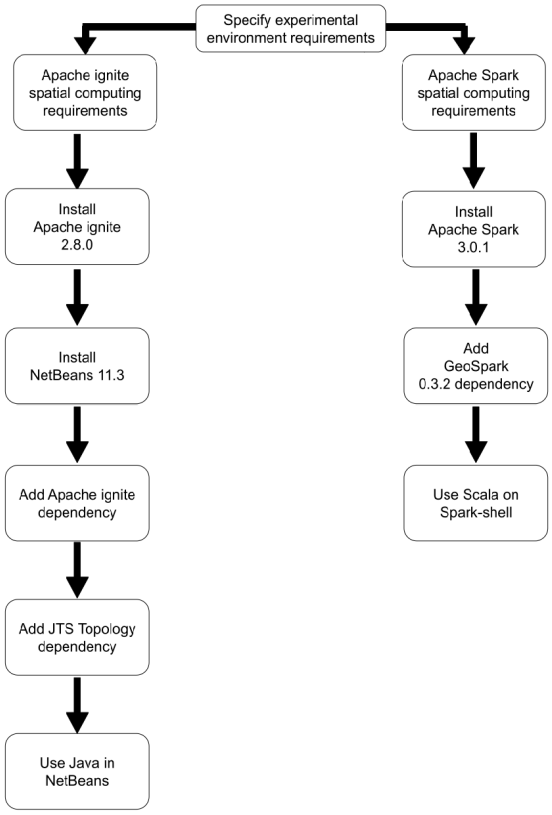


Fig. 2. The Experimental Environment Setup.

TABLE III. SUMMARY FOR THE SUPPORTED FEATURES IN GEOSPARK AND SPATIAL IGNITE

Features	GeoSpark	Spatial Ignite
geometry type	point, line, polygon	point, line, polygon
input format	CSV,TSV,WKT,GeoJSON	WKT
query language	SQL(2017)	Distributed SQL
indexing	R-tree, Quad-tree	R-tree
spatial analytic	No	Yes

indexing.

TABLE IV. SUMMARY OF SOME SUPPORTED SPATIAL FUNCTIONS AND QUERIES IN GEOSPARK AND SPATIAL IGNITE

Predicates	GeoSpark	Spatial Ignite
Equals	Yes	Yes
Intersects	Yes	Yes
Crosses	No	Yes
Envelope	Yes	Yes
ConvexHull	No	Yes
Within	No	Yes
Touches	Yes	Yes
KNN-query	Yes	No

3) Supported Spatial Functions and Queries: Table IV, shows a summary for some spatial functions and queries supported by GeoSpark and SpatialIgnite. Table IV confirms the existence of some similarities between GeoSpark and SpatialIgnite in the supported spatial predicates, in which that GeoSpark and SpatialIgnite are supported for the same predicates, such as Equals, Intersects, Envelope, Touches. On the other hand, KNN-query is only supported by GeoSpark, where Range-query and Join-query are supported by the two frameworks (GeoSpark and SpatialIgnite). However, from Table IV we have observed that SpatialIgnite supports more spatial functionality than GeoSpark, but GeoSpark in Apache Spark is more suitable with spatial data queries and indexing because it support complex spatial queries, such as kNN queries, which is not supported in SpatialIgnite on Apache ignite. Moreover, GeoSpark is an agile spatial data processing framework that can meet the changes in the requirements, and fault tolerance.

4) Performance and Execution Time: According to the results and assumptions produced by Md Mahbub Alam [12], which proposed that SpatialIgnite has addressed the best performance among SaptialHadoop and GeoSpark, and from this research observation GeoSpark has been widely used for big spatial data processing, and many spatial frameworks was been depending on GeoSpark, such as Apache Sedona which used GeoSpark to activate spatial big data computing. We have decided to conducted this part of the research on GeoSpark to understand the performance rate of GeoSpark, and how much it can be worst according to the assumption in [12]. To illustrate this, the study was started by installing Apache Spark and open Spark localhost, then on Spark shell, Scala programming language was used and the dependency was added to activated GeoSpark on Apache Spark. However, the Spark build-in session or Spark SQL context was started,

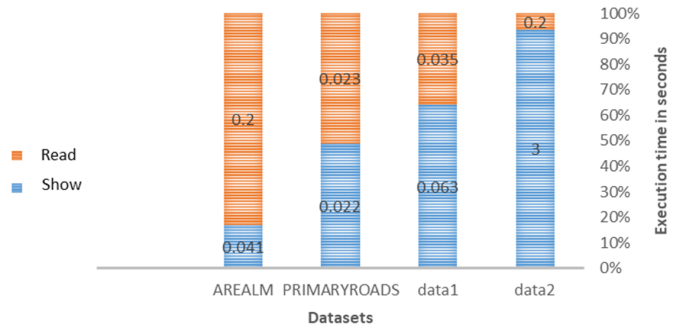


Fig. 3. A Compression among the Four Datasets when Read and Show their Contents on GeoSpark.

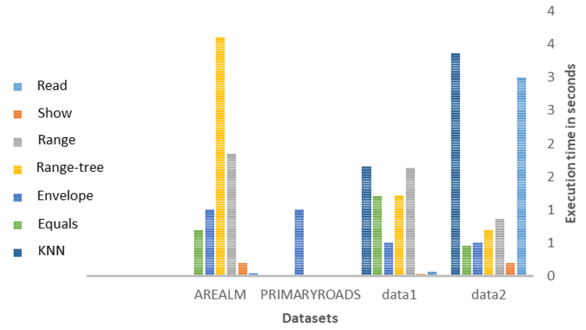


Fig. 4. Comparing the Execution Time for Different Functions and Queries on GeoSpark among the Four Datasets.

which is a class used for initializing the functionality of Spark SQL, where the Spark context is available as (sc), and Spark session is available as (spark). The datasets have been read using (spark.read), and execution time has been calculated for each dataset in the Apache Spark executor, and the comparison among the four datasets of the execution time of reading and show SQL functions in Fig. 3. However, as shown in Fig. 3 data 2 (8000 points geometry) has taken the highest reading time equals 3 seconds, where the show function for the same dataset takes only 0,2 seconds. On the other hand, the REALM dataset (129K polygon geometry) has recorded the lowest execution time for reading data on Spark equals 0.041 seconds. However, the ObjectRDD class was used for dealing with different spatial datasets geometries, such as PointRDD (data1 and data2), RectangleRDD (AREAL), and LineStringRDD (PRIMARYROA). However, two quires were examined with four operations on Geopark on the three datasets (data1, data2, AREALM), as shown in Table V. Additionally, PRIMARYROA datasets were presented multiple exceptions and errors so they excluded it and only datasets reading execution time was calculated. The Range query with tree indexing and without tree indexing has been applied on three datasets, KNN has applied only on the points datasets (data1 and data2), where Envelope and Equals have applied also on three datasets. After that, the execution time in seconds has estimated for each operation, and Table V shows more details about this step. Moreover, the obtained results were visualized in Fig. 4, which represented a comparison among the execution time required to examine the operations on the datasets. As a result, GeoSpark performance depends on the type of dataset geometries, but in general, the execution time of GeoSpark from our experimental study was acceptable.

IV. Results and Discussion

This section explained and discussed the main results that were observed from this research. However, the study was conducted on four main domains, and in each domain, was recorded some results, which were collected in this section. The observation was lead to that GeoSpark in Apache Spark supports better features than SpatialIgnite in Apache ignite, in which GeoSpark allow multiple file format, and multiple indexing, although that GeoSpark does not support spatial analysis as in SpatialIgnite. Moreover, SpatialIgnite supports more spatial functionality than GeoSpark, where KNN-query is only supported by GeoSpark. On the other hand, from experimental environment setup activation and working with spatial data computing on Apache Spark was more flexible than in Apache ignite, in which that preparing Apache Ignite experimental requirements have taken more time than Apache Spark, in which that spatial data computing on Apache Ignite required working on an IDE environment such as Netbeans, and adding some dependencies in the IDE environment, such as JTS Topology, was in Apache Spark does not depending on any IDE environment and the dependencies had been added on spark-shell in a simple code, and start working with the shell using Scala programming language. The last stage of the study built on the results and assumptions were produced by Md Mahbub Alam [12], which proposed that SpatialIgnite has addressed the best performance among GeoSpark, but from our observation, GeoSpark widely used for big spatial data processing, and many spatial frameworks have been depending on GeoSpark, such as Apache Sedona which used GeoSpark to activate spatial big data computing. Consequently, we have decided to conduct this part of the research on GeoSpark to understand the performance rate of GeoSpark. The result from this last step as it shows in Table V is mostly acceptable, but it depends on the size and the type of the dataset. However, the assumption by Md Mahbub Alam in his research can not be generalized on all Geospark and SpatialIgnite researches, in which multiple factors affected the performance of Geospark on Apache Spark, such as machine performance (CPU, and memory), dataset size, and type, programming environment (such as using spark-shell or IDE software). In the end, we have concluded that each one of these frameworks has its advantages over the other one, and each one of them can be used depending on the research area requirements and status.

V. Conclusion

Spatial computing is becoming increasingly relevant with the widespread use of mobile devices. The rise in scale and value of location data have led to the creation of a variety of specialized spatial data processing systems. In this research, we have conducted a comprehensive evaluation of big spatial data

TABLE V. COMPARING THE EXECUTION TIME FOR DIFFERENT FUNCTIONS AND QUERIES ON GEOSPARK AMONG THE FOUR DATASETS

Predicates	AREALM	PRIMARYROADS	data1	data2
Read	0.041	0.022	0.063	3
Show	0.2	0.023	0.035	0.2
Range	1.84	-	1.63	0.86
Range-tree	3.6	-	1.22	0.69
KNN	-	-	1.65	1.37
Envelope	1	-	0.5	0.5
Equals	0.70	-	1.2	0.46

computing on two data management systems Apache Ignite and Apache Spark. The comparative has been done on four different domains, experimental environment setup, supported features, supported functions and queries, and performance and execution time. The research concluded that each one of these frameworks has its advantages over the other one, and each one of them can be used depending on the research area requirements and status. However, the type and size of the dataset can have a large impact on the execution time. However, multiple factors affected the performance of Geospark on Apache spark, such as machine performance (CPU, and memory), dataset size and type, programming environment (such as using spark-shell or IDE software). From the observation GeoSpark has recorded more flexibility than SpatialIgnite on Apache ignite. Moreover, the spatial analytic features in SpatialIgnite on Apache ignite make it more suitable with spatial data analytic applications, such as analyzing spatial data, that extracted from social media. Additionally, GeoSpark in Apache Spark is more suitable with spatial data queries and indexing because it supports complex spatial queries, such as kNN queries, which is not supported in SpatialIgnite on Apache ignite. Moreover, GeoSpark is an agile spatial data processing framework that can meet the changes in the requirements, and fault tolerance.

REFERENCES

[1] Thulara N Hewage, Malka N Halgamuge, Ali Syed, and Gullu Ekici. Big data techniques of google, amazon, facebook and twitter. *Journal of Communications*, 13(2):94–100, 2018.

[2] Shadi A Noghabi, Kartik Paramasivam, Yi Pan, Navina Ramesh, Jon Bringhurst, Indranil Gupta, and Roy H Campbell. Samza: stateful scalable stream processing at linkedin. *Proceedings of the VLDB Endowment*, 10(12):1634–1645, 2017.

[3] Raghu Ramakrishnan, Baskar Sridharan, John R Douceur, Pavan Kasturi, Balaji Krishnamachari-Sampath, Karthick Krishnamoorthy, Peng Li, Mitica Manu, Spiro Michaylov, Rogério Ramos, et al. Azure data lake store: a hyperscale distributed file service for big data analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 51–63, 2017.

[4] Michael P Cope, Elena A Mikhailova, Christopher J Post, MA Schlautman, and P Carbajales-Dale. Developing and evaluating an esri story map as an educational tool. *Natural Sciences Education*, 47(1):1–9, 2018.

[5] Gouri Sankar Bhunia, Pravat Kumar Shit, and Debashish Sengupta. Free-open access geospatial data and tools for forest resources management. In *Spatial Modeling in Forest Resources Management*, pages 651–675. Springer, 2021.

[6] Malte Lech, Juha Ilari Uitto, Sven Harten, Geeta Batra, and Anupam Anand. Improving international development evaluation through geospatial data and analysis. *International Journal of Geospatial and Environmental Research*, 5(2):3, 2018.

[7] Salman Salloum, Ruslan Dautov, Xiaojun Chen, Patrick Xiaogang Peng, and Joshua Zhexue Huang. Big data analytics on apache spark. *International Journal of Data Science and Analytics*, 1(3-4):145–164, 2016.

[8] Michael A. Whitby, Rich Fecher, and Chris Bennight. Geowave: Utilizing distributed key-value stores for multidimensional data. In Michael Gertz, Matthias Renz, Xiaofang Zhou, Erik G. Hoel, Wei-Shinn Ku, Agnès Voisard, Chengyang Zhang, Haiquan Chen, Liang Tang, Yan Huang, Chang-Tien Lu, and Siva Ravada, editors, *Advances in Spatial and Temporal Databases - 15th International Symposium, SSTD 2017, Arlington, VA, USA, August 21-23, 2017, Proceedings*, volume 10411 of *Lecture Notes in Computer Science*, pages 105–122. Springer, 2017.

[9] Ahmed Eldawy, Louai Alarabi, and Mohamed F. Mokbel. Spatial partitioning techniques in spatial hadoop. *Proc. VLDB Endow.*, 8(12):1602–1605, 2015.

[10] Louai Alarabi, Mohamed F Mokbel, and Mashaal Musleh. St-hadoop: A mapreduce framework for spatio-temporal data. *GeoInformatica*, 22(4):785–813, 2018.

[11] Jia Yu, Zongsi Zhang, and Mohamed Sarwat. Spatial data management in apache spark: the geospark perspective and beyond. *GeoInformatica*, 23(1):37–78, 2019.

[12] Md Mahbub Alam, Suprio Ray, and Virendra C Bhavsar. A performance study of big spatial data systems. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, pages 1–9, 2018.

[13] Jia Yu, Zishan Fu, and Mohamed Sarwat. Dissecting geosparksim: a scalable microscopic road network traffic simulator in apache spark. *Distributed Parallel Databases*, 38(4):963–994, 2020.

[14] Louai Alarabi. Summit: a scalable system for massive trajectory data management. *ACM SIGSPATIAL Special*, 10(3):2–3, 2018.

[15] Yuanyuan Chen, Jingyi Yu, and Yong Gao. Detecting trajectory outliers based on spark. In *25th International Conference on Geoinformatics, Geoinformatics 2017, Buffalo, NY, USA, August 2-4, 2017*, pages 1–5. IEEE, 2017.

[16] Esteban Zimányi, Mahmoud Attia Sakr, and Arthur Lesuisse. Mobilitydb: A mobility database based on postgresql and postgis. *ACM Trans. Database Syst.*, 45(4):19:1–19:42, 2020.

[17] Aleem Akhtar. Role of apache software foundation in big data projects. *arXiv preprint arXiv:2005.02829*, 2020.

[18] Hadoop. Apache Hadoop. https://hadoop.apache.org/, 2021. [Online; accessed 23-Augst-2021].

[19] Spark. Apache Spark. https://spark.apache.org/, 2021. [Online; accessed 23-Augst-2021].

[20] Ignite. Apache Ignite. https://ignite.apache.org/, 2021. [Online; accessed 23-Augst-2021].

[21] Jia Yu, Jinxuan Wu, and Mohamed Sarwat. Geospark: A cluster computing framework for processing large-scale spatial data. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–4, 2015.

[22] Andrey Tapekhin, Igor Bogomolov, and Oleg Velikanov. Analysis of consistency for in memory data grid apache ignite. In *2019 Ivannikov Memorial Workshop (IVMEM)*, pages 46–50. IEEE, 2019.

[23] Yassine Benlachmi, Abdelaziz El Yazidi, and Moulay Lahcen Hasnaoui. A comparative analysis of hadoop and spark frameworks using word count algorithm. *International Journal of Advanced Computer Science and Applications*, 12(4), 2021.

[24] Cristiana-Stefania Stan, Adrian-Eduard Pandelica, Vlad-Andrei Zamfir, Roxana-Gabriela Stan, and Catalin Negru. Apache spark and apache ignite performance analysis. In *2019 22nd International Conference on Control Systems and Computer Science (CSCS)*, pages 726–733. IEEE, 2019.

[25] Zhibo Sun, Hong Zhang, Zixia Liu, Chen Xu, and Liqiang Wang. Migrating gis big data computing from hadoop to spark: an exemplary study using twitter. In *2016 IEEE 9th International Conference on Cloud Computing (CLOUD)*, pages 351–358. IEEE, 2016.

[26] Randall T Whitman, Michael B Park, Bryan G Marsh, and Erik G Hoel. Spatio-temporal join on apache spark. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–10, 2017.

[27] Jia Yu, Zongsi Zhang, and Mohamed Sarwat. Spatial data management in apache spark: The geospark perspective and beyond. *Geoinformatica*, 23(1):37–78, 2019.

[28] Jayati Gandhi, Nekita Chavhan, and Girish Kumar. Comparative study of spatial hadoop and geospark for geospatial big data analysis. *system*, 9:10, 2020.

[29] Mingjie Tang, Yongyang Yu, Qutaibah M Malluhi, Mourad Ouzzani, and Walid G Aref. Locationspark: A distributed in-memory data management system for big spatial data. *Proceedings of the VLDB Endowment*, 9(13):1565–1568, 2016.

[30] Francisco García-García, Antonio Corral, Luis Iribarne, George Mavrommatis, and Michael Vassilakopoulos. A comparison of distributed spatial data management systems for processing distance join queries. In *European Conference on Advances in Databases and Information Systems*, pages 214–228. Springer, 2017.

[31] Zhou Huang, Yiran Chen, Lin Wan, and Xia Peng. Geospark sql: An effective framework enabling spatial queries on spark. *ISPRS International Journal of Geo-Information*, 6(9):285, 2017.

[32] Panagiotis Nikitopoulos, Akrivi Vlachou, Christos Doulkeridis, and George A Vouros. Distrdf: Distributed spatio-temporal rdf queries on spark. In *EDBT/ICDT Workshops*, pages 125–132, 2018.

[33] Suprio Ray, Bogdan Simion, and Angela Demke Brown. Jackpine: A benchmark to evaluate spatial database performance. In *2011 IEEE 27th International Conference on Data Engineering*, pages 1139–1150. IEEE, 2011.

Carrot Disease Recognition using Deep Learning Approach for Sustainable Agriculture

Naimur Rashid Methun[1], Rumana Yasmin[2], Nasima Begum[3], Aditya Rajbongshi[4], Md. Ezharul Islam[5]
Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh[1,2,4,5]
Department of Computer Science and Engineering, University of Asia Pacific, Dhaka, Bangladesh[3]

***Abstract*—Carrot is a fast-growing and nutritious vegetable cultivated throughout the world for its edible roots. The farmers are still learning the scientific methods of carrot production worldwide. For the production of good quality carrots, modern technology is not being used to its fullest to detect carrot vegetable diseases in the farms. As a result, the farmers face difficulties now and then in continuous monitoring and detecting defects in carrot crops. Hence, this paper proposes an efficient carrot disease identification and classification method using a deep learning approach, especially Convolutional Neural Network (CNN). In this research, five different carrot diseases including healthy carrots have been examined and experimented with four different pretrained models of CNN architecture, i.e., VGG16, VGG19, MobileNet, and Inception v3. Among the four models, the Inception v3 model is selected as an efficient pretrained CNN architecture to build an effective and robust system. The Inception v3 based system proposed here takes carrot images as input and examines whether they are healthy or infected, and provides output accordingly. To train and evaluate the system, a robust dataset is used, which consists of original and synthetic data. In the Fully Connected Neural Network (FCNN), dropout is used to solve the problem of overfitting as well as to improve the accuracy of the system. The accuracy achieved from the method which uses Inception v3 is 97.4%, which is undoubtedly helpful for the farmers to identify carrot disease and maximize their benefits to establish sustainable agriculture.**

***Keywords*—Deep learning; convolutional neural network; Inception v3; carrot disease recognition**

I. Introduction

A very powerful concept that can alleviate extreme poverty is development in the agricultural sector. Many countries of the world are dependent on agriculture for their economic advancement. In the year 2018, 4% of the global GDP was added from the agricultural sector. In some developing countries, the amount of GDP that comes from agriculture is over 25% [1]. It is a livelihood for the vast majority of the population and directly affects the overall economy of many developing countries like Bangladesh, India, Pakistan, etc. As for example, in 2019, the share of Bangladesh, Pakistan, India and Morocco's GDP from agriculture was 12.68% [2], 22.04% [3], 15.96% [4] and 11.38% [5] respectively. In agriculture, especially for growing vegetables, early detection and classification of diseases allow the farmers to take preventive measures and reduce production loss as well as economic loss to ensure sustainability.

Experts have been identifying vegetable and fruit disease by naked eye observation over the last decades. However, this approach proves ineffective in many cases as it takes a lot of time to process, and experts are unavailable in rural areas. Since diseases leave some visible symptoms on the body of vegetables or fruits, it is possible to perform an imaging analysis of those visible defects on vegetables. Hence, Deep learning methods provide a solution to this issue.

Carrot (*Daucus carota*) is a vegetable that is rich in beta carotene, fiber, potassium, antioxidants, minerals, and vitamin k1. Frequent consumption of carrots decreases the risk of breast cancer [6]. It helps to prevent vitamin A deficiency and reduce heart diseases. Carrot is described as a functional food that provides benefits beyond essential nutrition, and it is even good for the eyes [7]. People living in rural areas of least developed countries face health problems due to nutritional deficiency, including vitamin A deficiency. Consumption of carrots can greatly contribute to a country's nutritional security where the standard of diets is relatively poor. Currently, carrot farmers mainly depend on human disease identification capability and experience to identify diseases that may produce inaccurate results. Hence it is indispensable to develop a fully automated carrot disease identification system.

In this work, we have discussed a solution based on deep learning approach that can correctly recognize the disease of carrots. We have experimented with five of the common carrot diseases, such as Black rot, Aster yellow, Sclerotinia rot, Root knot, and Growth crack. We have experimented with some pretrained models which include MobileNet CNN model [8], VGG16 [9], VGG19 [9], and Inception v3 [10] with the original and synthetic dataset. Finally, we have used Inception v3 in the proposed system because of the effectiveness of transfer learning for generating an accurate model in case of limited dataset. Our proposed system takes an image of carrot as input, the CNN layer automatically extracts features from that image, and Fully Connected Neural Network (FCNN) layer predicts the desired class it belongs to. The whole system was implemented using Python, TensorFlow [11], and Keras [12]. The main contribution of this research work is summarized as follows:

- We have proposed an efficient method to identify carrot diseases after analyzing images of infected carrots.
- We have expanded our dataset by generating synthetic data and applying some image processing techniques.
- We have experimented with four different CNN architectures and investigated the classification and detection results of carrot diseases.

- The accuracy rate achieved in this study outperforms the existing works in this domain.
- This work will provide assistance to detect carrot diseases earlier and solve the problems of farmers by taking measures to cure the diseases timely.

The organization of the rest of the paper is as follows: Section II contains a literature review of previous works and explains the major carrot diseases for recognition in this work. The system architecture is illustrated in Section III. In Section IV the research methodology is described, Section V shows the overall result, and Section VI concludes the work including future works.

II. Background and Literature Review

In this section we have discussed about six prominent diseases of carrot and then outlined the literature review.

A. Carrot Diseases Considered for Recognition

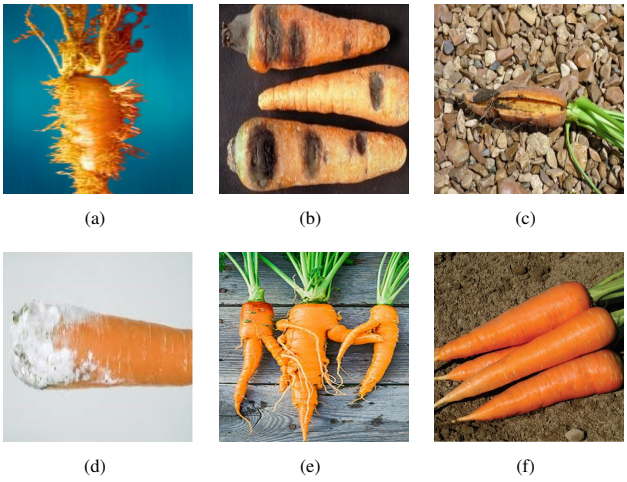


Fig. 1. (a) Aster yellow, (b) Black rot, (c) Growth crack, (d) Sclerotinia rot, (e) Root Knot, (f) Healthy carrot.

One of the main factors for deep learning method is the dataset. As we are working with carrot disease, we have selected six types of data, including a healthy carrot class and five other different diseases of carrot named as black rot, root knot, sclerotinia rot, growth crack, and aster yellow, as shown in Fig. 1. A short description of these diseases is as below:

(i) *Aster Yellow*: Aster yellow is a common disease in carrots caused by *Mycoplasma*. The affected carrots show symptoms such as yellowing leaves and the carrot tap root produces lots of fibrous side roots, tap roots become excessive hairy, tapered, and pale in color.
(ii) *Black Rot*: *Alternaria radicina* fungus causes Black rot in carrots [13]. This disease is mainly carried by seed and soil and is specified by a black decay at the crown area, which is shiny.
(iii) *Growth Crack*: Growth crack in carrots is caused by soil moisture. The carrot roots split in length.
(iv) *Sclerotinia Rot*: Sclerotinia rot, also known as white mold, is caused by *Sclerotinia sclerotiorum* fungus. The symptoms include fluffy white mycelial growth on the watery rot and black-colored sclerotia on the crown of the infected carrots.
(v) *Root Knot*: Root knot is generally caused by *Meloidogyne hapla* nematode. Forking, galls, hairiness, and stubby roots can be observed in the infected carrots. Very often, multiple tap roots form or roots are malformed.

B. Related Work

Several machine learning and deep learning approaches were developed through many years of research for object detection. But a few of them have been applied for detection of fruits and vegetables particularly focused on carrots.

A. Majumder *et al.* [14] developed a carrot disease recognition system using machine learning methods. In their work, they segmented the disease-affected region using k-means clustering algorithm. Then, they performed classification using Support Vector Machine (SVM) classifier. They used a total of 202 images and 11 feature sets in their work. This approach returned 96% accuracy. However, they have stated that if the quality of images is poor and has varying background color, it can cause distraction to provide results accurately.

S. Sasirekha *et al.* [15] described image processing techniques to identify and classify carrot vegetable disease. First, they converted the images from RGB to L*a*b color space. Then, they performed k-means clustering technique for segmentation purpose. A total of 13 features were extracted using texture and classification techniques. Then classification was done by multiclass SVM to identify the diseases of carrots.

G. C. Khadabadi *et al.* [16] used Probabilistic Neural Network (PNN) based classifier to recognize and identify disease in carrot vegetable. They applied Discrete Wavelet Transform (DWT) to extract features. Their proposed system gave an accuracy of 88.0% to classify carrot diseases.

H. Zhu *et al.* [17] proposed a carrot appearance quality identification based on deep learning. They have utilized AlexNet to extract features from carrot images and to identify carrot quality. The accuracy rate achieved in this work for binary classification recognition was 98.70%.

Rupali Saha *et al.* [18] proposed a method for the recognition of orange fruit disease involving a deep learning technique. They performed classification using CNN and used 8 feature set where the dataset size was 68. They claimed an accuracy of 93.21% in their proposed approach.

M. T. Habib *et al.* [19] applied machine learning method for detecting disease on papaya. First, they segmented the disease-attacked region using k-means clustering and performed classification using SVM. They used 10 feature set and the dataset size is 126. This approach returned an accuracy of 90.15%.

L. J. Rozario *et al.* [20] presented a method that involves identification of defective parts of fruits and vegetables. They experimented with four types of fruits and vegetables. They used modified k-means clustering and Otsu method for color-based segmentation of images. In this work, they used a total of 63 images. No classification was performed in this work.

S. A. Gaikwad *et al.* [21] presented a fruit disease detection and classification method based on image processing techniques. First, the images were segmented using k-means clustering algorithm. In the next step, some features were extracted from the segmented images. Finally, SVM classifier was used for classification purpose.

B. J. Samajpati *et al.* [22] experimented on three types of common apple diseases. They performed feature-level fusion. In total 13 features were extracted from every image. k-means clustering was applied for image segmentation. Random forest classifier was used for disease classification. They used a total of 80 images in their work. Their total accuracy ranged from 60-100% because of using various fusion of features. They did not carry out any performance analysis of the obtained results rather they showed only the differing accuracy.

M. Islam *et al.* in [23] proposed an integrated method combining image processing and machine learning technique for potato plant disease detection. Region of interest containing disease symptoms in the images were extracted in L*a*b* color spaces. In this work, 10 features were extracted. Multiclass SVM classifier was used to classify diseases over 300 images which gave an accuracy of 95%.

T. T. Mim *et al.* [24] worked on sponge gourd leaf and flower disease detection using CNN and image processing techniques. They used a very popular pretrained model called AlexNet for detecting diseases. Their proposed model architecture consisted of multiple layers of convolutional 2D layer and multiple ReLU activation function. They achieved 81.52% training accuracy and a loss of 0.5715 after thirty successful epochs.

M. T. Habib *et al.* [25] aimed to recognize jackfruit disease applying a machine vision based agromedical expert system. A total of 480 images were used in their work. First, they converted RGB images into L*a*b color space. Then k-means clustering algorithm was applied for image segmentation. They extracted 10 features from the images. Then they experimented with nine different classifiers for comparison purpose. Among all other classifiers, the random forest classifier produced 89.52% accuracy, which was the highest.

S. K. Behara *et al.* [26] presented a machine vision-based system to identify and measure the severity of disease in orange fruits. First, they separated the background, foreground, and defected region from the sample images using k-means clustering method. Then a total of 13 texture features were extracted from the defected region of images. After that, the extracted features were fed as input into the multiclass SVM classifier. Finally, a fuzzy logic model was used to compute the severity of disease. This work gave an accuracy of 90%.

Most of the existing works in the field of agriculture have been done with traditional machine learning. In most of the cases the authors have used separate segmentation algorithms and classifiers to extract features. However, no significant work has been done using the deep learning approach. In the proposed system, VGG16, VGG19, MobileNet, and Inception v3 models were used to train and test the infected carrot images among which the Inception v3 model outperformed the rest of the models.

III. System Architecture

The structural design of the proposed deep neural network based system for detecting and classifying carrot disease is shown in Fig. 2. At first, the farmer has to take an image of the infected carrot using a mobile phone or any other camera device. Then the image needs to be sent to the expert system via the internet. The expert system will analyze the image and compare it with the previously trained dataset to check if the carrot in the image is disease infected or not. If the carrot sample is disease infected, the system will identify the disease. Finally, the system will show the predicted disease name as output as well as give expert solutions from database.

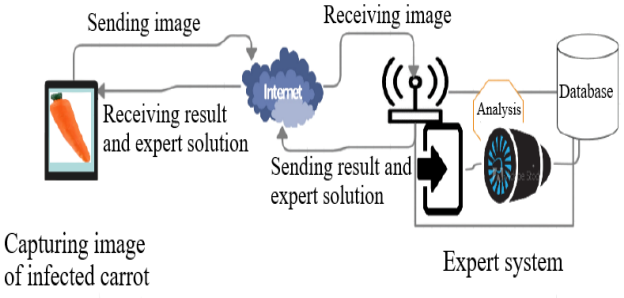


Fig. 2. Overall Architecture of the Proposed Carrot Disease Identification and Classification System.

IV. Research Methodology

In order to implement a carrot disease recognition system, a deep learning model has been built which is described in this section. Fig. 3 shows the detailed steps of building the framework for the system.

A. Image Acquisition and Preprocessing

Deep learning approaches need a large dataset to achieve more accuracy. In this research, a dataset of total 10,655 images are used for training including original, synthesized, and augmented data. A dataset of 480 original images has been created which includes 80 images per class (five disease classes and one healthy class). Some preprocessing techniques have been performed to increase the training dataset size. Another 1651 synthesized image dataset has been generated for six classes by changing the background of the original images. Synthesized data can be used and has a positive effect on training [28]. Data augmentation using TensorFlow and Keras is also applied in five different ways. The overall distribution of the dataset is shown in Table I. Both synthesized data and augmented data helped us to avoid overfitting. Before feeding the training images into the neural network, all the input images are resized (300x300) and rescaled (1/255). Then data augmentation is applied using *imageDataGenerator* [27] from TensorFlow. Finally, the dataset is ready for the training session.

B. Training of Images

We have experimented with some pretrained models, that includes MobileNet CNN model [8], VGG16 [9], VGG19 [9] and Inception v3 [10]. As the Inception v3 model provides

TABLE I. DISTRIBUTION OF DATASET

Class name	Collected data(c)	Synthetic data(s)	Image processing applied	Total data per class((s+c)*5)
Aster yellow	80	273	5 ways	1765
Black rot	80	274	5 ways	1770
Growth crack	80	276	5 ways	1780
Sclerotinia rot	80	276	5 ways	1780
Root Knot	80	276	5 ways	1780
Healthy carrot	80	276	5 ways	1780
Total	480	1651	5 ways	10655

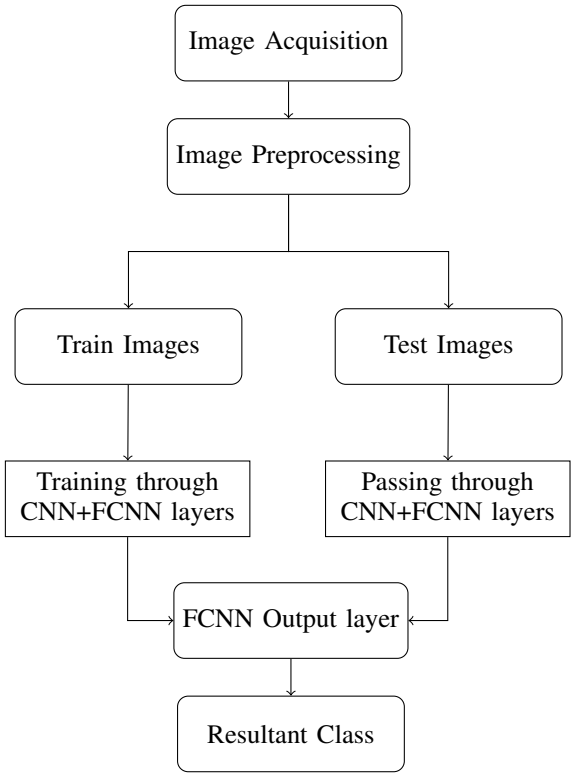


Fig. 3. Framework of the Proposed System.

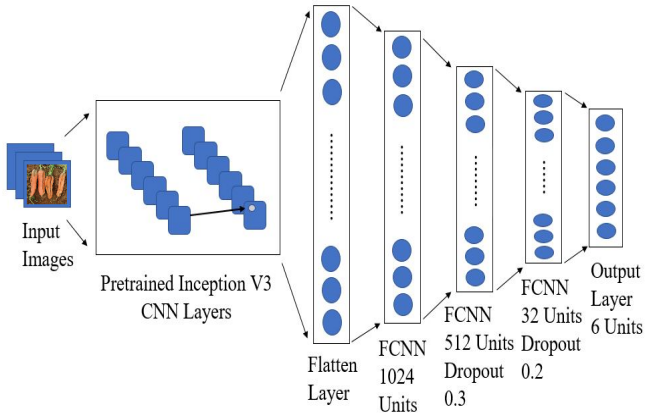


Fig. 4. Inception v3 based Neural Network Model for the Carrot Disease Recognition System.

better performance than the other models, the rest of this paper focuses on this model. However, the detailed comparison results are discussed in Section V. The FCNN Layers have been configured as shown in Fig. 4 to adapt the proposed system.

In the fully connected layer in FCNN, dropout [28] has been used to avoid overfitting. The last layer (the output layer) consists of six hidden units with the Softmax activation function. This function is used when we deal with multi-classification problems. It takes a value as an input and transforms it into a probability distribution whose total sum is 1.

The Softmax activation function is expressed as:

$$\sigma(\vec{x})_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}} \quad (1)$$

Where,

σ = Softmax,

$\vec{x}$ = input vector,

e^{x_i} = standard exponential function for input vector,

K = number of classes in the multi-class classifier, and

e^{x_j} = standard exponential function for output vector.

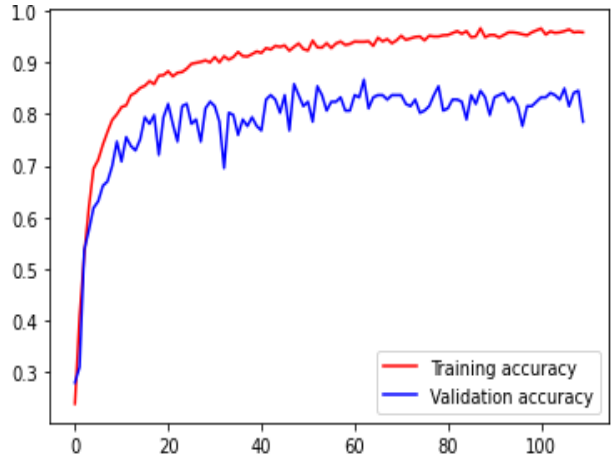
We used ReLu activation function in other layers. ReLu function can be expressed as:

$$f(y) = \begin{cases} 0, & \text{for } y < 0 \\ y, & \text{for } y \geq 0 \end{cases} \quad (2)$$

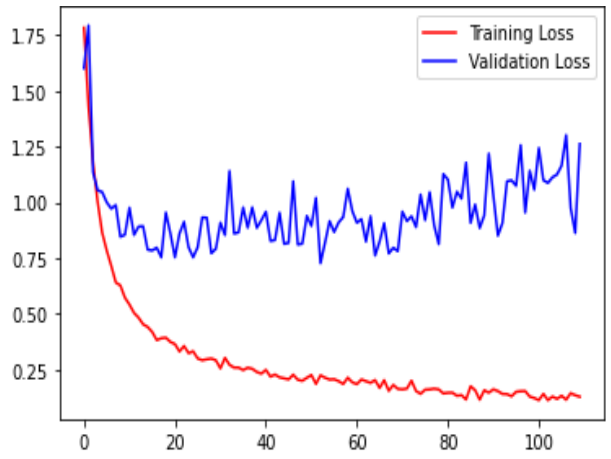
In this proposed system, RMSprop optimizer has been used as the optimization function and categorical cross-entropy has been used as the loss function. The learning rate has been set to 0.0001.

C. Output Generation

In this research, a total of 233 test images related to six classes have been used. Test images have been converted to the same form as training images by resizing and rescaling the images and fed into the proposed model. Next, the outputs of the

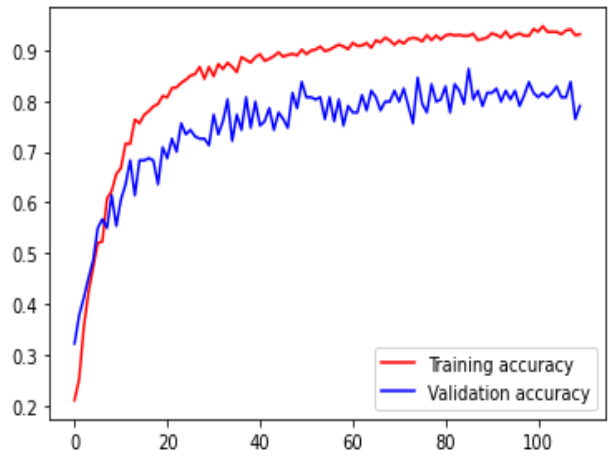


(a) Training and validation accuracy.

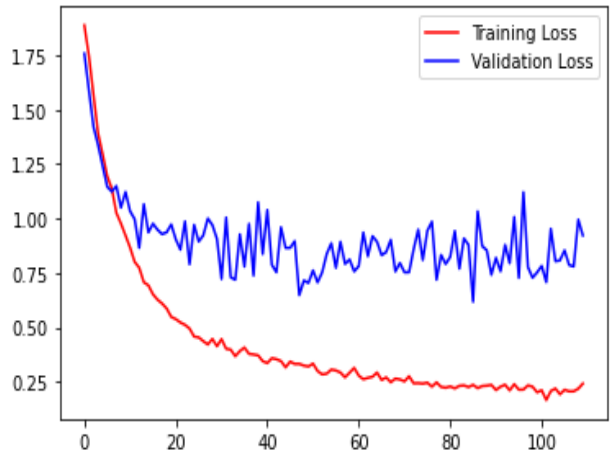


(b) Training and validation loss.

Fig. 5. Training and Validation Accuracy and Loss for VGG16 Model.



(a) Training and validation accuracy.



(b) Training and validation loss.

Fig. 6. Training and Validation Accuracy and Loss for VGG19 Model.

test images have been collected. To measure the performance of the classification system we need to compute *True Positive (TP)*, *False Positive (FP)*, *True Negative (TN)* and *False Negative (FN)*. These four cases of classification results can be represented by a confusion matrix. After testing and training the dataset, a 6×6 confusion matrix has been generated, which represents the correlation between the label and model's classification, that is, how successfully a classification model can predict.

Next, we have computed important performance metrics, i.e., accuracy, precision, recall, specificity, false positive rate (FPR), false negative rate (FNR), F-1 score from the confusion matrix so that the overall performance of the proposed system can be evaluated. The formula to calculate the performance metrics are given below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (5)$$

$$Specificity = \frac{TN}{TN + FP} \times 100\% \quad (6)$$

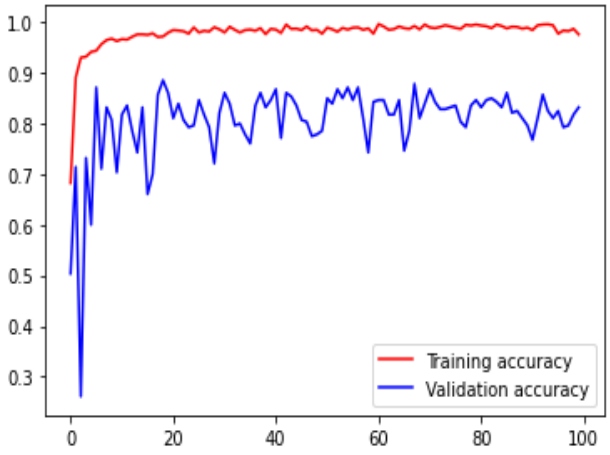
$$FPR = \frac{FP}{FP + TN} \times 100\% \quad (7)$$

$$FNR = \frac{FN}{FN + TP} \times 100\% \quad (8)$$

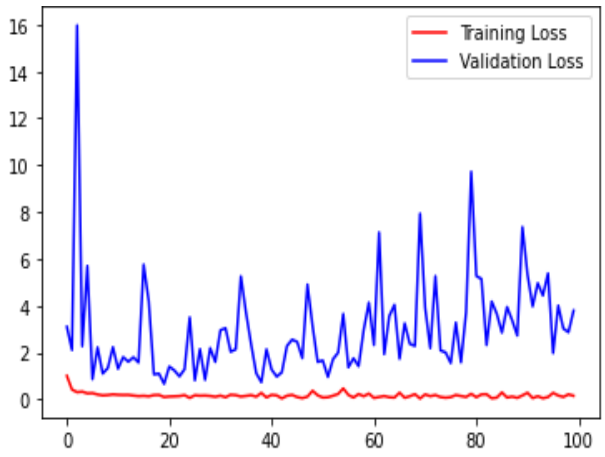
$$F1score = \frac{2 * Precision * Recall}{Precision + Recall} \times 100\% \quad (9)$$

V. Results and Discussion

The experimental analysis of the proposed system, model-wise analysis, and comparative analysis with other related works have been discussed in this section.

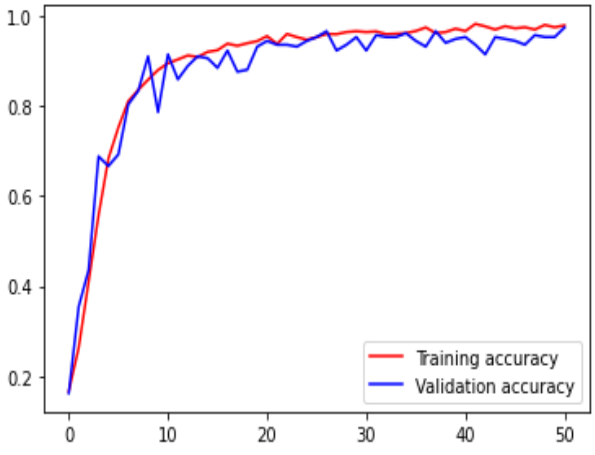


(a) Training and validation accuracy.

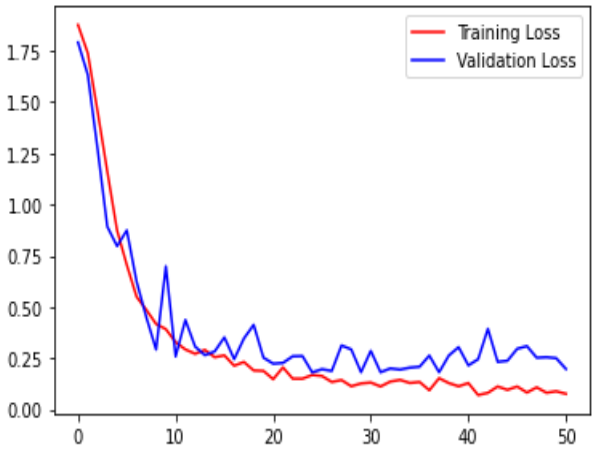


(b) Training and validation loss.

Fig. 7. Training and Validation Accuracy and Loss for MobileNet Model.



(a) Training and validation accuracy.



(b) Training and validation loss.

Fig. 8. Training and Validation Accuracy and Loss for Inception v3 Model.

A. Experimental Analysis of System

In this paper, we have worked with carrot disease. Therefore, six types of different data have been selected for the proposed system. Five of them are carrot diseases, and the other one is healthy carrot. After collecting different training and validation datasets, training data is fed in the final model to train the system, and validation data is fed to check its accuracy. A callback function is applied to save the best model with the highest accuracy. We have experimented with VGG16, VGG19, MobileNet, and Inception v3 models. The whole process has been implemented in Python, TensorFlow, and Keras. The epoch wise accuracy and loss for each of the models can be visualized in Fig. 5, Fig. 6, Fig. 7, and Fig. 8 respectively. It can be observed from Fig. 5, Fig. 6, Fig. 7, and Fig. 8 that the accuracy of the VGG16, VGG19, MobileNet and Inception v3 models is 86.6%, 86.2%, 89.2% and 97.4% respectively and the validation loss for the model VGG16 is more than 1.77%, VGG19 model gives 1.75% validation loss, MobileNet model gives the highest validation loss among the four models which is almost 16% and the Inception v3 model gives 1.76% validation loss.

After collecting the results on test dataset, a confusion matrix has been formed for each of the models. The confusion matrix generated from the models are shown in Fig. 9, Fig. 10, Fig. 11 and Fig. 12 respectively. The accuracy rate calculated from the confusion matrices in Fig. 9, Fig. 10 and Fig. 11 for the models VGG16, VGG19 and MobileNet is found to be 86.6%, 86.2% and 89.2% respectively. From the confusion matrix in Fig. 12, the accuracy of the proposed system based on Inception v3 model is found as high as 97.4%.

B. Model-wise Result Analysis

Before selecting Inception v3 as the model for transfer learning, few more state-of-the-art pretrained models have been experimented. VGG16 and VGG19 pretrained models provide us with 86.6% and 86.2% accuracy respectively. Nevertheless, we expected a more accurate system. Therefore, the MobileNet CNN model has been experimented which provides an accuracy of 89.2%. At last, the Inception v3 pretrained model has been applied, which provides the highest accuracy of 97.4%, as mentioned previously. Table II, Table III, Table IV and Table V shows the performance for each of the diseases of the models VGG16, VGG19, MobileNet, and Inception v3 respectively. All the performance metrics are calculated using the equations

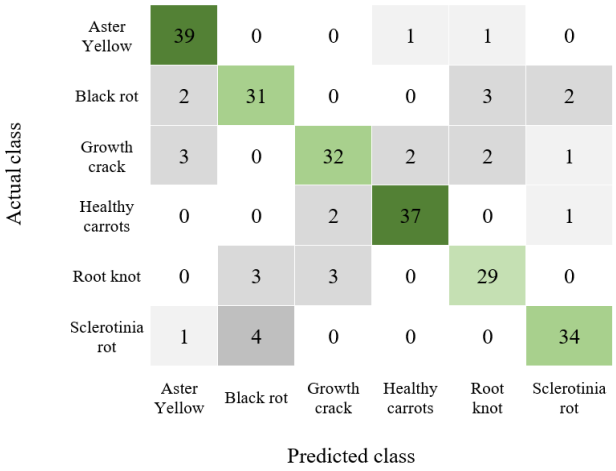


Fig. 9. Confusion Matrix for VGG16.

Fig. 10. Confusion Matrix for VGG19.

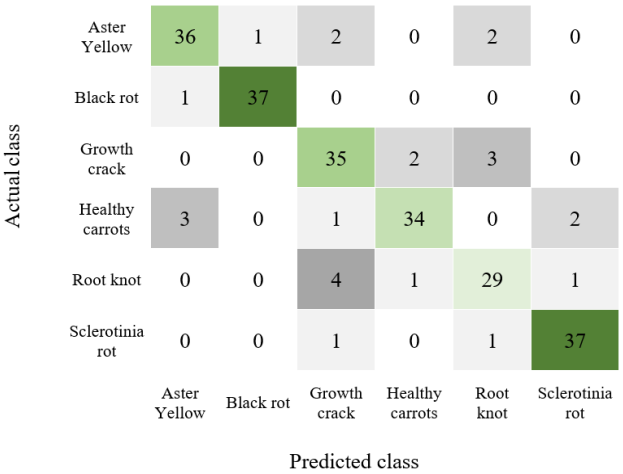


Fig. 11. Confusion Matrix for MobileNet.

Fig. 12. Confusion Matrix for Inception V3.

3, 4, 5, 6, 7, and 8 respectively and the results are shown in Table II, Table III, Table IV and Table V.

From the disease wise achieved performance for the model VGG16 in Table II, it can be observed that the healthy class gives the highest accuracy of 97.4% and the class black rot gives the lowest accuracy of 93.9%. The highest FPR is 3.5% for the class black rot, and the lowest FPR is 1.5% for the healthy class.

Table III summarizes the disease wise performance for the VGG19 model, and it can be seen that the healthy, growth crack, and sclerotinia rot classes give the highest accuracy rate of 96.1% and the class black rot gives the lowest accuracy of 93.5%. In this case, the highest FPR is 5% for the class root knot and the lowest FPR is 1% for the healthy class.

Table IV gives an outline of the disease wise performance for the MobileNet model in which the highest accuracy of 99.1% is found for the class black rot and the lowest accuracy rate for the class growth crack of 94.4%. The highest FPR is found to be 3% for the class root knot and the lowest FPR is 0.5% for the black rot.

It can be seen from the disease wise performance of the Inception v3 model in Table V that the class sclerotinia rot achieves the highest accuracy of 99.5% while the class root knot gives the lowest accuracy of 98.7%. The highest FPR is 1% for the classes growth crack and root knot, and the lowest FPR is found 0% for the healthy and sclerotinia rot classes, which justifies the robustness of the proposed methodology.

From the comparison of these tables, it can be noted that the model Inception v3 gives the highest overall system accuracy of 97.4% and the model VGG19 gives the lowest accuracy rate of 86.2%. Fig. 13 shows the overall comparison of accuracy, precision, recall, and F1 score among the four models on the dataset. The accuracy rate achieved mathematically justifies the validation accuracy graphs of the respective models in Fig. 5, Fig. 6, Fig. 7, and Fig. 8.

C. Comparative Analysis

The study of literature demonstrates that most of the papers have applied traditional machine learning methods. Very few

TABLE II. DISEASE WISE ACHIEVED PERFORMANCE FOR VGG16 MODEL

Evaluation Metrics	Healthy	Aster Yellow	Black Rot	Growth Crack	Root Knot	Scelerotinia Rot	System Performance
Precision(%)	92.5	86.6	81.5	86.4	82.8	89.4	86.6
Recall(%)	92.5	95.1	81.5	80	82.8	87.1	86.6
Accuracy(%)	97.4	96.5	93.9	94.4	94.8	96.1	86.6
Specificity(%)	98.4	96.8	96.4	97.4	96.9	97.9	97.3
FPR(%)	1.5	3.1	3.5	2.5	3	2	2.6
FNR(%)	7.5	4.8	18.4	20	17.14	12.8	13.3

TABLE III. DISEASE WISE ACHIEVED PERFORMANCE FOR VGG19 MODEL

Evaluation Metrics	Healthy	Aster Yellow	Black Rot	Growth Crack	Root Knot	Scelerotinia Rot	System Performance
Precision(%)	94.4	86	87	91.8	76.7	84	86.2
Recall(%)	82.5	90.2	71	85	94.2	94.8	86.2
Accuracy(%)	96.1	95.7	93.5	96.1	94.8	96.1	86.2
Specificity(%)	98.9	96.8	97.9	98.4	94.9	96.3	97.2
FPR(%)	1	3.1	2	1.5	5	3.6	2.7
FNR(%)	17.5	9.7	28.9	15	5.7	5.1	13.7

TABLE IV. DISEASE WISE ACHIEVED PERFORMANCE FOR MOBILENET MODEL

Evaluation Metrics	Healthy	Aster Yellow	Black Rot	Growth Crack	Root Knot	Scelerotinia Rot	System Performance
Precision(%)	91.89	90	97.3	81.3	82.8	92.5	89.2
Recall(%)	85	87.8	97.3	87.5	82.8	94.8	89.2
Accuracy(%)	96.1	96.1	99.1	94.4	94.8	97.8	89.2
Specificity(%)	98.4	97.9	99.4	95.8	96.9	98.4	97.8
FPR(%)	1.5	2	0.5	4.1	3	1.5	2.1
FNR(%)	15	12.1	2.6	12.5	17.1	5.1	10.7

TABLE V. DISEASE WISE ACHIEVED PERFORMANCE FOR INCEPTION V3 MODEL

Evaluation Metrics	Healthy	Aster Yellow	Black Rot	Growth Crack	Root Knot	Scelerotinia Rot	System Performance
Precision(%)	100	97.6	97.3	95.2	94.4	100	97.4
Recall(%)	95	95.6	97.3	100	97.1	97.4	97.4
Accuracy(%)	99.1	99.1	99.1	99.1	98.7	99.5	97.4
Specificity(%)	100	99.4	99.4	98.9	98.9	100	99.4
FPR(%)	0	0.5	0.5	1	1	0	0.5
FNR(%)	5	2.3	2.6	0	2.8	2.5	2.5

of them have applied deep learning approach. Clearly, the machine learning approach works well in the case of a limited dataset. Consequently, synthetic data can be useful for deep learning approaches. To assess the merit of the stated system, we have compared it with the results of some most recent relevant research. A. Majumder *et al.* [14] used k-means clustering for segmentation and SVM as a classifier to recognize carrot disease, which achieved 96% accuracy. S. Sasirekha *et al.* [15] employed image processing techniques to identify and classify diseases of carrot vegetables. However, the authors did not mention any accuracy of their proposed work. G. C. Khadabadi *et al.* [16] used PNN classifier to recognize and identify carrot diseases with an accuracy rate of 88.0%. The accuracy gained from these methods is not satisfactory. So, we have proposed an efficient system employing the Inception v3 model to recognize and classify carrot diseases. Comparing these outcomes with the proposed system, it can be said that this work has achieved a higher accuracy which is 97.4%. The detailed comparison of our proposed model with other related works is presented in Table VI.

TABLE VI. COMPARISON WITH RELATED WORKS

Reference	Problem Domain	Object	Total size of Dataset	Segmentation Algorithm	Classification Performed	Classifier	No of features detected	Accuracy
This work	Disease Detection and Classification	Carrot	2131	—	Yes	CNN	—	97.4%
A. Majumder *et al.* (2019) [14]	Disease Detection and Classification	Carrot	202	k-means clustering	Yes	SVM	11	96%
S. Sasirekha *et al.* (2019) [15]	Disease Recognition	Carrot	Not mentioned	k-means clustering	Yes	Multiclass SVM	13	Not mentioned
G. C. Khadabadi *et al.* (2015) [16]	Disease Detection and Classification	Carrot	50	Not mentioned	Yes	PNN	4	88.0%
Rupali Saha *et al.* (2020) [18]	Disease Detection and Classification	Orange	68	Color threshold segmentation	Yes	CNN	8	93.21%
M. T. Habib *et al.* (2020) [19]	Disease Detection and Classification	Papaya	129	k-means clustering	Yes	SVM	10	90.15%
L. J. Rozario *et al.* (2016) [20]	Segmentation and disease Identification	Apple, Banana, Tomato, Potato	63	k-means clustering, modifies k-means clustering, otsu method	No	Not mentioned	Not mentioned	Not mentioned
S. A. Gaikwad *et al.* (2017) [21]	Disease Detection and Classification	Fruit	Not mentioned	k-means clustering	Yes	SVM	Not mentioned	Not mentioned
B. J. Samajpati *et al.* (2016) [22]	Disease Detection and Classification	Apple	80	k-means clustering	Yes	Random Forest Classifier	13	60-100%
M. Dhakate *et al.* (2015) [29]	Disease Detection and Classification	Pomegranate	500	k-means clustering	No	Not mentioned	8	90%
M. Islam *et al.* (2017) [23]	Disease Detection and Classification	Potato leaf	300	Not mentioned	Yes	Multiclass SVM	10	95%
T. T. Mim *et al.* (2020) [24]	Disease Recognition	Sponge Gourd Leaf and Flower	6000	Not mentioned	Yes	CNN	Not mentioned	81.52%
M. T. Habib *et al.* (2020) [25]	Disease Recognition	Jackfruit fruit and leaf	480	k-means clustering	Yes	Random Forest Classifier	10	89.59%
S. K. Behara *et al.* (2018) [26]	Disease Classification and measurement of severity of disease	Orange	20	k-means clustering	Yes	Multiclass SVM	13	90%

VI. CONCLUSION AND FUTURE WORK

In this research, we have compared four different CNN architectures, i.e., VGG16, VGG19, MobileNet, and Inception v3 to identify and classify five major carrot diseases, namely Aster yellow, Black rot, Growth crack, Root knot, and Sclerotinia rot. Observing the experimental results of all four models, the Inception v3 model has shown its worth in detecting and classifying carrot diseases, assuring 97.4% accuracy. The methodology stated in this paper will help the farmers to identify carrot disease and take real-time actions. We hope that it would promote farmers worldwide and agricultural experts to start smart farming which will save their time and reduce economic loss as well as achieve sustainable agriculture.

In the future, we can extend this research by increasing the number of identified diseases with a larger dataset and experiment with various other classification methods. We wish to work on identifying disease severity as well. As IoT and mobile technology are gaining popularity among the general mass, we plan to create a web-based application as well as a multilingual mobile application implementing the proposed method wherein the rural farmers can be benefited from real-time solutions.

REFERENCES

[1] "Overview." WorldBank. [Online]. Available: https://www.worldbank.org/en/topic/agriculture/overview, Sep.30 2020. Accessed on 2020-11-25.

[2] "Bangladesh: Share of economic sectors in the gross domestic product (gdp) from 2009 to 2019." [Online]. Available: https://www.statista.com/statistics/438359/share-of-economic-sectors-in-the-gdp-in-bangladesh/, Nov 2020. Accessed on 2020-11-25.

[3] "Pakistan - gdp distribution across economic sectors 2019." [Online]. Available: https://www.statista.com/statistics/383256/pakistan-gdp-distribution-across-economic-sectors/, 18 Nov 2020. Accessed on 2020-12-19.

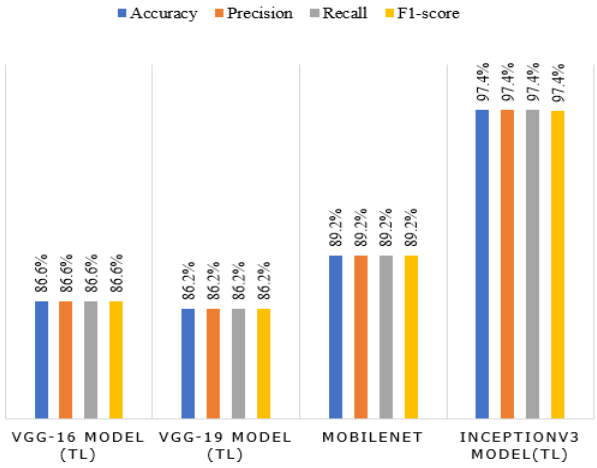


Fig. 13. Measurement Metrics for Different Deep Learning Models.

[4] "India: Distribution of gross domestic product (gdp) across economic sectors from 2009 to 2019." [Online]. Available: https://www.statista.com/statistics/271329/distribution-of-gross-domestic-product-gdp-across-economic-sectors-in-india/, 18 Nov 2020. Accessed on 2020-12-19.

[5] "Morocco: Distribution of gross domestic product (gdp) across economic sectors from 2009 to 2019." [Online]. Available: https://www.statista.com/statistics/502771/morocco-gdp-distribution-across-economic-sectors/, 18 Nov 2020. Accessed on 2020-12-19.

[6] M. P. Longnecker, P. A. Newcomb, R. Mittendorf, E. R. Greenberg, and W. C. Willett, "Intake of carrots, spinach, and supplements containing vitamin a in relation to risk of breast cancer," Cancer Epidemiology and Prevention Biomarkers, vol. 6, no. 11, pp. 887–892, 1997.

[7] S. A. Arscott and S. A. Tanumihardjo, "Carrots of many colors provide basic nutrition and bioavailable phytochemicals acting as a functional food," Comprehensive Reviews in Food Science and Food Safety, vol. 9, no. 2, pp. 223–239, 2010.

[8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826, 2016.

[11] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., "Tensorflow: A system for large-scale machine learning," in 12th USENIX symposium on operating systems design and implementation (OSDI16), pp. 265–283, 2016.

[12] F. Chollet et al., "Keras documentation,"Keras. io, 2015.

[13] F. C. Meier, C. Drechsler, E. Eddy, et al., "Black rot of carrots caused by alternaria radicina n. sp.," Phytopathology, vol. 12, no. 4, 1922.

[14] A. Majumder, M. T. Habib, P. H. Lima, S. Sourav, and R. N. Nandi, "Automated carrot disease recognition: a computer vision approach," International Journal of Engineering Technology, vol. 7, no. 4, pp. 5790–5797, 2019.

[15] S. Sasirekha and K. B. Suganthy, "An approach for detection of diseasein carrot using k-means clustering," International Journal of Research in Engineering, Science and Management (IJRESM), vol. 2, no. 2, pp. 527–530, February-2019.

[16] G. C. Khadabadi, A. Kumar, and V. S. Rajpurohit, "Identification and classification of diseases in carrot vegetable using discrete wavelet transform," in 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), pp. 59–64, IEEE, 2015.

[17] H. Zhu, L. Deng, D. Wang, J. Gao, J. Ni, and Z. Han, "Identifying carrot appearance quality by transfer learning," Journal of Food Process Engineering, vol. 42, no. 6, p. e13187, 2019.

[18] R. Saha and S. Neware, "Orange fruit disease classification using deep learning approach," International Journal, vol. 9, no. 2, 2020.

[19] M. T. Habib, A. Majumder, A. Jakaria, M. Akter, M. S. Uddin, and F. Ahmed, "Machine vision based papaya disease recognition," Journal of King Saud University Computer and Information Sciences, vol. 32, no. 3, pp. 300–309, 2020.

[20] L. J. Rozario, T. Rahman, and M. S. Uddin, "Segmentation of the region of defects in fruits and vegetables," International Journal of Computer Science and Information Security, vol. 14, no. 5, p. 399, 2016.

[21] S. A. Gaikwad, K. S. Deore, M. K. Waykar, P. R. Dudhane, and G. Sorate, "Fruit disease detection and classification," International Research Journal of Engineering and Technology, vol. 4, pp. 1151–1154, 2017.

[22] B. J. Samajpati and S. D. Degadwala, "Hybrid approach for apple fruit diseases detection and classification using random forest classifier," in 2016 International Conference on Communication and Signal Processing (ICCSP), pp. 1015–1019, IEEE, 2016.

[23] M. Islam, A. Dinh, K. Wahid, and P. Bhowmik, "Detection of potato diseases using image segmentation and multiclass support vector machine," in 2017 IEEE 30th Canadian conference on electrical and computer engineering (CCECE), pp. 1–4, IEEE, 2017.

[24] T. T. Mim, M. H. Sheikh, S. Chowdhury, R. Akter, M. A. A. Khan, and M. T. Habib, "Deep learning based sponge gourd diseases recognition for commercial cultivation in bangladesh," in International Conference on Artificial Intelligence & Industrial Applications, pp. 415–427, Springer, 2020.

[25] M. T. Habib, M. J. Mia, M. S. Uddin, and F. Ahmed, "An in-depth exploration of automated jackfruit disease recognition," Journal of King Saud University Computer and Information Sciences, 2020.

[26] S. K. Behera, L. Jena, A. K. Rath, and P. K. Sethy, "Disease classification and grading of orange using machine learning and fuzzy logic," in 2018 International Conference on Communication and Signal Processing (ICCSP), pp. 0678–0682, IEEE, 2018.

[27] F. Chollet, "Building powerful image classification models using very little data," Keras Blog, 2016.

[28] P. Baldi and P. J. Sadowski, "Understanding dropout," Advances in neural information processing systems, vol. 26, pp. 2814–2822, 2013.

[29] M. Dhakate and A. Ingole, "Diagnosis of pomegranate plant diseases using neural network," in2015 fifth national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG), pp. 1–4, IEEE, 2015.

Data Security: A New Symmetric Cryptosystem based on Graph Theory

Khalid Bekkaoui, Soumia Ziti, Fouzia Omary
Intelligent Processing and Security of Systems(IPSS)
Faculty of Sciences, Mohammed V University
in Rabat, Morocco

***Abstract*—Sharing private data in an unsecured channel is extremely critical, as unauthorized entities can intercept it and could break its privacy. The design of a cryptosystem that fulfills the security requirements in terms of confidentiality, integrity and authenticity of transmitted data has therefore become an unavoidable imperative. Indeed, a lot of work has been carried out in this regard. Although many cryptosystems have been proposed in the published literature, it has been found that their robustness and performance vary relatively from one to another. Adopting this reflection, we address in this paper the concept of block cipher, which is a major cryptographic solution to guarantee confidentiality, by involving the properties of graph theory to represent the plaintext message. Our proposal is in fact a new symmetric encryption block cipher that proceeds by representing plaintext messages using disjoint Hamiltonian circuits and then dealing with them as an adjacency matrix in a pre-encryption phase. The proposed system relies on a particular sub-key generator that has been carefully designed to produce the encryption keys according to the specifications of the system. The obtained experimental results demonstrate that our proposed cryptosystem is robust against statistical attacks, particularly the DIEHARD test, and presents both good confusion and good diffusion.**

***Keywords*—Cryptosystem; graph theory; hamiltonian circuits; adjacency matrix; block cipher; encryption**

I. Introduction

Cryptography is a component of cryptology that is based on a number of methods and principles for converting a readable message to a totally unreadable one. This field is dealing with many security problems such as the confidentiality of communications via non-secure channels, the privacy of individuals, the data storing on unsecured mediums, and so on. Cryptography refers to the study and analysis of data encryption systems intended to reduce the impact of hackers and to prevent, as best as possible, any unauthorized attempts to gain access to these confidential data. The main principles of information security, notably confidentiality, integrity, authentication, and non-repudiation [1].

Confidentiality is a crucial part of security. It can be ensured by an encryption process, whereby the data becomes non-intelligible to any non-authorized parties trying to gain access to it. The idea behind of encryption process is to transform a plaintext into a ciphertext, so only authorized parties can obtain the message in its original format by reversing the encryption process, known as decryption. Technically, decryption should be extremely difficult for any unauthorized and unqualified parties attempting to perform it.

Over the years, cryptography has continued to be improved and has progressively became an indispensable part for private data sharing. All contributions dedicated to this field of research have aroused great interest. In the literature, cryptography can be classified into three categories: Symmetric Key Cryptography, which is an encryption system where both the transmitter and the recipient of the message use one common key such as DES [2], AES [3], or IDEA [4], to encrypt and decrypt the messages. The second category is Asymmetric Key Cryptography. In this system, a couple of keys(private and public keys) are used in order to encrypt and decrypt the messages such as RSA [5], ElGamal [6], Diffie-Hellman [7], etc. The last category is Hybrid key Crypthography, Which consist of using an encryption mode that utilizes both symmetric and asymmetric public key encryption. This method benefits from public key cryptography for key sharing and from the speed of symmetric encryption for message encryption. Nowadays, cryptology is able to handle a substantial set of mathematical tools, that allowed for improvements in terms of efficiency and performance. In particular, graph theory is a field that is considered very promising in this regard, since it provides concepts that could be useful in solving problems in every network related areas.

Graph theory in mathematics refers to the study of graphs, which are a major object of discrete mathematics. Generally, a graph is represented as a set of vertices linked by edges. They are thus mathematical structures used for modelling pair-wise relationships between objects. It can be found in road networks, electrical circuits, constellations, etc. Graphs provide a way of thinking that can be used for modeling a vast range of problems. They are the foundation of numerous computer programs that allow communication and advanced technological processes. The seven bridges of Konigsberg (1736) [8] is a mathematical problem well known for having established the foundations of the theory of graphs. Graph theory is a relatively new concept that has been successfully incorporated and has enabled the development of stronger encryption algorithms that have proven to be difficult to break, even for the latest software solutions. In fact, it consists of modeling encryption problems by graph representation, so that they eventually become problems in graph theory where the solutions are usually well-known. Although solutions to graph problems can be fairly easy and efficient (with respect to the time required for computational processing which is reasonable), they can also be rather difficult (relative to the processing time increases exponentially). This resulted in the application of concepts introduced in graph theory to large-scale cryptography, since many NP-hard problems are derived

from this theory.

Considering the above mentioned points, the design of cryptosystems based on the concepts of graph theory is of utmost importance. In this work, we present a new cryptosystem that takes advantage of the principles of graph theory, which enable a high degree of security while maintaining the performance of data processing. The main idea of our approach is to represent the plaintext with all disjoint Hamiltonian circuits as a pre-encryption phase, then using our own sub-keys generator following the cipher block chaining mode of operation to encrypt the plaintext.

The rest of the paper is structured as follows. Section 2 presents preliminary knowledge. Section 3 presents a literature review of related work. Section 4 details the proposed scheme. Security analysis and experimental results are elaborated in Section 5, and lastly, the conclusion and future works is given in Section 6.

II. Preliminary Knowledge

- **Graph**: A graph G is a set of points called vertices V and a set of lines called edges E that connect some vertices together. The graph is defined as a set of vertices and edges that form a pair $G = (V, E)$.
- **Simple graph**: A graph in which each pair of vertices is linked by at the very most one edge and where no vertex has a loop.
- **Undirected Graph**: An undirected graph G is a pair (V, A) where V is a finite set of vertices and A is a set of unordered pairs of vertices. Also, loops are not allowed in undirected graphs.
- **Cycle**: A chain whose start and end nodes are the same and which does not use the same link more than once.
- **Hamiltonian Path**: A path that passes once and only once through each of the vertices of an undirected graph.
- **Hamiltonian Circuit**: simple cycle passing through all the vertices of a graph one and only once.
- **Adjacency Matrix**: Let G be an undirected graph with m vertices from 1 to m. We call the adjacency matrix of graph the matrix $A = (a_{jk})$ where a_{jk} is the total number of edges joining vertex j to vertex k:

$$\begin{cases} a_{jk} = w & \text{if and only if j and k are adjacent.} \\ a_{jk} = 0 & \text{if not.} \end{cases} \tag{1}$$

  with w is the weight of the edge (j, k).

- **Blum Blum Shub (BBS)**: is a pseudo-random number generator first proposed in [9]

$$x_{n+1} = x_n^2 \mod M \tag{2}$$

  With $M = pq$ the product of two large primes p and q.
  The complexity of the factorization of M is the main basis for the security of this generator, which means that the two primes must be carefully chosen to guarantee robustness.

III. Related Work

The application of graph theory in cryptography has become more emergent. However various encryption techniques have been proposed in this context.
A technique has been proposed by Amudha et al [10] that encodes clear messages through the Euler graph, the key used to protect the data in this approach is a kind of Hamilton circle. The authors in [11] sequentially construct three different graphs on the basis of an unconventional mapping, conjectured to be a one-way trapdoor function and designed specifically for graph structures. Some work focuses on the application of graph theory principles in computer networks and its potential to tackle the challenges of provisioning in secure cloud computing environments [12]. Two graph based public key cryptosystems have been suggested to secure sensitive Data in the work of Sensarma et al [13], where one is based purely on the properties of matrices, while the other is based on graph codes. In the work described in [14], the authors proposed a hybrid Cipher Block Chaining encryption system for e-mail protection. The suggestion was predicated on the integration of encryption technologies. Yousif et al [15], introduced a process to produce a new key on the basis of chaotic maps that are utilized to encode images. Within the work in [16], the emphasis is on the possibility of employing the Euler graph as a method object used in the remote method invocation (RMI) technique.

Among the most recent works, we mention the work presented in [17], where a block cipher system has been proposed using disjoint Hamiltonian circuits to present the data as a graph. Also in [18], a double vertex graph has been suggested to encrypt a word. At first, the given message was encrypted using the encryption table. The plaintext was then converted into a path graph. From the latter, a double vertex graph is constructed. We also mention the work [19], in which the original message is converted into several graphs. The ciphertext is obtained from the projection of the adjacency matrices representing the graphs into the secret key. A number of other proposals were suggested in the same thematic area [20], [21].

The originality of our work lies in the fact that our proposed system was able to blend both the concept of block ciphers, which is a major category of symmetric cryptography, and graph theory properties for representing plaintext, in particular Hamiltonian circuits, unlike the majority of works from the literature that rely only on graph theory properties to conceive their encryption systems.

IV. Proposed Approach

The primary aim that drives the system put forward in this paper, is to propose a robust variant of the encryption scheme proposed in [17] while maintaining the performance levels. The main concept on which our approach is based is inspired by the divide-and-conquer design method, which consists in dividing an initial problem into sub-problems and then addressing every component of the resulting subset independently. The final solution of the initial problem is then deduced from the solutions found to the sub-problems.

The system described in this work has been designed in such a way that it takes into account the complexity of

the processing that the plaintext messages are subjected to during their encryption process. Indeed, this is the objective of the contribution in this paper, which is to improve the processing of the plaintext by making it more difficult and more complex than [17], using mainly all the Hamiltonian circuits that represent the plaintext.

The scheme proposed in [17] used a block of 25-characters length, which can be represented by 2 disjoint Hamiltonian circuits in a graph of order 13, given that a graph of order 13 contains 6 disjoint Hamiltonian circuits (Theorem 1). In contrast to [17], which used only 2 of the 6 circuits, the concept put forward in this approach makes use of all the disjoint Hamiltonian circuits of the graph (6 circuits), which allows the representation of blocks with 78-characters length in a single graph.

Theorem 1: In a complete graph with n vertices there are $(n-1)/2$ edge-disjoint Hamiltonian circuits, if n is an odd number strictly greater than 3 [22].

Considering a message which consist of 78 characters, the formula for splitting into blocks in [17] would be as follows: $78 = 25 \times 3 + 3$. This means that four blocks will be transformed into four adjacency matrices. The formula used in the proposed algorithm is limited to a single block, which in turn will be partitioned into six sub-blocks to form a single graph with six disjoint Hamiltonian circuits, thus forming a single adjacency matrix (FIG. 1 illustrates the difference between both systems).

Generally, the pre-encryption process of the plaintext message involves several steps: First the plaintext is converted into ASCII values and then divides into several blocks of size 78 (referring by $Block_i$). This operation uses the following formula: $n = 78k+r$, where n is the size of the plaintext, r ($r \in [0,77)$] the remainder of the division of n over 78) represents the remainder of the plaintext after block partitioning, and k is the number of blocks (refers to the quotient).

$$\begin{cases} k' = k & \text{If the division is exact.} \\ k' = k+1 & \text{otherwise.} \end{cases} \quad (3)$$

Where k' represents the total number of blocks resulting from the division. Each $Block_i$ is partitioned into 6 sub-blocks of size 13(each sub-block is represented by $subBlock_{ij}$), which are then converted into Hamiltonian circuits where the weights of the edges of the graph G_i are represented by the ASCII values of the characters that compose them. Finally, the resulting graph is converted into an adjacency matrix M_i.

The main process involving in our proposed system are presented in the following:

A. Key Generation / Re-Generation Algorithms

The generation of the sub-key K_i occurs in four steps. The first involves the random selection of a character $Char$ from the $Block_i$. The second step consists in using the position corresponding to the ASCII value of $Char$ in two ways, to construct the vector of positions VP that is necessary for the decryption, as well as to recover the value N located in the same position in the master key KEK (of size 256), which will be used as the seed of the BBS generator. The third step

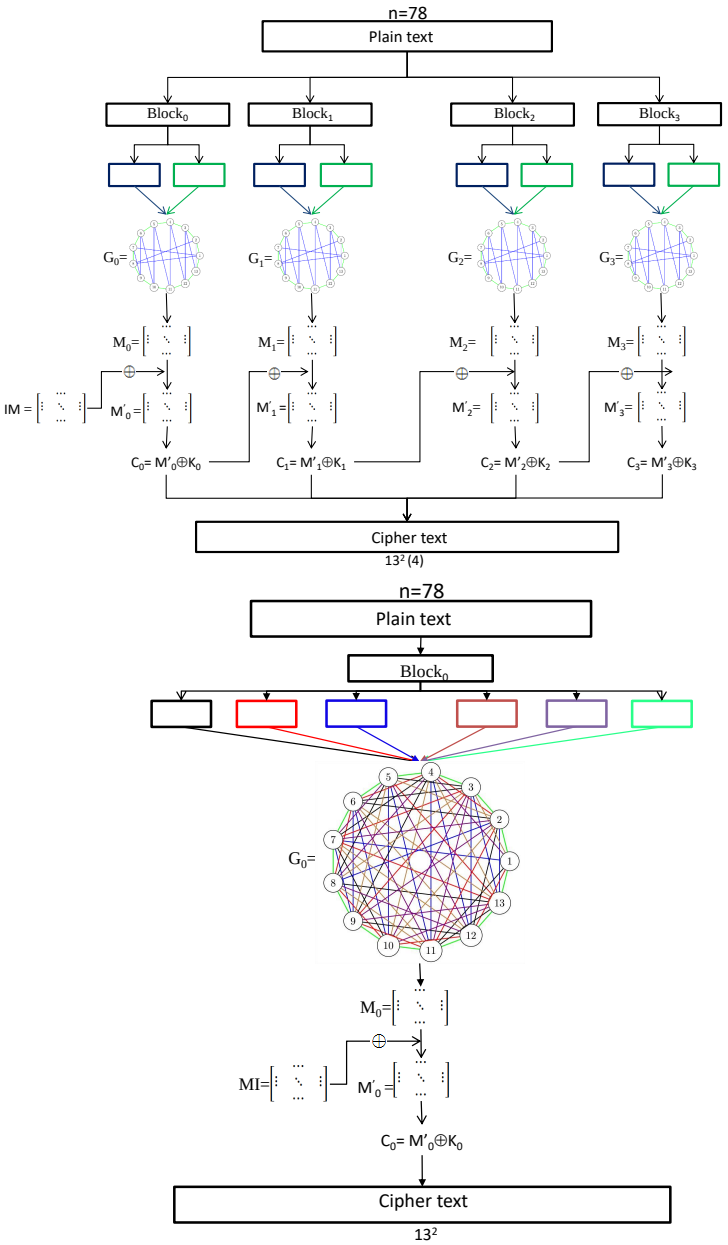


Fig. 1. Comparaison between the Encryption Process in [17] and the Proposed One.

in the process allows the generation of a vector S_i of size 13 from N using BBS generator. The fourth and final step uses the resulting S_i to generate the sub-keys K_i as a square matrix of order 13. The all sub-keys K_i(i = 0,. . ., k'-1) that are generated constitute the set $SK_{k'}$. This process is illustrated in FIG. 2.

The regeneration of K_i during the decryption process begins with the use of VP to recreate a key Key of size $13^2k'$ from KEK. Key is then divided into sub-vectors S_i of size 13 which are subsequently used to generate the sub-keys K_i as square matrices of order 13. This process is described in FIG. 3.

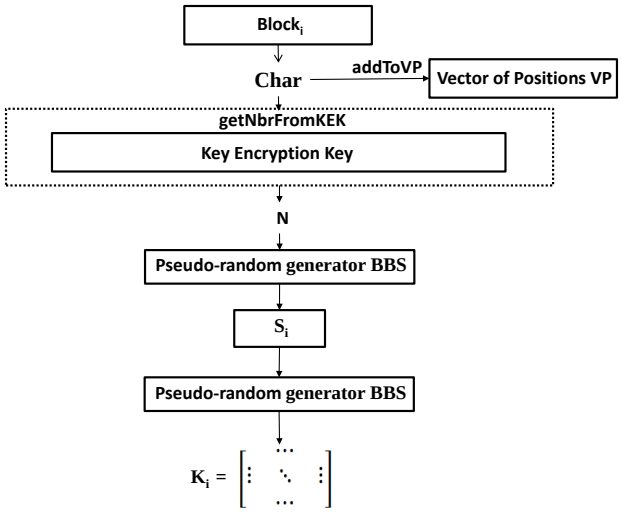


Fig. 2. Sub-keys Generator in Encryption Process.

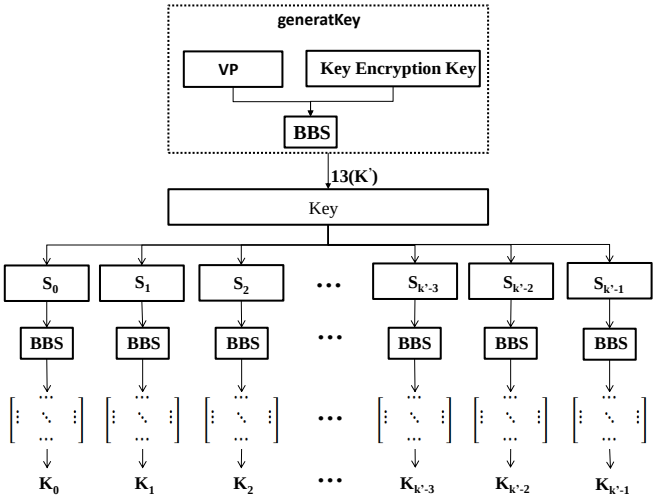


Fig. 3. Sub-keys Generator in Decryption Process.

Algorithm 1: Sub-keys generator in encryption process (GenerateSubKeys)

input : Clear message of n characters CMC_n , master key KEK, k' number of blocks
output: sub-keys $SK_{k'}$, the vector VP

begin
  /* Converts each character of the message into its ASCII value. */
  $CMA_n \leftarrow$ ***convertMessage***(CMC_n, n);
  /* Splits the message into k' $Block_i$ forming the set $BlockSet_{k'}$, where $BlockSet_{k'} = \{Block_0, \ldots, Block_{k'-1}\}$. */
  $BlockSet_{k'} \leftarrow$ ***parseMessage***(CMA_n, k');
  for *element* $Block_i$ *of the set* $BlockSet_{k'}$ **do**
    /* Randomly selects a character $Char$ from $Block_i$. */
    $Char \leftarrow$ **getCharFromBlock**$(Block_i)$;
    /* Feeds the vector VP with the ASCII value of the character $Char$. */
    $VP \leftarrow$ **addToVP**$(Char)$;
    /* Returns the content in the position p of the master key KEK, where p represents the ASCI code of the character concerned. */
    $N \leftarrow$ **getNbrFromKEK**$(KEK, Char)$;
    /* Generates from the seed N a vector S_i of size 13. */
    $S_i \leftarrow$ **generateSeed**(N, BBS);
    /* Takes as input the vector S_i and returns the sub-key K_i as a square matrix of order 13. */
    $K_i \leftarrow$ **generateSubKey**(S_i, BBS);
    /* Feeds the set $SK_{k'}$ with the sub-key K_i. */
    $SK_i \leftarrow$ **putSubKey**(K_i);
  end
end

B. Encryption Process

The encryption process begins with the pre-encryption step described above. Cipher block chaining (CBC) is used as mode of operation in our approach. The chaining uses a feedback method, in the sense that the result of the encryption of the previous block C_{i-1} is reused for the purpose of encrypting the current block M_i. More specifically, an exclusive or (XOR) operation is applied between the current block M_i and the previous block of ciphertext C_{i-1} as shown below:

$$M'_i = C_{i-1} \oplus M_i \tag{4}$$

A second XOR operation is then performed between the result of operation (3) and the sub-key K_i generated by the pseudorandom generator to compute the cipher C_i of the current block:

$$C_i = M'_i \oplus K_i \tag{5}$$

Since the first block does not have an antecedent. We generate an random matrix referring to IM(initialization matrix) which allows to perform the XOR operation with M_0. Each encrypted block consequently depends not only on the corresponding plaintext block, but also on all the encrypted blocks that precede it. The rows of the matrix C_i are concatenated to form a vector $eBlock_i$ of size 13^2, representing each encrypted block.

The resulting vectors $eBlock_i$ (i = 0,. . ., k'-1) generated from all blocks are then concatenated to form a single vector EM of size $13^2k'$. The encryption process, as shown in FIG. 4, ends with the transmission of the encrypted message EM in addition to the vector VP that is related to the decryption process.

Algorithm 2: Sub-keys generator in decryption process (GenerateSubKeys)

input : master key KEK, the vector of positions VP
output: sub-keys $SK_{k'}$

begin
  /* Generates a key Key of size 13k' from the vector of positions VP and the master key KEK. */
  $Key \leftarrow$ ***generateKey***(KEK, VP);
  /* Divides the key Key into k' vectors S_i (i = 0,..., k'-1). */
  $S_{k'} \leftarrow$ ***parseKey***(Key);
  for *element* S_i *of the set* $S_{k'}$ **do**
    /* Takes as input the vector S_i and returns the sub-key K_i as a square matrix of order 13. */
    $K_i \leftarrow$ **generateSubKey**(S_i);
    /* Feeds the set $SK_{k'}$ with the sub-key K_i. */
    $SK_i \leftarrow$ **putSubKey**(K_i);
  end
end

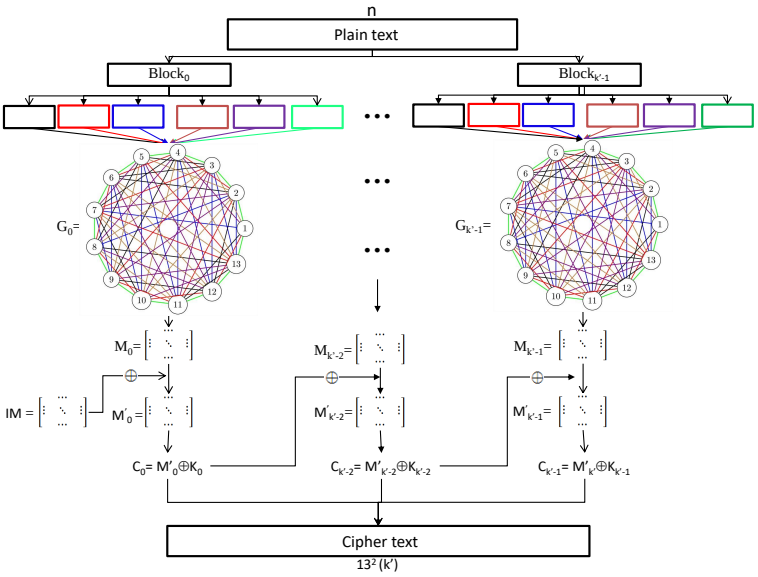


Fig. 4. Encryption Scheme.

Algorithm 3: Encryption Algorithm Using Disjoint Hamiltonian Circuits

input : Clear message of n characters CMC_n, master key KEK, Initialization Matrix IM of size 13
output: Encrypted message EM

begin
  $SK_{k'} \leftarrow$ ***GenerateSubKeys***(CMC_n, KEK, k');
  (IV-A).
  /* Converts each character of the message into its ASCII value. */
  $CMA_n \leftarrow$ ***convertMessage***(CMC_n);
  /* Splits the message into k' $Block_i$ forming the set $BlockSet_{k'}$, where $BlockSet_{k'} = \{Block_0, \ldots, Block_{k'-1}\}$. */
  $BlockSet_{k'} \leftarrow$ ***parseMessage***(CMC_n);
  for *element* $Block_i$ *of the set* $BlockSet_{k'}$ **do**
    /* Divides each block $Block_i$ into six sub-blocks $subBlock_{ij}$ (j = 0,...,5) of size 13, all forming the set $subBlockSet_i$, where $subBlockSet_i = \{subBlock_{i0}, \ldots, subBlock_{i5}\}$. */
    $subBlockSet_i \leftarrow$ **parseBlock**$(Block_i)$;
    /* Converts the sub-blocks into disjoint hamiltonian circuits in a graph G. */
    $G_i \leftarrow$ **blockToGraph**$(subBlockSet_i, 13)$;
    /* Transforms the graph G_i into an adjacency matrix M_i of order 13. */
    $M_i \leftarrow$ **graphToMatrix**(G_i);
    if *i=0* **then**
      $M'_0 \leftarrow IM \oplus M_0$;
    else
      $M'_i \leftarrow C_{i-1} \oplus M_i$;
    end
    $C_i \leftarrow M'_i \oplus SK_i$;
    /* Concatenates the rows of the matrix C_i to form the vector $eBlock_i$ of size 13^2. */
    $eBlock_i \leftarrow$ **transformMatixToVector**(C_i);
  end
  /* Forms a single vector EM of size $13^2 k'$ by concatenating the resulting vectors $eBlock_i$ (i = 0,..., k'-1). */
  $EM \leftarrow$ **concatenateEncryptionBlock**$(eBlock_{k'})$;
end

C. Decryption Process

In general, the process of decryption corresponds to the process of encryption performed in reverse order (ALGORITHM 4). In the decryption process described in this paper, the ciphertext EM refers to the input of the algorithm. EM is decomposed into k' vectors ($eBlock_i$) and then gathered to constitute the set $eBlockSet_k$. The $eBlock_i$(i = 0,. . ., k'-1) are subsequently transformed into a matrix C_i. The number of blocks k' is calculated as follows:

$$k' = m \div 13^2 \quad (6)$$

with m is the size of the ciphertext.

The sub-key generation algorithm presented in ALGORITHM 2 makes use of the provided vector VP to produce a key of size $13^2 k'$ from the master key KEK. Each block C_i(i = 0,. . ., k'-1) is decrypted using its own sub-key K_i using the following formula:

$$M_i = C_{i-1} \oplus M'_i \quad (7)$$

Where

$$M'_i = C_i \oplus K_i \tag{8}$$

and

$$M_0 = IM \oplus M'_0 \tag{9}$$

At this stage, the decrypted blocks M_i are transformed into a graph G_i and then into $Block_i$. Finally, the plaintext message is formed by concatenating the $Block_i$(i = 0,. . ., k'-1) as shown in FIG. 5.

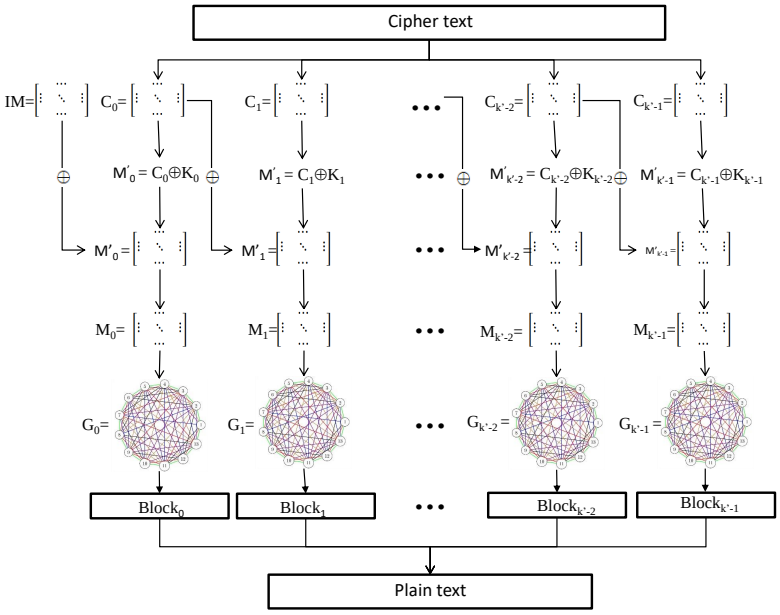


Fig. 5. Decryption Scheme.

Algorithm 4: Decryption Algorithm Using Disjoint Hamiltonian Circuits

input : Encrypted message EM, master key KEK, the vector of positions VP
output: Clear message of n characters CMC_n

begin
  $SK_{k'} \leftarrow$ ***GenerateSubKeys***(KEK, VP);
  (IV-A).
  /* Divides the encrypted message into k' $eBlock_i$ wich are then concatenate to form a set $eBlockSet_{k'}$. */
  $eBlockSet_{k'} \leftarrow$ ***parseEncryptedMessage***(EM);
  for *element* $eBlock_i$ *of the set* $eBlockSet_{k'}$ **do**
    /* Form the matrix C_i of order 13 from the vector $eBlock_i$. */
    $C_i \leftarrow$ **transformVectorToMatrix**$(eBlock_i)$
    $M'_i \leftarrow C_i \oplus K_i$;
    if $i = 0$ **then**
      $M_0 \leftarrow IM \oplus M'_0$;
    else
      $M_i \leftarrow C_{i-1} \oplus M'_i$;
    end
    /* Transforms the adjacency matrix M_i into a graph G_i. */
    $G_i \leftarrow$ **matrixToGraph**(M_i);
    /* Returns the $Block_i$ represented by the disjoint hamiltonian circuits inside the graph G_i. */
    $Block_i \leftarrow$ **graphToBlock**(G_i);
  end
  /* Forms a single block that forms the plaintext message by concatenating the resulting blocks $Block_i$ (i = 0,..., k'-1). */
  $CMC_n \leftarrow$ **concatenateBlock**$(Block_{k'})$;
end

V. Security Analysis and Experimental Results

The evaluation of the encryption system addressed the reliability of the suggested algorithm. For this purpose, we study the system reaction in terms of performance and security according to fundamental criteria. For which we also perform different statistical tests. This evaluation is described in the following sections.

A. Confusion and Diffusion Tests

Diffusion and confusion are very important as aspects of the functioning of a secure encryption which were first identified in 1949 by Claude Elwood Shannon [23]. In his original definitions:
Confusion means making the relationship between key and ciphertext as complicated and as involved as is feasible, whereas in this case refers to the property that redundancy in the plaintext's statistics is "dissipated" in the ciphertext's statistics.
Diffusion is related to the reliance of the output bits upon the input bits. In a cipher with proper diffusion, the changing of an input bit is expected to change every output bit with a probability of half (this is referred to as the strict avalanche criterion). Accordingly, the used equation (10) is:

$$bits_{diff} = (1 \div (13^2 \times 16) \times w(C \oplus C') \tag{10}$$

$$= (1 \div (2704) \times w(C \oplus C') \tag{11}$$

Where w is the hamming weight, C and C' are respectively the original and modified inputs, and the value 16 refers to the number of bits representing each element in the cipher.

B. Plaintext Sensitivity Test

The diffusion property is intended to produce an avalanche effect [24] between the plaintext and the encrypted messages. The sensitivity test of the bit change in the plaintext is used to verify the diffusion property of a particular algorithm.
Given pairs of plaintext and secret keys, we generate the ciphertext corresponding to each pair (plaintext, secret key) through our cryptosystem, changing one or more bits (Knowing that a change at character level implies a change of bit) in the randomly generated plaintext, and by retaining the key unchanged.
Subsequently, we calculate the average of the percentage of

bit difference by the equation (10) as illustrated in the FIG. 6. Over 50% of the bits in the cipher text are changed. We can clearly see that the average of the percentages of bit difference is between 48.16% and 51% for our encryption system and between 47.1% and 50.80% for AES-128. These percentages demonstrate that our encryption system offers a good diffusion compared to AES-128.

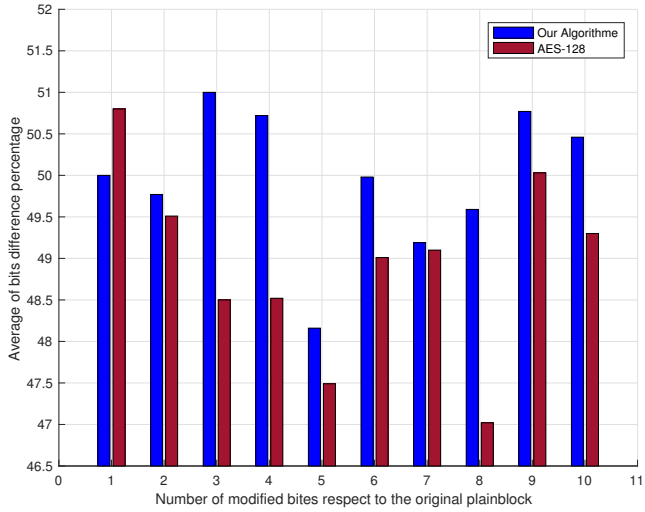


Fig. 6. Number of Modified Bits with respect to the Original Plaintext.

C. Key Sensitivity Test

The confusion property establishes a relation between the key and the ciphertext. The key sensitivity test ensures this property. Indeed, we consider a set of pairs of plaintext and secret keys. Each pair is encrypted by applying the proposed cryptosystem. Then, we modify one or more bits in the different randomly generated keys while the clear text still fixed. Afterward we calculate the average of the percentage of bit difference by applying the equation (10).

FIG. 7 represents the results obtained by using our encryption system and our generator to produce the encryption keys. We can notice that more than 50% of the bits are modified. Specifically, the average percentage of bit difference is between 49.64% and 50.79% for our encryption system and between 48.25% and 50.73% for AES-128. Thus, according to the experimental results, it can be said that the key generation via our algorithm is more robust than AES-128.

D. Statistical Tests

In order to study the quality of the random generation of the suggested encryption block cipher, we apply the well-known DIEHARD test [25]. The primary objective of this test is to demonstrate whether our cryptosystem is able to withstand statistical attacks. In other terms, the output of a secure block cipher must be indistinguishable statistically from a random output using the encryption function. To perform this test, a randomly generated cipher sequence is initially binarized to generate a bitstream of over 10 MB. Thereafter, the bitstream is analyzed statistically by putting it under the

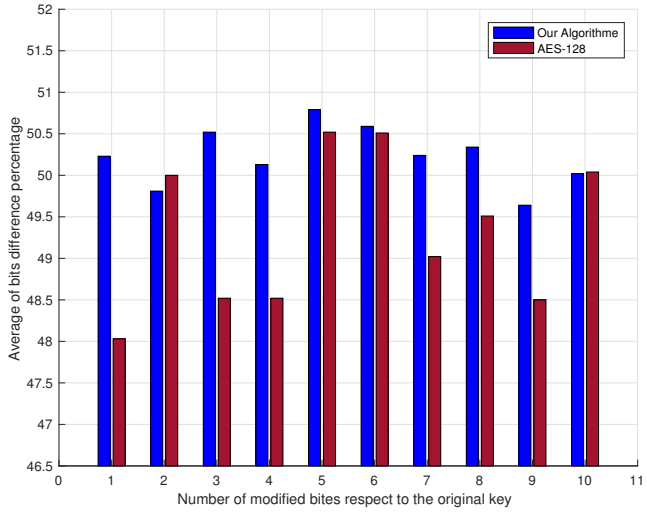


Fig. 7. Number of Modified Bits with respect to the Original Key.

DIEHARD tests. The DIEHARD tests check the p-value of the values generated randomly, with the p-value in the range [0.025, 0.975]. The average values found are summed up in TABLE I. The results indicate that the bitstream generated with our encryption system successfully passed all DIEHARD tests. Moreover, our encryption system shows a satisfactory and statistically indistinguishable random behavior.

E. Brute Force Attack

Brute force attacks are a mean to get all possible key arrangements with a fast prediction tool. Assuming that a high-performance machine that spends 10^{-10} seconds on testing the validity of every key is used, and that the numbers entailed in the master key range from 1 to 1000.
Given that algorithm has 1000^{256} potential keys. A brute force attack would require about $10^{-10} \times 1000^{256}$ seconds to find the appropriate key. Therefore, a brute force attack with an exhaustive search over the key possibility space is not feasible in a reasonable amount of time.

To find a 78-character message when a single block is used, it normally takes 1000 possibilities to find one of the master key numbers, which will represent the seed of the BBS generator involved in producing the S_0 vector. However, the prime numbers used as the input parameters for the generator are not easily determined (due to factorization problems). As a result, it is nearly impossible to figure out the sub-key if the pq product is large enough.

F. Time Analysis

Table II represents the performance test of our cryptosystem, compared to other known block ciphers such as triple DES [26] and AES [27] in terms of their CPU time consumption. The computations are run on a computer with on an Intel Core i7-6600U processor, 64-bit OS, 2.81 GHz with 20 GB of RAM. It can be seen from TABLE II that our algorithm can achieve good results in terms of run time over the other standard encryption systems.

TABLE I. DIEHARD TEST

Test Name	P-value	Interpretation
diehard bitstream	0.59537390	
diehard squeeze	0.97442749	
diehard sums	0.11133210	
diehard count 1s str	0.60934773	
diehard count 1s byt	0.78478421	
diehard parking lot	0.55915630	
diehard birthdays	0.03222200	
diehard operm5	0.75636037	
diehard oqso	0.33566335	
diehard dna	0.45051943	
diehard 2dsphere	0.53656799	
diehard 3dsphere	0.62980562	
diehard rank 32x32	0.40775458	
diehard rank 6x8	0.45554634	
diehard opso	0.44037399	
diehard runs	0.86351847	PASSED
diehard craps	0.15275419	
rgb bitdist	0.69014502	
rgb minimum distance	0.57113046	
rgb permutations	0.60422228	
rgb lagged sum	0.60927830	
rgb kstest test	0.26054914	
dab bytedistrib	0.68169231	
dab dct	0.25149694	
dab filltree	0.88848873	
dab filltree2	0.29185197	
dab monobit2	0.74899931	
sts monobit	0.68441660	
sts runs	0.37246909	
sts serial	0.50145101	
marsaglia tsang gcd	0.47467308	

TABLE II. ENCRYPTION TIME COMPARISON BETWEEN OUR BLOCK CIPHER AND OTHERS BLOCK CIPHERS USING DIFFERENT MESSAGE SIZE

Message Size (Kilo Byte)	AES (ms)	3DES (ms)	Our encryption algorithm
3	248.07	247.47	4.9
10	951.2	614.9	10.4
20	1972	1096	21.2

VI. CONCLUSION AND FUTURE WORK

The work presents a new cryptosystem that takes advantage of the principles of graph theory, which enable a high degree of security while maintaining the performance of data processing. Our proposed encryption block cipher using in particular the disjoint Hamiltonian circuits that have been adopted to represent the plaintext in a pre-encryption phase. the process makes use of a specific sub-key generator that has been set up to generate the encryption keys according to the requirements of the proposed system. We have performed different statistical tests, specifically the DIEHARD, confusion and diffusion tests to prove the security and performance of our cryptosystem. The experiments results proved the good behaviour of our proposed design in terms of robustness and CPU time compared to 3DES and AES. In a future work, we intend to use another pseudo-random generator, such as the one proposed in [28] known as PSOCA, which is mainly based on cellular automata, and we also investigate other properties of graph theory for a more discriminating and robust representation of the data.

REFERENCES

[1] A. J. Menezes, J. Katz, P. C. Van Oorschot, and S. A. Vanstone, *Handbook of applied cryptography*. CRC press, 1996.

[2] P. FIPS, "81, des modes of operation," *Issued December*, vol. 2, p. 63, 1980.

[3] V. Rijmen and J. Daemen, "Advanced encryption standard," *Proceedings of Federal Information Processing Standards Publications, National Institute of Standards and Technology*, pp. 19–22, 2001.

[4] W. Meier, "On the security of the idea block cipher," in *Workshop on the Theory and Application of of Cryptographic Techniques*. Springer, 1993, pp. 371–385.

[5] N. P. Smart, "The "naive" rsa algorithm," in *Cryptography Made Simple*. Springer, 2016, pp. 295–311.

[6] ——, "Public key encryption and signature algorithms," in *Cryptography Made Simple*. Springer, 2016, pp. 313–347.

[7] A. J. Menezes, J. Katz, P. C. Van Oorschot, and S. A. Vanstone, *Handbook of applied cryptography*. CRC press, 1996.

[8] G. Alexanderson, "About the cover: Euler and königsberg's bridges: A historical view," *Bulletin of the american mathematical society*, vol. 43, no. 4, pp. 567–573, 2006.

[9] L. Blum, M. Blum, and M. Shub, "A simple unpredictable pseudo-random number generator," *SIAM Journal on computing*, vol. 15, no. 2, pp. 364–383, 1986.

[10] P. Amudha, A. C. Sagayaraj, and A. S. Sheela, "An application of graph theory in cryptography," *International Journal of Pure and Applied Mathematics*, vol. 119, no. 13, pp. 375–383, 2018.

[11] S. G. Akl, "The graph is the message: design and analysis of an unconventional cryptographic function," in *From Parallel to Emergent Computing*. CRC Press, 2019, pp. 425–442.

[12] K. D. Rangaswamy and M. Gurusamy, "Application of graph theory concepts in computer networks and its suitability for the resource provisioning issues in cloud computing-a review." *J. Comput. Sci.*, vol. 14, no. 2, pp. 163–172, 2018.

[13] D. Sensarma and S. S. Sarma, "Application of graphs in security," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 10, pp. 2273–2279, 2019.

[14] S. H. Hashem, "Proposal hybrid cbc encryption system to protect e-mail messages," *Iraqi Journal of Science*, vol. 60, no. 2, pp. 157–170, 2019.

[15] A. Yousif and A. H. Kashmar, "Key generator to encryption images based on chaotic maps," *Iraqi Journal of Science*, vol. 60, no. 2, pp. 362–370, 2019.

[16] T. A. Khaleel and A. A. Al-Shumam, "A study of graph theory applications in it security," *Iraqi Journal of Science*, vol. 61, no. 10, pp. 2705–2714, 2020.

[17] K. Bekkaoui, S. Ziti, and F. Omary, "A robust scheme to improving security of data using graph theory," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 5, 2020.

[18] C. Beaula and P. Venugopal, "Cryptosystem using double vertex graph," *Indian Journal of Science and Technology*, vol. 13, no. 44, pp. 4483–4489, 2020.

[19] P. Perera and G. Wijesiri, "Encryption and decryption algorithms in symmetric key cryptography using graph theory," *Psychology and Education Journal*, vol. 58, no. 1, pp. 3420–3427, 2021.

[20] S. G. Akl, "How to encrypt a graph," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 35, no. 6, pp. 668–681, 2020.

[21] P. Venugopal, "Encryption using double vertex graph and matrices," *Solid State Technology*, vol. 64, no. 2, pp. 2486–2493, 2021.

[22] N. Deo, *Graph theory with applications to engineering and computer science*. Courier Dover Publications, 2017.

[23] C. E. Shannon, "Communication theory of secrecy systems," *The Bell system technical journal*, vol. 28, no. 4, pp. 656–715, 1949.

[24] J. C. H. Castro, J. M. Sierra, A. Seznec, A. Izquierdo, and A. Ribagorda, "The strict avalanche criterion randomness test," *Mathematics and Computers in Simulation*, vol. 68, no. 1, pp. 1–7, 2005.

[25] G. Marsaglia, "Diehard test suite," *Online: http://www. stat. fsu. edu/pub/diehard*, vol. 8, no. 01, p. 2014, 1998.

[26] D. Coppersmith, D. B. Johnson, and S. M. Matyas, "A proposed mode for triple-des encryption," *IBM Journal of Research and Development*, vol. 40, no. 2, pp. 253–262, 1996.

[27] N. P. Smart and N. P. Smart, *Cryptography made simple*. Springer, 2016.

[28] C. Hanin, F. Omary, B. Boulahiat, and S. Elbernoussi, "Design of new pseudo-random number generator based on non-uniform cellular automata," *International Journal of Security and Its Applications*, vol. 10, no. 11, pp. 109–118, 2016.

A New Algorithm to Reduce Peak to Average Power Ratio in OFDM Systems based on BCH Codes

Brahim BAKKAS[1], Reda Benkhouya[2], Toufik Chaayra[3], Chana Idriss[4], Hussain Ben-Azza[5]
Ecole Nationale d'Art et Méties (ENSAM-Meknès), Moulay Ismail University, Meknès 50050, Morocco[1,5]
MISC-Faculty of Sciences, Ibn Tofail University, Kenitra 14000, Morocco[2]
Department of Mathematics and Computer, Multidisciplinary Faculty of Nador, Mohammed I University, Morocco[3]
High School of Technology, Moulay Ismail University, Meknès 50050, Morocco[4]

***Abstract*—Orthogonal Frequency Division Multiplexing (OFDM) has a great peak-to-average power ratio (PAPR). This will reduce the performance of the power amplifier (PA). Therefore, PAPR deteriorates the overall energy efficiency of an OFDM system. Peak Insertion (PI) is one of the most commonly used methods to reduce PAPR, it gives the best reduction in PAPR. Therefore, it causes a strong degradation in Bit Error Rate (BER). To solve this problem, we propose a new algorithm called BCB-OFDM based on Bose Chaudhuri Hocquenghem Codes (BCHs) and PI. BCB is implemented in OFDM system with Quadrature Amplitude Modulation (QAM) and two coding rates 1/2 and 1/4 over an Additive White Gaussian Noise (AWGN) channel. Simulation results show that the BCB is very interesting and achieve a good value in terms of PAPR reduction with keeping good performance compared with PI and normal OFDM. In addition, BCB algorithm is simple, robust, and leaves no requirement side information with more flexibility to choose between PAPR reduction and BER performances.**

Keywords*—Orthogonal Frequency Division Multiplexing (OFDM); Peak to Average Power Ratio (PAPR); Bit Error Rate (BER); Peak Insertion (PI); Coding; Bose Chaudhuri Hocquenghem (BCH)

I. Introduction

Orthogonal Frequency Division Multiplex (OFDM) is a multi-carrier technique that has shown its effectiveness, robust against interference problems caused by multipath. Thanks to its simplicity of implementation through the use of Fourier Transform. OFDM is the basic technology used in wireless communication such as WiFi, WiMAX [1], [2], [3], 4G [4] and 5G [5], [6], [7], [8]. However, it has a major drawback caused by the high value of Peak To Average Power Ratio (PAPR) defined as the ratio between the maximum power and the average power of an OFDM signal. The high value of PAPR value forces the Power Amplifier (PA) to work in the non-linear region and cause a degradation of the signal and need a large consumption in energy [9], [10].

Several approaches to resolving this problem have been offered, the most widely utilized approaches are: Selective Mapping Technique (SLM)[11], Partial Transmit Sequences (PTS) [12], [13], companding [14], [3], Clipping [15], Palm Clipping [16], Peak Insertion (PI)[17] and Tone reservation [18]. In [19], [20], the authors compare some PAPR reduction methods in terms of PAPR and BER. The obtained result in this comparison is that the Peak Insertion technique proposed gives the highest PAPR reduction, but it causes a high fort degradation [21]. Also, linear codes give the best result in terms of BER with small PAPR reduction by finding the best code with a diminished PAPR however it produces a computational complexity to find the best code word with a low PAPR value [22], [23], [24]. As a result, to choosing the relevant PAPR reduction approach, a compromise between PAPR reduction and BER performances must be made.

The main contributions of this article are: (1) Proposed a new algorithm to reduce PAPR in OFDM Systems called BCB algorithm based on BCH codes and Peak Insertion. (2) Using BCH codes to explore the Peak Insertion to improve its signal degradation. (3) BCB method does not require any site information from the transmitter.

The remainder of this paper is organized as follows: The System model and PAPR definition and related work about Peak Insertion are presented in Section II. The proposed method and algorithm, its principle and parameters are detailed in Section III. In Section IV, we provide and discuss the simulation results in terms of PAPR reduction and BER performance compared with the normal OFDM and PI method. The conclusion and future works are presented in Section V.

II. OFDM System and Related Work

In this section, we present the OFDM system model and related work about peak insertion method.

A. OFDM System Model and PAPR Definition

The principle of the OFDM system is composed by two parts as presented in Fig. 1. At the transmitter side, the input data are mapped by one of the modulation schemes. The obtained signal is modeled by using the Inverse Fast Fourier Transformation (IFFT) algorithm with N-points before being transferred to the channel. At the receiver side, the process inverse is performed. The discrete OFDM signal in the time-domain is defined by:

$$x(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k * e^{\left(\frac{j2\pi nk}{N}\right)} \quad , \quad 0 \leq n \leq N-1. \quad (1)$$

Where N is the number of subcarriers, and X_k is the kth subcarrier of the same OFDM symbol. The PAPR is defined as the ratio of maximal power and the average power :

$$PAPR = \frac{\max\left(|x(n)|^2\right)}{E\{|x(n)|^2\}}, \quad 0 \leq n \leq N-1, \quad (2)$$

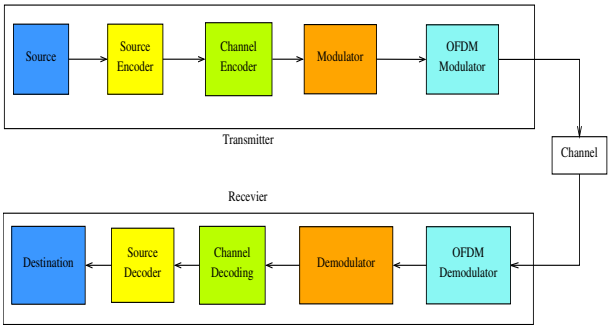


Fig. 1. Communication Digital Systems Block with OFDM.

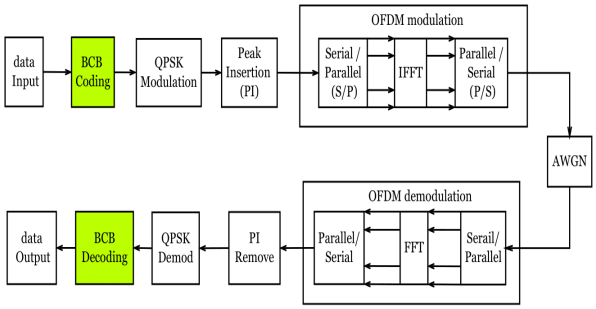


Fig. 2. BCB Peak Insertion in OFDM Systems.

where $E\{.\}$ represents the expectation value. PAPR performance is represented by the Complementary Cumulative Distribution Function (CCDF), which is the probability of PAPR exceeding a threshold, defined as follows:

$$CCDF(\gamma) = \Pr[PAPR \geq \gamma], \tag{3}$$

where γ is a threshold value.

B. Peak Insertion Method

Peak Insertion (PI) method proposed in [17] is implemented by adding an impulse α to the original OFDM symbol in the frequency domain to reduce the PAPR value. The resulting signal Y_k of a signal X_k is given by:

$$Y_k = X_k + \delta(k - k_\alpha), \tag{4}$$

where X_k is the original OFDM symbol, α is a positive real number representing the strength of the inserted peak, k_α is the position of the inserted peak, N is the number of subcarriers [17]. The peak α is inserted to increase the PAPR of the signal in the frequency domain. However, after PI the mean square value of this signal will increase, but with a greater mean square value [19]. The power of this signal is then controlled by scaling it by a real β such that its mean square value is reduced to a suitable level without affecting the PAPR reduction [17]. At the receiver, To restore the original OFDM, reverse operations are carried out. PI method has a benefit in PAPR reduction with a gain of 8dB [23], [17]. However, it presents the drawback of signal degradation, in this area we present a new algorithm to reduce PAPR with good performances in terms of BER .

III. Proposed Method BCB PAPR

In this Section, we presents the principle of the proposed method based on BCH codes and PI technique. The BCH codes form a large class of error correcting cyclic codes, This class of codes is used for purpose of multiple-error corrections. For any positive integers m with $m \geq 3$, and, t with $t < 2^{m-1}$, there exists a binary BCH code with the following parameters:

n : Block length $n = 2^m - 1$
k : BCH Code length $k = 2^m - m - 1$
$(n-k)$: Number of parity-check digits $(n-k) \geq mt$
d : Minimum distance $d \geq 2t + 1$

We call this code a t-error-correcting BCH code, and we note in this paper the $BCH(n,k)$ a BCH code.

To improve the Peak Insertion technique, we propose a new method combined with BCH codes noted in this paper BCB-OFDM reduction. Let $BCB(n,k,v,\alpha)$ be the new block coding scheme with peak insertion α where:

n, k : Block length and code length respectively
v : Number of bits information
α : Peak insertion value
$(k-v)$: length of zeros padding

Let $m^{(i)} = b_1^{(i)}, b_2^{(i)}, \ldots, b_v^{(i)}$ is the i^{th} block bits of size v, where $i = 1, 2 \ldots, nbits$, and $nbits = log_2(M)$ number of bits/symbols. The $m^{(i)}$ message with length v bits is padded with $(k-v)$ to have a block with k bits noted by $B^{(i)}$ where k is the BCH code length. The block $B^{(i)}$ is now represented as bellow:

$$B^{(i)} = b_1^{(i)}, \ldots, b_v^{(i)}, \overbrace{0, \ldots, 0}^{(k-v)} \tag{5}$$

We apply the BCH(n,k) encoding, the encoded message $C^{(i)}$ of $B^{(i)}$ is given as fellow:

$$C^{(i)} = b_1^{(i)}, \ldots, b_v^{(i)}, \overbrace{0, \ldots, 0}^{(k-v)}, c_{n-k+1}^{(i)}, \ldots, c_n^{(i)} \tag{6}$$

With 4-QAM modulation we have 2 bits for each symbol, then the coded message C is now construct respectively from two blocks $C^{(1)}$ and $C^{(2)}$ by flipping the zeros block of each $C^{(i)}$ to the middle of the codded message C as follows:

$$\begin{aligned} C = b_1^{(1)}, \ldots, b_v^{(1)}, c_{n-k+1}^{(1)}, \ldots, c_n^{(1)}, \overbrace{0, \ldots, 0}^{2(k-v)+1}, \\ b_1^{(2)}, \ldots, b_v^{(2)}, c_{n-k+1}^{(2)}, \ldots, c_n^{(2)} \end{aligned} \tag{7}$$

The coded message C is now mapped by 4QAM modulation, $X = [X_0, X_1, \ldots, X_{N-1}]$, a peak real α is inserted before using the N-points IFFT algorithm to obtained the OFDM message $x = IFFT(X^T)$ where $(.)^T$ is the complex transpose. Then the signal OFDM is passed through an AWGN channel. At the received part, the inverses process is applied. Perform Serial to Parallel conversion, then the N-point FFT is applied. The peak α is removed from the signal before applying 4QAM demodulation. The decoding algorithm is applied and the bits informations is recovered and compared

with the original bits to compute the BER for each SNR as shown in Fig. 2. The main steps of the proposed method have been widely described in the following Algorithm 1.

Algorithm 1 The proposed BCB method in OFDM systems

Require: $Nsym, NFFT, M, n, k, v, \alpha, E_bN_0et$
Ensure: $PAPR, BERSet$
  $n \leftarrow log_2(NFFT)$
  $nbits \leftarrow log_2(M)$
  for each snr $\in E_bN_0set$ **do**
    for iter=1 **to** Nsym **do**
      $B \leftarrow randint(nbits, v)$
      $C \leftarrow BCBcod(B, n, k, v)$
      $X \leftarrow qam(C, M)$
      $Y \leftarrow PIadd(X, \alpha)$
      $x \leftarrow OFDMMod(Y, NFFT)$
      $PAPR \leftarrow CCDF(x)$
      $y \leftarrow AWGN(x, snr)$
      $\hat{Y} \leftarrow OFDMDemod(y, NFFT)$
      $\hat{X} \leftarrow PIremove(\hat{Y}, \alpha)$
      $\hat{C} \leftarrow qamdem(\hat{X}, M)$
      $\hat{B} \leftarrow BCBdecod(\hat{C}, n, k, v)$
    end for
    $BERSet(iter) \leftarrow BERcompute(B, \hat{B})$
  end for

IV. Simulation Results and Discussion

The performance of the proposed algorithm is investigated in terms of PAPR reduction and BER performances. We start this section by examining the impact of α and coding rates (1/4 and 1/2) respectively. The results are represented for each case. Finally, we compare the proposed algorithm with the normal OFDM and the Peak Insertion. The simulations are carried out with 4QAM (QPSK) as modulation schemes with N-points $NFFT$ size equal to 256 over an AWGN channel. A number $Nsym$ of OFDM symbols are generate randomly, we note that in this system we have neglected the other parameters of the OFDM systems such as cyclic prefix and guard interval. Table I detailed the parameters of simulation, and Table II regroups the possible value of BCH coding, we choose two values k equal to 207 and 147 (respectively with low and high capacity errors correcting).

TABLE I. Simulation Parameters of the Proposed Algorithm BCB in OFDM Systems

Parameter	values
Number of symbols (Nsym)	10000
Number of bit errors (Nberr)	100
FFT size (NFFT)	256
Peak insertion α	0, 29, 50 and 80
Peak insertion β	1
BCH (n,k,t)	n=255, k=207 and 147 (see Table II)
Message Rate (h)	1/2 and 1/4
Modulation scheme	4-QAM (QPSK)
Channel model	AWGN

A. PAPR Performances of the Proposed BCB

In this subsection, we fix some parameters such as FFT size $NFFT = 256$, information size (message) $v = 64$ and $v = 128$. we choose two value of coding k equal to 147 and 207. Also, we varying the value of the peak α to 0, 29, 50 and 80 as mentioned in Table I. The results are depicted by using the CCDF curves for each cases. Firstly, we set the rate to 1/4, the results are presented with the CCDF curves with variable α in Fig. 3 and Fig. 4. Secondly we change the rate to 1/2 the results are showed in Fig. 5 and Fig. 6.

TABLE II. Possible values of $BCH(n = 255, k, t)$ codes

k	t	k	t	k	t	k	t
247	1	187	9	123	19	63	30
239	2	179	10	115	21	55	31
231	3	171	11	107	22	47	42
223	4	163	12	99	23	45	43
215	5	155	13	91	25	37	45
207	**6**	**147**	**14**	87	26	29	47
199	7	139	15	79	27	21	55
191	8	131	18	71	29	13	59

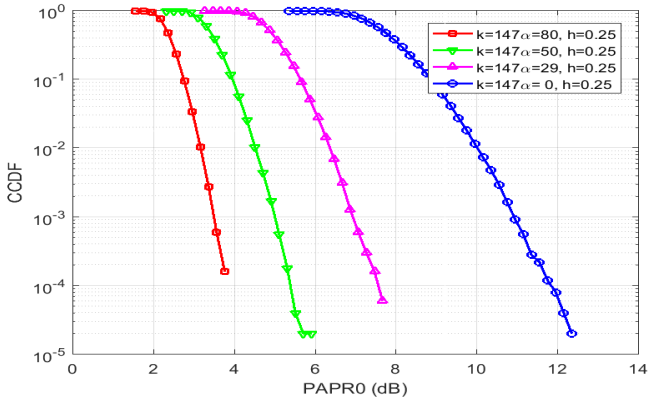


Fig. 3. CCDF Curves of the Proposed BCB-OFDM NFFT=256, 4QAM, $BCB(n = 255, k = 147, v = 64, \alpha)$ with Variable α.

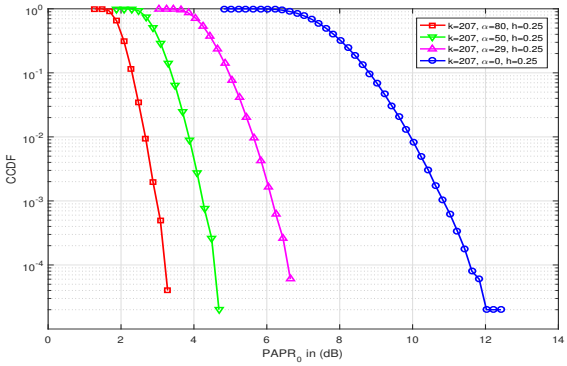


Fig. 4. CCDF Curves of the Proposed BCB-OFDM NFFT=256, 4QAM, $BCB(n = 255, k = 207, v = 64, \alpha)$ with Variable α.

In the first case, $BCB(n = 255, k = 147, v = 64, \alpha)$, we set k to 147, v to 64 and the number of zero padding is $2 * (k - v)$ with variable α. The result are depicted in terms of PAPR in Fig. 3. From the obtained results a significant reduction in terms of PAPR is observed for α is equal to 29 where the 8dB is attend. Also, we achieves 6dB when α is great then 29.

Fig. 4 shows the CCDF with variable α of the second case

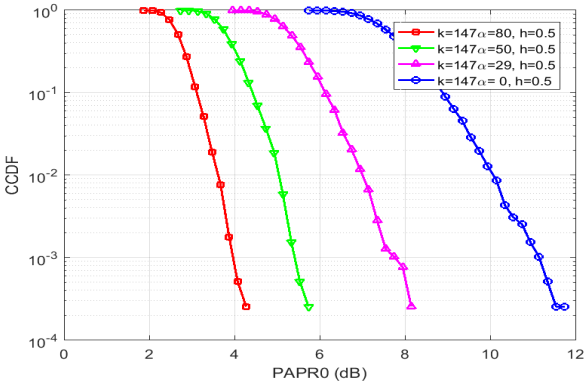


Fig. 5. CCDF Curves of the Proposed BCB-OFDM NFFT=256, 4QAM, $BCB(n = 255, k = 147, v = 128, \alpha)$ with Variable α.

$BCB(n = 255, k = 207, v = 64, \alpha)$. In this case we set k to 207 and we keep other parameters $NFFT = 256$, $v = 64$. From this figure we note that proposed method achieve a good reduction even we augment the peak α. The value of PAPR is reduce to half (more then 6dB) when α is great then 50, the gain is very important and it is between $7dB$ and $9dB$.

Now, we repeat the same simulation realized in the first and second cases by changing the message size from $v = 64$ to $v = 128$ with a rate equal to $1/2$ and two values of BCH codes length k.

In the third case, $BCB(n = 255, k = 147, v = 128, \alpha)$ the value coding length k is set to 147 and the number of zero padding is 2*(k-v) with variable α. The result are depicted in Fig. 5 in terms of PAPR with variable α. As shown in this figure, the, one can clearly see that the PAPR decreases when α increases, a value of 4.2dB achieved at CCFD=10^{-4} when $\alpha = 80$.

In the forth case, $BCB(n = 255, k = 207, v = 128, \alpha)$ we set the coding length k to 207 and the number of zero padding is $2 * (k - v)$. with variable α. The CCDF curves of the proposed algorithm is showed in Fig. 6. In This figure, we notice that the PAPR values reduced when the α increases. The value 3.8dB is achieved at CCDF=$10^{-}4$ when α equal to 80.

B. Comparative Studies

To validate the proposed method in terms of PAPR and BER performances, we compare the BCB method with Peak Insertion and normal OFDM by varying the code rate and value of the peak α. Firstly, we depict the CCDF and BER curves with variable α, coding rate, and information size, the results are depicted in Fig. 7 for each value of α. Finally, the gain in terms of PAPR at CCDF=10^{-4} versus variable α are summarized in Table III, while Table IV regroups the SNR loss in dB calculated in eq. (8) with variable α, coding length k and information size v.

$$SNR_{loss} = SNR_{method} - SNR_{Normal} \quad (8)$$

According to the previous results depicted in Fig. 7, and summarize results in Tables IV and III. The value of PAPR

TABLE III. PAPR GAIN AT CCDF=10^{-4} OF THE PROPOSED METHOD, PEAK INSERTION AND NORMAL OFDM

Method	$\alpha = 0$	$\alpha = 29$	$\alpha = 50$	$\alpha = 80$
OFDM		0		
PI(α)	0	6.4	8.2	9.3
BCB (255,147,64,α)	0	4.2	6.2	7.9
BCB (255,207,64,α)	0	5.4	6.6	8.8
BCB (255,147,128,α)	0	3.8	6.0	7.6
BCB (255,207,128,α)	0	4.2	6.2	7.9

diminishes where α increases. It is worth remarking here that the best gain in terms of PAPR is achieved when the value of α is great than 50. The introduction of block coding let us correct some errors with the same PAPR as the normal OFDM. Overall we see the technique proposed here easily outperforms the other approaches in terms of BER with good values in terms of PAPR.

TABLE IV. SNR LOSS IN dB OF THE PROPOSED METHOD, PEAK INSERTION AND NORMAL OFDM AT BER= 10^{-4}

Method	α=0	α=29	α=50	α=80
OFDM		0		
PI(α)	0	6.4	10.4	14.4
BCB (255,147,64,α)	-4.5	0.3	3.4	7.5
BCB (255,207,64,α)	-3.0	1.0	5.5	9.5
BCB (255,147,128,α)	-5.7	0.5	4.5	8.5
BCB (255,207,128,α)	-5.5	2.0	6.0	10.0

V. CONCLUSION

In this paper, we have proposed a new method to reduce PAPR and BER in the OFDM system based on BCH codes and Peak Insertion. The proposed algorithm achieved a good reduction in terms of PAPR reduction with good performances in terms of BER compared with Peak Insertion. Varying the parameter α we achieve a good reduction with a gain of 8.8dB when α is equal to 80 with a gain of 6dB in terms of BER compared with PI (with $\alpha = 80$). The simulation results show that the proposed method is simple, robust, and does not need side information with more flexibility to choose between PAPR reduction and BER performances. Future work is planned to study the proposed algorithm more in-depth by using Genetic

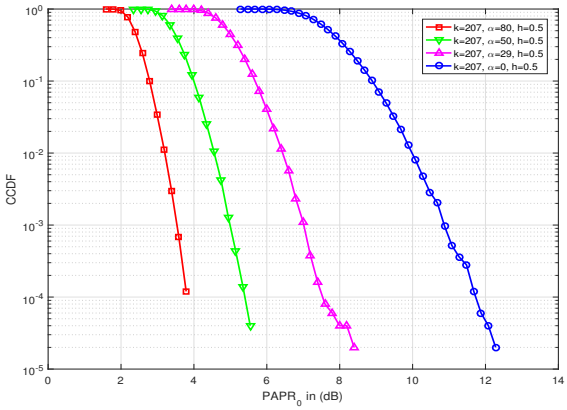


Fig. 6. CCDF Curves of the Proposed BCB-OFDM FFT=256, 4QAM, $BCB(n = 255, k = 207, v = 128, \alpha)$ with Variable α.

Algorithms to enhance the proposed algorithm to find the best comprise between PAPR reduction and BER performance. Also, we investigate other coding schemes especially the Goppa codes.

REFERENCES

[1] H. Bolcskei, "Mimo-ofdm wireless systems: basics, perspectives, and challenges," *IEEE wireless communications*, vol. 13, no. 4, pp. 31–37, 2006.

[2] K. Fazel and S. Kaiser, *Multi-carrier and spread spectrum systems: from OFDM and MC-CDMA to LTE and WiMAX*. John Wiley & Sons, 2008.

[3] M. N. Aarab and O. Chakkor, "Mimo-ofdm for wireless systems: An overview," in *International Conference on Artificial Intelligence and Symbolic Computation*. Springer, 2019, pp. 185–196.

[4] A. Svensson, A. Ahlén, A. Brunstrom, T. Ottosson, and M. Sternad, "An ofdm based system proposal for 4g downlinks," in *Multi-Carrier Spread-Spectrum*. Springer, 2004, pp. 15–22.

[5] B. Farhang-Boroujeny and H. Moradi, "Ofdm inspired waveforms for 5g," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2474–2492, 2016.

[6] N. K.P and C. K. N, "Studying applicability feasibility of ofdm in upcoming 5g network," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 1, 2017. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2017.080128

[7] F. Sandoval, G. Poitau, and F. Gagnon, "Hybrid peak-to-average power ratio reduction techniques: Review and performance comparison," *IEEE Access*, vol. 5, pp. 27 145–27 161, 2017.

[8] R. Ahmed, F. Schaich, and T. Wild, "OFDM enhancements for 5G based on filtering and windowing," in *Mult. Access Tech. 5G Wirel. Networks Beyond*. Springer, 2018, pp. 39–61.

[9] C. Rapp, "Effects of hpa-nonlinearity on a 4-dpsk/ofdm-signal for a digital sound broadcasting signal," *ESASP*, vol. 332, pp. 179–184, 1991.

[10] P. Banelli, G. Baruffa, and S. Cacopardi, "Effects of hpa nonlinearity on frequency multiplexed ofdm signals," *IEEE Transactions on Broadcasting*, vol. 47, no. 2, pp. 123–136, 2001.

[11] S. Gupta and A. Goel, "Improved selected mapping technique for reduction of papr in ofdm systems," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, 2020. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2020.0111016

[12] Y. A. Al-Jawhar, N. S. M. Shah, M. A. Taher, M. S. Ahmed, and K. N. Ramli, "An enhanced partial transmit sequence segmentation schemes to reduce the papr in ofdm systems," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 12, 2016. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2016.071209

[13] Y. A. Al-Jawhar, K. N. Ramli, A. Mustapha, S. A. Mostafa, N. S. Mohd Shah, and M. A. Taher, "Reducing PAPR with Low Complexity for 4G and 5G Waveform Designs," *IEEE Access*, vol. 7, pp. 97 673–97 688, 2019.

[14] B. Bakkas, I. Chana, and H. Ben-Azza, "PAPR reduction in MIMO-OFDM based on polar codes and companding technique," in *Proc. - 2019 Int. Conf. Adv. Commun. Technol. Networking, CommNet 2019*. IEEE, apr 2019, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8742379/

[15] K. Anoh, C. Tanriover, and B. Adebisi, "On the Optimization of Iterative Clipping and Filtering for PAPR Reduction in OFDM Systems," *IEEE Access*, vol. 5, pp. 12 004–12 013, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7938298/

[16] B. Bakkas, R. Benkhouya, I. Chana, and H. Ben-Azza, "Palm date leaf clipping: A new method to reduce papr in ofdm systems," *Information*, vol. 11, no. 4, p. 190, apr 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/4/190

[17] A. I. Siddiq, "PAPR reduction in OFDM systems using peak insertion," *AEU - Int. J. Electron. Commun.*, vol. 69, no. 2, pp. 573–578, feb 2015. [Online]. Available: http://dx.doi.org/10.1016/j.aeue.2014.11.009

[18] B. Wang, Q. Si, and M. Jin, "A novel tone reservation scheme based on deep learning for papr reduction in ofdm systems," *IEEE Communications Letters*, vol. 24, no. 6, pp. 1271–1274, 2020.

[19] P. P. Ann and R. Jose, "Comparison of papr reduction techniques in ofdm systems," in *2016 International Conference on Communication and Electronics Systems (ICCES)*. IEEE, 2016, pp. 1–5.

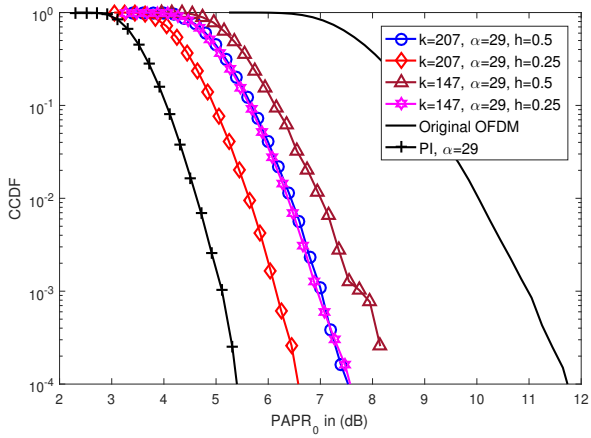
[20] W. G. Abera, "Comparative study of the performances of peak-to-average power ratio (papr) reduction techniques for orthogonal frequency division multiplexing (ofdm) signals," in *International Conference on Information and Communication Technology for Develoment for Africa*. Springer, 2017, pp. 56–67.

[21] P. Preenu Ann and R. Jose, "Comparison of PAPR reduction techniques in OFDM systems," *Proc. Int. Conf. Commun. Electron. Syst. ICCES 2016*, vol. 1, no. 4, pp. 40–49, 2016.

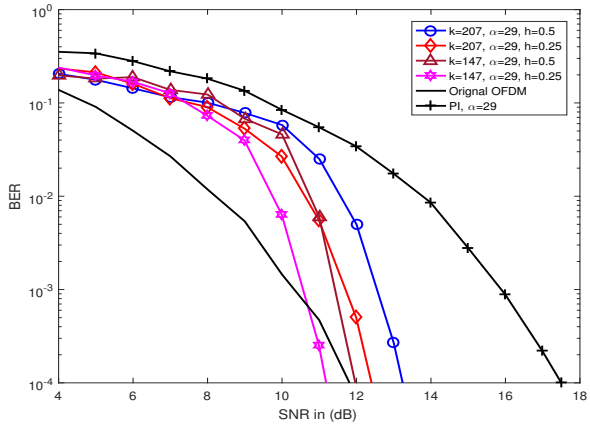
[22] N. Sabna, R. Revathy, and P. S. Pillai, "Bch coded ofdm for undersea acoustic links," in *2015 International Symposium on Ocean Electronics (SYMPOL)*. IEEE, 2015, pp. 1–6.

[23] P. Gupta, B. A. Kumar, and S. K. Jain, "Peak to average power ratio reduction in ofdm using higher order partitioned pts sequence and bose chaudhuri hocquenghem codes," in *2015 International Conference on Signal Processing and Communication Engineering Systems*. IEEE, 2015, pp. 443–447.

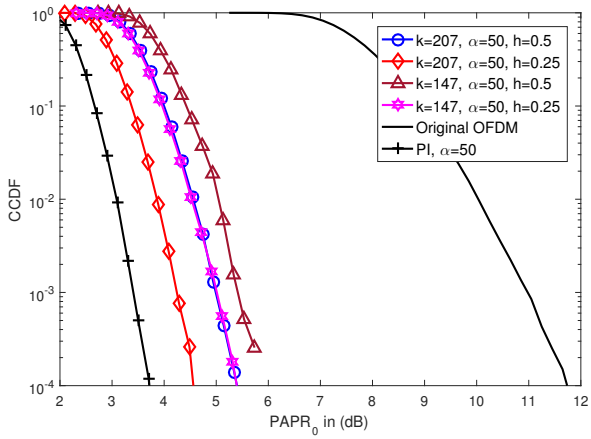
[24] P. Kumar, A. K. Ahuja, and R. Chakka, "Bch/hamming/cyclic coding techniques: comparison of papr-reduction performance in ofdm systems," in *International Conference on Intelligent Computing and Applications*. Springer, 2018, pp. 557–566.



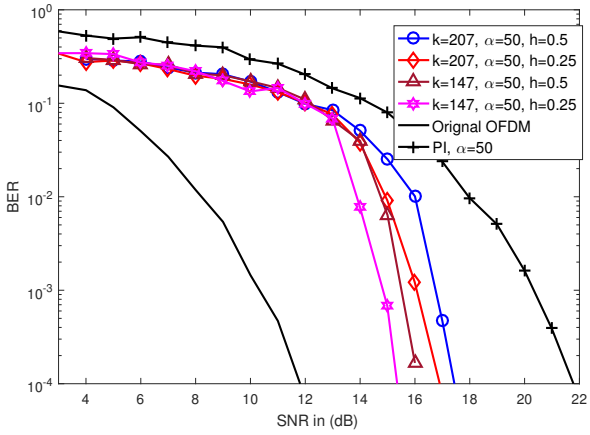
(a) CCDF $\alpha = 29$



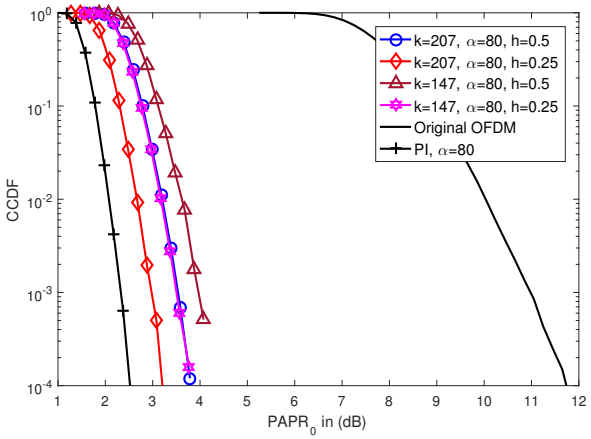
(b) BER, $\alpha = 29$



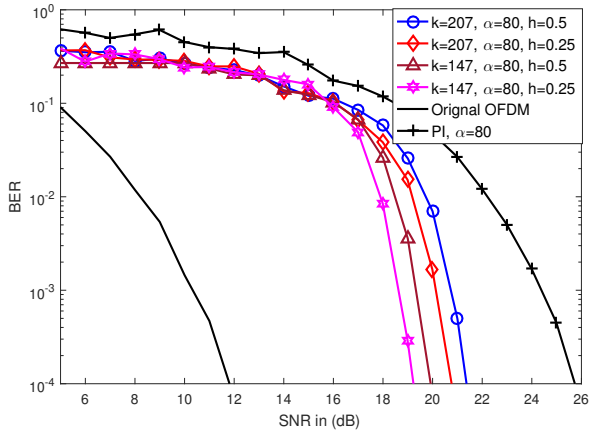
(c) CCDF, $\alpha = 50$



(d) BER, $\alpha = 50$



(e) CCDF, $\alpha = 80$



(f) BER, $\alpha = 80$

Fig. 7. CCDF and BER Performances of the proposed BCB-OFDM Compared with Peak Insertion and Normal OFDM, NFFT=256 with QAM Modulation Versus α.

Collision Resolution Techniques in Hash Table: A Review

Ahmed Dalhatu Yusuf[1], Saleh Abdullahi[2], Moussa Mahamat Boukar[3], Salisu Ibrahim Yusuf[4]
Nigerian Communications Commission[1]
Department of Computer Science, Nile University of Nigeria, Abuja, Nigeria[2,3,4]

***Abstract*—One of the major challenges of hashing is achieving constant access time $O(1)$ with an efficient memory space at a high load factor environment when various keys generate the same hash value or address. This problem causes a collision in the hash table, to resolve the collision and achieve constant access time $O(1)$ researchers have proposed several methods of handling collision most of which introduce a non-constant access time complexity at a worst-case scenario. In this study, the worst case of several proposed hashing collision resolution techniques are analyzed based on their time complexity at a high load factor environment, it was found that almost all the existing techniques have a non-constant access time complexity. However, they all require an additional computation for rehashing keys in a hash table some of which is as a result of deadlock while iterating to insert a key. It was also found out that there are wasted slots in a hah table in all the reviewed techniques. Therefore, this work, provides an in-depth understanding of collision resolution techniques which can serve as an avenue for further research work in the field.**

***Keywords*—Hashing; collision resolution; hash table; hash function; slot**

I. Introduction

Hashing is a data structure for searching an element from a collection with the primary goal of achieving a constant time complexity$O(1)$ [6], [7], [1]. It uses a hash function $h(key)$ to generate an address or a hash value of an element in a hash table. A hash table is a collection of slots in memory defined for storing a set of keys. A number of hashing techniques exist, all of them use a hash function to identify an address of a key in order to achieve constant access time both for insertion and searching a key. However, In a situation where a hash function $h(key)$ generates the same address/hash value for more than one key that introduces a collision in the hash table. The constraint of a hash table is only one element or key can be placed in a single slot. Therefore, to resolve the collision researchers have proposed several collision resolution techniques such as probing techniques, double hashing, separate chaining, cuckoo hashing etcetera for handling the colliding keys in the hash table. A number of these techniques introduce a non-constant time complexity in a high load factor environment.

Existing work on the hashing techniques generally focus on security oriented hashing [23], applications of hashing [9], [20], [8], and general comprehensive knowledge of hashing techniques. However, in this work we focused on runtime complexity of the existing technique and identifying the problems of the existing techniques so that a research gap will be provided which can serve as an avenue to improve upon collision resolution techniques. This is due to the importance hashing demonstrated in insertion, searching, and matching in many areas of computer [9], [8], [3]. In cryptography such as password verification, message digest, and Rabin-Karp algorithm. Therefore, a development in the area of hashing will advance the efficiency of several applications across many areas of computing.

In this work relevant proposed collision resolution techniques in hash table were reviewed. Highlighting the hash function employed in each method, how key is hashed into a hash table, key retrieval strategies and costs based on worse-case runtime complexity, alongside problems associated with each existing technique and we also provided a research gap in hashing technique for researchers to improve on.

II. Related Work

The research effort of Lianhua *et al* [9] proffered the main ideas on the prevailing hashing techniques for various "data and applications", it investigated the similarities and robustness of each technique in the two categories demonstrated in their work: data-situated and security-situated hashing and also present a brief application domain that requires hashing. The work was conducted due to an increase in the volume of data generated every day from diverse areas such as social network activities, daily transactions in the business domain, data from IoT applications, and other numerous domains. This increment of data has led to significant issues in analyzing and processing data and hashing strategy has been an efficient approach for fast data access for decennaries. The techniques and applications reviewed in their work have shown the positive impact hashing has on the performance of various applications such as networking, image classification, text classification and thus makes it an interesting area of study in order to make real-world applications very efficient.

The work of Tom [15] Indicated that several algorithms are implemented using dictionary complex type that most high-level programming comes with. This dictionary type could be implemented in several ways with a different data structure. However, a study has shown that for better lookup performance it should be implemented with a hash table. Python dictionary takes advantage of that, it uses a hash table with open addressing [30]. But the problem with that is whenever a collision occurs probing method has to happen. Therefore, In an environment where the collision is high then the lookup performance reduces. The trivial method is to use chainning for dictionary implementation.

Abhay [20] presented a technique that minimized the memory space used in implementing the hash table by compressing

the key for data-item in the hash table. The hash value $h(key)$ can be generated by any hash function and subsequently, the compressed value is generated to mapped the input key.

Sailesh *et al* [8] proposed a technique that try to achieve constant time for retrieving keys at high load factor environment and memory minify bandwidth in high-performance networking subsystem. The method uses a hash table with many multiple logical chunks given a key n likely slot in memory. An item will be mapped with $h(k)$, which inserted the key in the search space U in the scope of the chunked subscripts of the hash tables, i.e. *h:* $U = \{0, 1, ..., |U| - 1\}$, where $|U|$ is the length of the chunked hash table. An item can be inserted in a bucket $h(k)$ in any of the chunked hash tables. In the event where all the chunked buckets for *h(k)* are not empty then a collision is inevitable.

Yuanyuan *et al* [4] improved upon open address cuckoo hashing by overcoming the problem of an infinite loop at a time of insertion, which reduces the efficiency of query processing. They proposed a better technique called SmartCuckoo, which presents the relationship of hashing using directed pseudo forest and uses it subsequently to indict element placement for the correct determination of the existence of an infinite loop. SmartCuckoo can also predict insertion failure without going through some probing steps. However, in some environment the prediction might not be accurate. Therefore, there is still a change for an infinite loop. The work has been implemented on a "cloud storage system" the source code has been released for public users.

Peter *et al* [12] presented an approach for resolving collision in one-dimensional array. The technique concatenated (dot (.)) the key and the *h(k)* and insert it in the first empty bucket in the hash table. For example, in a hash table of *size* 7, *h(23) = 23* mod *7*, will be placed 2.23 at the first available cell in the hash table. Searching an element in this technique is linear *O*(n), hence the *h(k)* will not help in locating the key from the hash table.

Randomized hashing was proposed by Shai *et al* [11] as a process that takes a message and returns a hash value of the message that can be used in digital signature without any modification in traditional hash function such as SHA. The objective of their work is to free "digital signature schemes" from their dependence on collision contention.

A. Collision Resolution Strategies

The circumstance where a hash value calculated using a hash function matches with another hash value that is already involved within the hash table is named as collision [24]. To place all the colliding keys in the hash table that could be achieved using a method called collision resolution. Collision resolution refers to a situation when two items hash to the same slot, and a systematic method must be used to insert the second item in another slot in the hash table. An example of collision is described in Fig. 1.

1) Open Addressing: is a technique that resolve colliding keys in the hash table by looking for an empty slot using some sequence of probing techniques to find a new slot for an element that caused the collision [28]. This hash table has a probe sequence which is usually in the form: *(h(k) = [h(k) +*

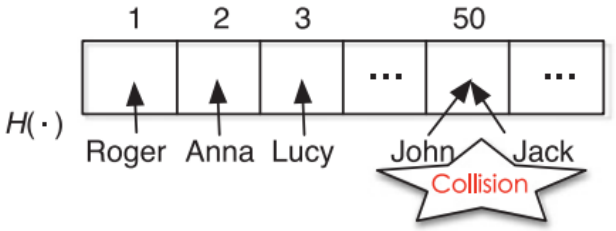


Fig. 1. Example of Collision [9].

c(i) mod *n, for i=0,1,..,n-1)* where *h* is the hash function and *n* is the size of the hash table. The function *c(i)* is required to iterate through *n-1*.

The probe sequence method include the following:

- Linear probing, in this method if a collision occurs it resolve it by finding the next empty slot in the hash table and place a key. To search a key say y, the approach is to look for it in a hash table starting with an index of $h(y)$ and continue to the next slot in the hash table i.e. $h(y)+1, h(y)+2, ...$, until an empty slot is reached or a slot whose content is y. If an empty slot is reached then the key is not in the hash table. The problem with this method is clustering, (likelihood of one collision causing neighbouring bucket collision) at a high load factor which degrades its performance [25]. The method was founded by "Gene, Elaine and Samuel" [26].
- Quadratic probing, is another open addressing collision resolution method in which the interval to place a key if collision occur is quadratic i.e $h(key) + 1^2, h(key) + 2^2, ..., h(key) + n^2$. This technique considers better than open addressing with linear probing since it keeps away from the problem of clustering, even though it is not resistant to it [19]. The major problem with this method is finding an empty bucket is challenging when the hash table is $> 59\%$ full [27].
- Double hashing, this method resolve a collision by using another hash function to determine the interval to insert a key. For instance given, two different hash functions h_x and h_y, the position i of key in the hash table of size $|n|$ slots is; $h_x(key) + i * h_y(key)\%|n| for\ i = 1, ..., n - 1$, where h_x and $h_y \in \cup = \{h_a, h_b, ..., h_z\}$ [29].

The work of Benjamin *et al.* [10] demonstrated the usage and efficiency of open addressing with quadratic probing to handle collision in communication between applications that use different communication design, computation, and data structures. For example, data can be distributed to many processes but each process will carry different tasks independently on the data. The work offered a Berkeley Container Library BCL; a cross-platform data structure library for a "one-sided communication" environment for parallel applications. The BCL is composed in C++ programming and its data structures phase are intended to be sans coordination, utilizing one-sided communication primitives that can be executed utilizing RDMA equipment while not requiring coordination with remote CPUs. Along these lines, BCL is steady with the soul

of Partitioned Global Address Space GAS language, however profferer efficient add and search operations in the hash table, instead of reading and read operation of PGAS languages. BCL provides a central data structure for all the processes and can be used by each process in a parallel program.

2) Separate Chaining: This strategy uses a collection of nodes known as a linked list or list data structure to resolve the colliding keys in the hash table whenever a collision occurs as described in Fig. 2. In a high load factor environment this method provide a non-constant time complexity of $O(n)$ for inserting and retrieving a key from the hash table and tends to cause problem of tracking linked list [5]. However, another method invented by Dhar *et al.* [16] provide better performance from $O(1+n)$ to $O(logn)$. The technique uses a binary search tree to chain a collide key rather than using a list or linked list which reduces the time for searching a key. The problem of this method is an additional cost of balancing BST when the inserted keys cause a skewed binary tree.

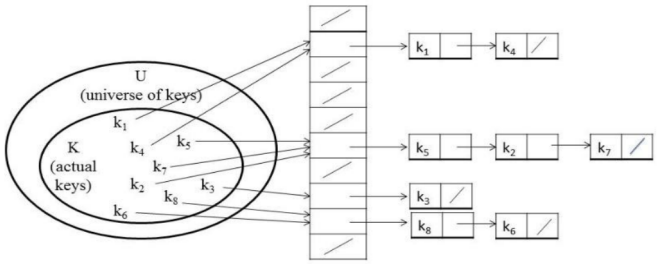


Fig. 2. Separate Chaining using Linked list [12].

3) Coalesced Hashing: is approach takes advantage of two different collision resolution techniques to handle collision in a hash table, an open addressing and chaining. It uses similar insertion procedure as open addressing to insert an element in the hash table using $h(k)\ mod\ n$. When a collision occurs at *i* position in the hash table, coalesced hashing resolve it similar to separate chaining by inserting the key that causes the collision in the first empty slot from the bottom of the hash table i.e. *for (i ¡ n, i ¡=0, i–)* where *i* is an index and *n* is the size of the hash table. It then chains the colliding key original hash value to hash value the colliding key is inserted using a pointer instead of creating a linked list like separated chaining. It minimizes space usage but constant time lookup is not achievable at a high load factor like an open addressing and separate chaining [2]. Any open addressing method can be used to identify a position to insert a key that collides in coalesced hashing.

4) Cuckoo Hashing: This is another open addressing technique that was first introduced in 2004 by "Flemming and Rasmus" [13]. The method is ubiquitous and uses in an array of real-life applications [14], [17], [18], [21]. It uses two or more hash functions to insert key to slot in a hash table, which means any key in *U = {k1, k2... kn}* can be in more than one slot. Any key can also be relocated to another slot in the hash table. Insert a key has a number of hash function options say, *h1(k)* and *h2(k)*. Relocation of a key can be done if *h1(k)* and *h2(k)*. are not free for insertion. This problem can be overcome by relocating an existing key to a new slot using another *h(k)* and supersede the new key into relocating key position hashing. If the relocating key *h(k)* is not empty, then repeat relocating key supersede another key until the method gets a free slot. In a situation it iterates through the hash table without resolving the problem, all the keys will be rehashed with different *h(k)*. *N* number of rehashes might be conducted in order for cuckcoo to achieve. However, "MinCounter" technique presented in [22] reduced the number of rehashing by superseding a new key with a rarely accessed key to address collision in a hash table instead of superseding any random key. Each slot in a MinCounter method has a counter variable that keeps track of the number relocation that occurs at a slot. To insert a new key it checks the counter variable and inserts it into a slot with a minimum value rather than iterating through the hash tables for an empty slot to place a key. In a situation of insertion failure, a key is placed in the "memory cache" to avoid rehashing. "MinCounter" provides better performance for inserting and query response in cloud services. The structure of this algorithm is described in Fig. 3.

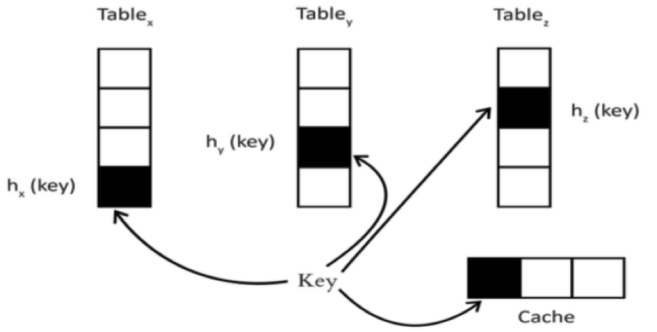


Fig. 3. Example of MinCounter Hashing Technique [22].

Jeyaraj *et al* [14] research effort improve the performance of eliminating duplicate fingerprint date in data duplication method for a backup system using cuckoo hashing. The approach in the work is implemented with two tables T_i and T_j. To insert a new fingerprint. We use fp to denote fingerprint, It will check; **if** $T_i[h(fp)] = fp$ then it will skip the insertion. Else it checks; **if** $T_i[h_i(fp)]$ = NULL then; $T_i[h_i(fp)] = fp$ else; **if** $T_i[h_i(fp)] \neq$ NULL then check; **if** $T_j[h_j(fp)]$ = NULL then; $T_j[h_j(fp)] = fp$ It continue with normal cuckoo process until all the available fingerprint are distributed into the two hash table depending on the when cuckoo succeed. The research work also support parallel insertion into the hash tables which gives the technique a better throughput and minimises memory space compare to [5] but add cost for inserting a key similar to techniques in [28], [10], [2], [13].

III. Research Gap

Hashing is indeed a very important algorithm that depicted interest from many areas of storage systems. It must be efficient for retrieving an element at all times and the memory space should be utilized efficiently due to the limited nature of the storage system. The worse-case time complexity, mapping method, key retrieval approach and problem of most of the important techniques is shown in Table I. Reviewed research efforts above tried to use hashing in various domains to

TABLE I. SUMMARY OF SOME OF THE IMPORTANT REVIEWED HASH COLLISION RESOLUTION TECHNIQUES

Technique	Mapping	Lookup	Problem	Time complexity of lookup & mapping (worst-case)
Linear probing [26]	$x \leftarrow h(key) \bmod n,\ i \leftarrow 0$ **while** $(x \neq NULL)\{\ i \leftarrow i+1,\ x \leftarrow x+i \bmod n\}$ insert 'key' at x position. Where n is the size of the hash table.	$x \leftarrow h(key) \bmod n,\ i \leftarrow 0$ **while** $(x < n)\{$**if**$(x = key)$ print 'Key found' **break loop** $i \leftarrow i+1,\ x \leftarrow x + i \bmod n$ **if**$(x = NULL)$ print 'Key doesn't exist' **break loop**$\}$.	• Clustering problem • Wasted slot(s) • Rehashing	$O(n),\ O(n)$
Quadratic probing [19]	This works similar to linear probing but move in quadratic form to resolve a collision. $x \leftarrow h(key) \bmod n,\ i \leftarrow 0$ **while** $(x \neq NULL)\{\ i \leftarrow i+1,\ x \leftarrow x + i^2 \bmod n\}$ insert 'key' at x position. Where n is the size of the hash table.	$x \leftarrow h(key) \bmod n,\ i \leftarrow 0$ **while** $(x < n)\{$**if**$(x = key)$ print 'Key found' **break loop** $i \leftarrow i+1,\ x \leftarrow x + i^2 \bmod n$ **if**$(x = NULL)$ print 'Key doesn't exist' **break loop**$\}$.	• This technique keeps away from clustering problem although is not resistant to it. • Wasted slot(s) • Rehashing • Finding empty slot if hash table is $> 59\%$ occupied is challenging.	$O(n),\ O(n)$
Double hashing [29]	This technique uses two different hash functions h_x and h_y, the position i of key in the hash table of size $\|n\|$ slots is; $h_x(key) + i * h_y(key)\%\|n\|$**for** $i = 1, ..., n-1$, where h_x and $h_y \in \cup = \{h_a, h_b, ..., h_z\}$	$i \leftarrow 0,\ x \leftarrow h_x(key),\ y \leftarrow h_y(key)$ **while**$(T[(x+i*y) \bmod n] \neq key)\{$ **if**$(T[(x+i*y) \bmod n] = -1)\{$ print "key does not exist" **break**$\}$ $i \leftarrow i+1\ \}$ print "Key found"	• Wasted slot(s) • Rehashing	$O(n),\ O(n)$
Coalesced hashing [2]	$x \leftarrow h(key) \bmod n,\ i \leftarrow n$ **while** $(x \neq NULL\ AND\ i > -1)\{\ i \leftarrow i-1,\ x \leftarrow x + i \bmod n\}$ insert 'key' at x position then set a pointer from the colliding position. Where n is the size of the hash table.	$x \leftarrow h(key) \bmod n,\ i \leftarrow n$ **while** $(x < n\ AND\ i > -1)\{$**if**$(x = key)$ print 'Key found' **break loop** $i \leftarrow i-1,\ x \leftarrow x + i \bmod n$ **if**$(x = NULL)$ print 'Key doesn't exist' **break loop**$\}$.	• Wasted slot(s) • Rehashing	$O(n),\ O(n)$
Cuckoo hashing [13]	This method uses two hash tables (T_1 & T_2) and two hash functions $h_1(key)$ & $h_2(key)$). $h_1(key) = key \bmod n$ and $h_2(key) = (key/n) \bmod n$. **if** $(T_1[h_1(key) = key\ OR\ T_2[h_2(key) = key]])$ print "Key already exist" **else while**(true)**{** key swap $T_1[h_1(key)]$ **if** $key = NULL\ key$ swap $T_2[h_2(key)]$ **if** $key = NULL$**}** rehash all keys then try inserting the key.	Similarly, retrieve a key uses the two hash tables (T_1 & T_2) and two hash functions $h_1(key)$ & $h_2(key)$). $h_1(key) = key \bmod n$ and $h_2(key) = (key/n) \bmod n$. **if** $(T_1[h_1(key) = key\ OR\ T_2[h_2(key) = key]])$ print "Key already exist" **else** print "Key does not exist"	• Wasted slots • Rehashing • High cost of insertion which could lead to deadlock	$O(1),\ O(n)$
Separate chaining with linked list hashing [5]	This method uses $h(key) \bmod n$ to insert a key like linear probing. But resolve the colliding keys by using linked list.	**if** $(h(key) \bmod n) = key$ print "Key found" **else**$\{\ p \leftarrow head$ **while**$(p \neq NULL$ AND $p.info \neq key)\{$ access $p.info\quad p = p.link\ \}$	• Wasted slots • Rehashing	$O(n),\ O(n)$
Separate chaining with binary search tree hashing [16]	This works similar to separate chaining with a linked list but it resolve collision using a binary search tree. The time complexity for searching a key from a binary search tree is $O(log\ n)$.	$node \leftarrow start$ **while** $(node \neq NULL)\{$ **if**$(key[node] = key)$ return y else if $key[node] < key$ then $node \leftarrow$ right[node] else $node \leftarrow$ left[node] $\}$ print "Key not found". Where $start$ is the root node.	• Wasted slots • Rehashing • Computation for balancing skew tree	$O(log\ n),\ O(log\ n)$

improve the performance of retrieving, matching, and inserting data. However, these existing collision resolution techniques both have their pros and cons. In this work we identified three major issues of the current techniques mention below:

Issues I, All the existing technique mentioned reviewed in this work suggested using prime number as a size of a hash table $|T|$, usually in form:

$$s = (x \times |key|)$$
$$|T| = > p(s)$$

where $> p$, means next prime number of s and $x > 1.0$. The problem with this method it creates a wasted slot in the hash table and in that case it does not achieve the goal of the hashing that is primarily designed to utiliaze minimum amount of memory to store data, which is less than the amount to store the actual data [26].

Theorem: For n arbitrary set of keys K to hash into a hash table T

$$|T| = > p\{x \times |K|\} \Rightarrow \exists_{wastedSlot} \in T$$

Proof: Consider a set of positive integers $K = \{k_1, k_2, ..., k_n\}$ to hash into a hash table

$$|T| = > p(x \times |K|)$$
$$Let\ I_{i=1}^{|K|} = \begin{pmatrix} i{+}{+} & if\ h(k_i) \rightarrow T \\ i{=}i{+}0 & else \end{pmatrix}$$
$$h(k_i) = k_i\ mod\ |T|$$
$$|T| > |K|$$

So, after mapping k_n into T, I will be equal to $|K|$
Therefore, $wastedSlot = |T| - I$

Issues II, Rehashing is a problem with all the existing techniques, which has to be done whenever there is an additional element to hash as a result of determining the size of hash table described in *issues I* and also when a certain threshold is reached which was considered to be set for some hashing technique like, double hashing technique [29]. For any arbitrary key is map into T

$$K = \{k_1, k_2, ..., k_n\} \rightarrow T[0, 1, ..., m-1]$$

Every ith location of a key is determine with:

$$h(key)\ mod\ |T|$$

Therefore, any change in $|T|$ all the element need to be remapped into T:

$$h(key) \rightarrow T\ using\ h(key)\ mod\ |T_{new}|$$

Issues III, The better performance technique among the reviewed works is cuckoo hashing, which has a time complexity of $O(1)$ to search for an element but has a deadlock problem which is a result of a high number of relocations before inserting an element. To resolve this problem of deadlock entire rehashing of the keys has to be done which is quite time-consuming and not efficient. It also has a high amount of wasted slots compare to other reviewed technique this because it uses two hash tables.

IV. Conclusion

In conclusion, this work reviewed a number of different collision resolution techniques in the hash table. These techniques were employed in many areas of computer science such as IP address lookup, job balancing, security, etc. The time complexity for inserting and retrieving an element of all the collision resolution techniques was identified. However, we found that achieving constant access time with a good insertion performance is still challenging with all the current collision resolution techniques. This work provided runtime complexity and the major problems associated with the existing collision resolution techniques which can serve as an avenue for further research in the field. Here, the analysis was based on worse-case time complexity of the respective algorithms. However, other future research should consider other different aspect of algorithm analysis such as space complexity and most suitable conditions with respect to input size through mathematical notation or simulation.

Acknowledgment

I will like to express my gratitude to my parents for their advice and support. Many thanks goes to Nile University of Nigeria for the 90% financial support opportunity provided to carryout this work.

References

[1] Black, Paul E. "DADS: The On-Line Dictionary of Algorithms and Data Structures," NIST: Gaithersburg, MD, USA, 2020.

[2] Sriram, Ranjena, et al. "Efficient Data Cleaning Algorithm and Swift Unique User Identification Algorithm Using Coalesced Hashing and Binary Search Techniques for Web Usage Mining," International Journal of Pure and Applied Mathematics 118.18, 2018.

[3] Brenton Lessley and Hank Childs. "Data-Parallel Hashing Techniques for GPU Architectures," IEEE Transactions on Parallel and Distributed Systems Volume: 31, Issue: 1, 2020.

[4] Sun, Yuanyuan, et al. "SmartCuckoo: a fast and cost-efficient hashing index scheme for cloud storage systems." 2017 {USENIX} Annual Technical Conference ({USENIX}{ATC} 17). 2017.

[5] Brad Miller, David Ranum. "Problem Solving with Algorithms and Data Structures," 2013

[6] Necaise, Rance D. "Data structures and algorithms using Python," Wiley Publishing, 2010.

[7] Cormen, Thomas H., et al. Introduction to algorithms. MIT press, 2009.

[8] Sailesh Kumar, Patrick Crowley. "Segmented Hash: An Efficient Hash Table Implementation for High Performance Networking Subsystems," 2005 Symposium on Architectures for Networking and Communications Systems (ANCS), 2005.

[9] Lianhua Chi, Xingquan Zhu. "Hashing techniques: A survey and taxonomy," ACM Computing Surveys, Vol. 50, No. 1, Article 11, 2017.

[10] Brock, Benjamin, Ayd?n Buluc, and Katherine Yelick. "BCL: A cross-platform distributed data structures library," Proceedings of the 48th International Conference on Parallel Processing. 2019.

[11] Halevi, Shai, and Hugo Krawczyk. "Strengthening digital signatures via randomized hashing," Annual International Cryptology Conference. Springer, Berlin, Heidelberg, 2006.

[12] Nimbe, Peter, Samuel Ofori Frimpong, and Michael Opoku. "An efficient strategy for collision resolution in hash tables," International Journal of Computer Applications 99.10, 2014.

[13] Pagh, Rasmus, Flemming Friche Rodler. "Cuckoo hashing," Journal of Algorithms 51.2, 2004.

[14] Jane Rubel A. Jeyaraj, Sundarakantham Kambaraj and Velmurugan Dharmarajan. "High-speed data deduplication using Parallelized Cuckoo Hashing," Turkish Journal of Electrical Engineering & Computer Sciences 2018

[15] Van Dijk, Tom. "Analysing and improving hash table performance," 10th Twente Student Conference on IT. University of Twente, Faculty of Electrical Engineering and Computer Science, 2009.

[16] Dhar, Siddharth, et al. "A tree based approach to improve traditional collision avoidance mechanisms of hashing." 2017 International Conference on Inventive Computing and Informatics (ICICI). IEEE, 2017.

[17] Debnath, Biplob K., Sudipta Sengupta, and Jin Li. "ChunkStash: Speeding Up Inline Storage Deduplication Using Flash Memory," USENIX annual technical conference, 2010.

[18] A. Kirsch and M. Mitzenmacher, "The power of one move: Hashing schemes for hardware," IEEE/ACM Transactions on Networking, vol. 18, no. 6, pp. 1752?1765, 2010.

[19] Konheim, Alan G. "Hashing in computer science: Fifty years of slicing and dicing," John Wiley & Sons, 2010.

[20] Abhay Kulkarni. "Efficient Hash Table Key Storage," Avago Technologies International Sales Pte . Limited, Singapore (SG), 2019.

[21] Y. Hua, B. Xiao, and X. Liu. "Nest: Locality-aware approximate query service for cloud computing," Proceedings of the 32nd IEEE International Conference on Computer Communications(INFOCOM), pp. 1327-1335, 2013.

[22] Sun, Yuanyuan, et al. "MinCounter: An efficient cuckoo hashing scheme for cloud storage systems," 2015 31st Symposium on Mass Storage Systems and Technologies (MSST). IEEE, 2015.

[23] Arvind K. Sharma ; S.K. Mittal. "Cryptography & Network Security Hash Function Applications, Attacks and Advances: A Review,". 2019 Third International Conference on Inventive Systems and Control (ICISC), 2019

[24] Joux, Antoine. "Multicollisions in iterated hash functions. Application to cascaded constructions." Annual International Cryptology Conference. Springer, Berlin, Heidelberg, 2004.

[25] Goodrich, Michael T., and Roberto Tamassia. "Algorithm design and applications," Hoboken: Wiley, 2015.

[26] Knuth, Donald E. "Sorting and searching (6. printing, newly updated and rev. ed.). Boston [ua]." 2000

[27] Weiss, Mark Allen, "Data Structures and Algorithm Analysis in C++," Pearson Education. ISBN 978-81-317-1474-4, 2009.

[28] Agrawal, Anand, Sriram Bhyravarapu, and Nuthalapati Venkata Krishna Chaitanya. "Matrix hashing with two level of collision resolution." 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2018.

[29] Phillip G.Bradford and Michael N. Katehakis, "A Probabilistic Study on Combinatorial Expanders and Hashing", SIAM Journal on Computing, 37 (1): 83-111, doi:10.1137/S009753970444630X, 2017

[30] Kumar, Arun, and Supriya P. Panda. "A survey: how python pitches in IT-world." 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). IEEE, 2019.

Future Friend Recommendation System based on User Similarities in Large-Scale on Social Network

Md. Amirul Islam[1], Linta Islam[2], Md. Mahmudul Hasan[3], Partho Ghose[4],
Uzzal Kumar Acharjee[5], Md. Ashraf Kamal[6]
Department of Computer Science and Engineering, Jagannath University, Dhaka, Bangladesh[1,2,3,4,5]
Department of Computer Science and Engineering, World University of Bangladesh, Dhaka, Bangladesh[1,6]

***Abstract*—Friendship is one of the most important issues in online social networks (OSN). Researchers analyze the OSN to determine how people are connected to a network and how new connections are developed. Most of the existing methods cannot efficiently evaluate a friendship graphs internal connectivity and decline to render a proper recommendation. This paper presented three proposed algorithms that can apply in OSN to predict future friends recommendations for the users. Using network and profile similarity proposed approach can measure the similarity among the users. To predict the user similarity, we calculated an average weight that indicates the probability of two users being similar by considering every precise subset of some profile attributes such as age, profession, location, and interest rather than taking the only average of the superset profile attributes. The suggested algorithms perform a significant enhancement in prediction accuracy 97% and precision 96.566%. Furthermore, the proposed recommendation frameworks can handle any profile attribute's missing value by assuming the value based on friends' profile attributes.**

***Keywords*—Social networks; recommendation framework; profile similarity; network similarity**

I. Introduction

OSN is a platform for sharing information (such as opinions, news, views), communication, business over the internet and related with the connectivity of people. The popularity of OSN increases day by day in recent times, and OSN data are referred to as one of the most important sources of information over the internet [1]. It is a great medium to be connected with more similar and both known and unknown people to share their own opinions and do audio or video chat. The OSN provides the facility for spreading news over the internet, helps to smooth communication with individuals or the community quickly, and helps continued other internet-based activities (such as online shopping and blogging).

An OSN is interpreted as a graph $g = (v, e)$, where users are denoted as node v, and the relation between users is denoted as edge e. Online social networking encloses networking for business, pleasure, and all points in between users. Networks themselves have different objectives, and their online reproduction work in many ways. A social network allows people to share information with friends and familiarity, both old and new. Therefore, every user creates a personal network based on some user properties and wants to broaden their network to create a new friendship using profile and network similarity. A user in a network creates a new friendship link with others for communication or sharing views, opinions or other information after creating an account. Mainly two points are involved in creating a new relationship in the social network by the theory of Homophily [2]. The first point is that users try to establish relationships with other users based on who is closer to them on the social graph. And the second point is users form relationships with users who are comparable to them and have particular properties like occupation, age, religion, hobby, gender, etc.

In social networks, how users are connected can be known by analyzing social networks. We can find the hidden patterns of the network and which path is best for spreading news, advertisements and political opinions. Analyzing online social networks helps us how to impact social media on human behaviours and how social media use in a convenient way [3]. Facebook[1], Twitter[2], VKontakte[3], Flickr[4], YouTube[5] are the most popular OSN attracted people by their impact on the internet as an excellent media for sharing news, opinions, interest, pictures, videos and a great communication medium.

Network similarity indicates the similarity among different networks rather than other nodes in a social network. Each user in a social network has their sub-network with friends and friendship links, and with time, users want to broaden their network for information sharing by including new friends. People want to establish new friendships with others who are closer to each other in OSN. Two graphs are used to compute the network similarity in the social graph. One is a friendship graph, and another one is the mutual friends' graph. The profile contents are unstructured keywords such as education, profession, gender, age, interest, and one or more of these are used for finding similarities between profiles. The string matching method is used to calculate the profile similarity among the profile attributes. In the paper [4], authors calculated the profile similarity of Facebook by handled only the individual profile value (Interest) of the users. On the other hand, authors in the paper [1] calculated the profile similarity of Facebook by considered user occupation, education, and gender.

Today OSN has become a large field of online activities, and it continues broadening day by day. So, it becomes a complex topic for a user to find a similar account from an extensive network. Information interchange is a common phenomenon in social networks. Those activities bring an

[1]https://www.facebook.com/
[2]https://twitter.com/
[3]https://vk.com/
[4]https://www.flickr.com/
[5]https://www.youtube.com/

excess of messages that make users confused about what they want. OSN recommendation system [5] is a framework that recommends the user to others by analyzing their available social information. Information of users in online social networks are largely available, mining profile information which can be able to predict user personality, that is an essential issue for finding user preference for recommending products, songs, online game mining social data [6]. In social networks, users are introduced to each other in several ways. Friend matching is a technique that can help to find friends on social media. Users connect using similar profile information such as similar educational background, similar home town or same interest.

In most cases, a recommendation system fails to recommend a similar use on the web because of missing profile content [7]. For constructing a recommendation system, it is essential to confirm that the user profile contents are available. Some profiles in OSN cannot provide us with all attributes information; in this case, the recommendation system fails to recommend appropriately. In order to overcome this problem, we need a technique that can infer profile missing value. Therefore the primary aim of this paper is to mine an extensive group of social data and discover the more related people or users for the recommendation. Our suggested approach computes the similarity by utilizing various similarity measures among all probable peoples and recommends them if they are not friends still now.

Following are the contributions of this paper:

- Analysis of online social networks to find user similarity between two users by combining missing profile items, network similarity, and the weight of each attribute set for profile similarity.
- To propose prediction of profile disappeared value (Algorithm 1) to handle missing profile values.
- To construct a better future friend recommendation system for users, we presented two modified profile weight calculations methods named Feature Weight Computation System (FWCS) and Friend Matching System (FMS).

The remaining part of this paper is outlined as follows: In Section II, literature review described. Our proposed methods and algorithm with calculation of user similarities for friend recommendation system are discussed in Section III. Experimental results and discussion is shown in Section IV and in Section V, briefly concludes our research effort with future research directions.

II. Literature Review

Focusing on profoundly significant work, we audit some current related work and afterwards sum up standard techniques for recommendation framework. This section briefly presents a few studies handle in recent years by different strategies.

Friendship is a fundamental relationship in social networks and suggests friends are practical activities to overcome this, [8] proposed a friendship recommendation solution by profile matching. This work assigns different weights in different items and developed a mining model to discover different factors actual degree of influence by measuring the profile attributes using some similarity measures. The proposed framework yielded an accuracy of 95%. In recent years, a similarity measure between nodes is defined based on the features of their neighbourhood information from many users in an extensive network. In the paper [9], suggested parametric system for neighborhood-based similarity is applied to calculate several similarities and calculative costs among neighborhood nodes. A unique multi-feature SVM based friend recommendation model (MF-SVM) is introduced by Xin et al. [10]. This proposed model is a binary classification problem. It can handle the sparse situation of user location and user-user formation in the location-based social network. The authors extracted three features using their proposed model but did not consider or handle missing values, network similarities, and weight calculation. To evaluate the MF-SVM model, two real-world data sets, Foursquare and Gowalla are chosen. The model achieved an accuracy of 90%.

In 2020, Qader et al. [11] suggested a Dual-Stage FR model to recommends users to other users based on user interests. The model applies the double stage technique on unlabeled information of 1241 users collected from OSN users via the online survey. The authors mainly combined user-based collaborative filtering (UBCF) and graph-based FR in their proposed model. However, the drawback of this technique is that the computational cost linearly increases with the user. The accuracy of the model was 86%. In 2020, Soni et al. [12] presented a novel FR framework based on their similar choices, activities, preference and locations. The authors replaced k-means clustering with hierarchical clustering in their proposed model and principal component analysis (PCA) techniques applied to the dataset for dimensionality reduction. However, the limitation of this model is the cost of PCA calculation when the matrices become high. The model achieved an accuracy of 89.47%.

Kumar et al. [13] introduced a graph-based FRS utilizing two CF systems: the number of mutual users and the influence factor. Then, it assigned a score number to every conceivable friend to track down the higher closeness between clients dependent on the highest score number. The datasets utilized are Stanford SNAP, which individually consists of 4039 and 81,306 clients from Facebook and Twitter. The model achieved an accuracy of 97.2%. In the research paper [14], a new framework is proposed called multi-step resource allocation (MSRA) to predict the implicit relationships. The authors are mainly combined three sources of information: a user-item matrix, explicit and implicit associations. To evaluated the proposed method, two real datasets are used (Last.Fm and Ciao). The proposed MSRA model achieved an accuracy of 95.80%. To predict the future friends in the social networks, Shabaz et al. [15] proposed a new approach called Shabaz–Urvashi Link Prediction (SULP). This new technique can solve the problem of linking isolated or missing nodes in social networks and connect the nodes in a network faster than any other link prediction algorithm that exists. For this reason, this novel approach can reduce the connection time and resources involved in it. The proposed SULP model achieved a precision of 76%, recall of 82% and TRP of 88%. In 2021, Berkani et al. [16] proposed a unique recommendation framework for users in social networks. This method mainly based on semantic and social-based classification of the user

profiles. The authors have used two classifications techniques: the K-means algorithm and K-Nearest Neighbours algorithm to optimize the performances of the recommendations systems. This proposed model used two datasets, one is the Yelp datasets, and another one is the Rich Epinions datasets. The proposed method achieved an accuracy of 95%.

Apart from this, researchers continuously contribute to developing an efficient system to recommend friends to the users. We have taken some recent papers, and their contribution in a different part of similarity measure is shown in Table I.

It is shown that different parts of similarity measurement are not fulfilled. For this reason, existing techniques could not measure user similarity efficiently, and the recommendation system could not recommend properly. In this paper, we proposed an efficient technique for measuring user similarity between two users, combining all parts (inferring missing profile item, network similarity, the weight of each attribute set for profile similarity) of similarity measurement and an efficient recommendation system. The existing method measuring network similarity uses mutual friendship, and target user friendship graph edges only. In a new friendship formation, two users have the same influence. In our network similarity, method friendship graph edges two of them are used. This work used only observed frequency measure in profile similarity and set the same weight to each profile attribute. Utilizing weight computation of every profile attribute's performance by considering only profile similarity, authors recommended future friends for the users in their paper [8]. But our proposed framework has diverged from other research contributions because firstly, we computed the weight for every set of profile attributes and then merged it with the network similarity. For creating future friendships among the users', the profile attributes set contributed a vital influence. In our proposed method combine feature computerization systems based on the supervised learning strategy.

III. Research Methodology

This section represents the main portion of the paper: the proposed architecture design and development of the proposed algorithms.

A. Proposed Architecture Design

Friendship and profile information from the online social network is used in our suggested model to calculate the similarity of several users who do not belong to the friendship graph. There are four phases in our proposed model: The first phase is used to extract the user's features based on the user profile and handle the user's profile attributes if there are any missing data by assuming the value. Data mining technology is used in the second phase. In the third phase, the friend matcher method recommends the future friend for the user by predicting the user's similarity. And in the final phase, the feature automation technique (Supervised Learning-Based) is used to formation the friendship by identifying the most prominent attributes. In Fig. 1 demonstrated the proposed model architecture. All parts of this model are described in detail below.

1) User Profile: Each user in a social network has two types of information: profile information and friendship information. The profile holds some user personal information such as name, home location, date of birth, gender, profession, interests, educational information, etc. This paper considers only four profile attributes and friendship information to unlock the critical fact of making a new friendship. A profile of four attributes: home location, profession, date of birth and interest.

2) Features Extraction: In this step, our proposed model can extract the information (home location, profession, date of birth and interest) from a social network by using some API.

3) Handling Missing Value: A social network consists of a large number of user profiles. Every user profile has personal information such as name, email, location, hometown, date of birth, personal interests, profession, gender etc. Some attributes have multiple values (e.g. interests = [programming, football, reading, gardening]). It is highly possible that some individuals do not possess all types of attributes, and some specific attribute's values may be missing in the profile. Thus, while comparing two user profiles, it is great to have all the information to measure similarity or dissimilarity. If one or more fields of a profile are missed, comparison cannot be performed, and hence it does not allow similarity among an individual's profile. For this reason, inferring missing profile items is an essential part of similarity measurement.

Handling missing values of a profile is just like making a data preprocessing. In [18], the authors proposed inferring personal items of a profile. The approach of calculating missing items is usually made by taking into account all of the information of friends or by searching a user's group membership. To disclose political views or sexual relationships can be obtained by accumulating all friends' profile information or group membership. In most cases, for security purposes, the social network does not allow this access, and it is not possible to extract all information of friends or group membership. Moreover, information cannot be extracted because some social users hide their sensitive information. As a result, some information about any user or his friends cannot be extracted. It is possible to retrieve user profile information with the help of social network $APIs$.

Considering all those limitations, we propose a method to overcome the problem of missing profile items. To infer a user profile information by using all of his/her friends profile information, we find the rank of each attribute's value of this missing profile item. The highest rank is considered as a value of this missing profile attribute. We have used a modified page rank algorithm to find out the highest rank of missing profile items. We calculate the vote of each friend according to his profile attribute value which is missed by a user u. The modified page rank formula is defined in Equation 1.

$$R(D_i) = P(D_i) + \frac{\sum_{j=1}^{n} \frac{P(D_i)}{1+\neg P(D_i^j)}}{n} \quad (1)$$

Here, $R(D_i)$ $P(D_i)$ refers rank and probability of a attribute respectively, $\neg P(D_i^j)$ represent probability of the other attributes except the attribute D_i of j^{th} user. Here, we add 1 with $\neg P(D_i^j)$ because of $\neg P(D_i^j)$ may zero when one user not connected with different profile attribute value user.

TABLE I. CONTRIBUTION OF DIFFERENT PAPER IN DIFFERENT PORTION

Existing Paper	Missing Item	Network Similarity	Determining Weight	Feature Automation
Akcora et. al. [1]	✓	✓	✗	✗
Msazhari et. al. [8]	✗	✗	✓	✗
Xin et al. [10]	✓	✗	✗	✓
Qader et al. [11]	✓	✓	✗	✓
Soni et al. [12]	✗	✓	✗	✗
Kumar et al. [13]	✗	✓	✗	✓
Al-Sabaawi et al. [14]	✓	✗	✗	✓
Shabaz et al. [15]	✗	✗	✓	✓
Berkani et. al. [16]	✗	✓	✓	✓
Razis et. al. [17]	✗	✗	✗	✓

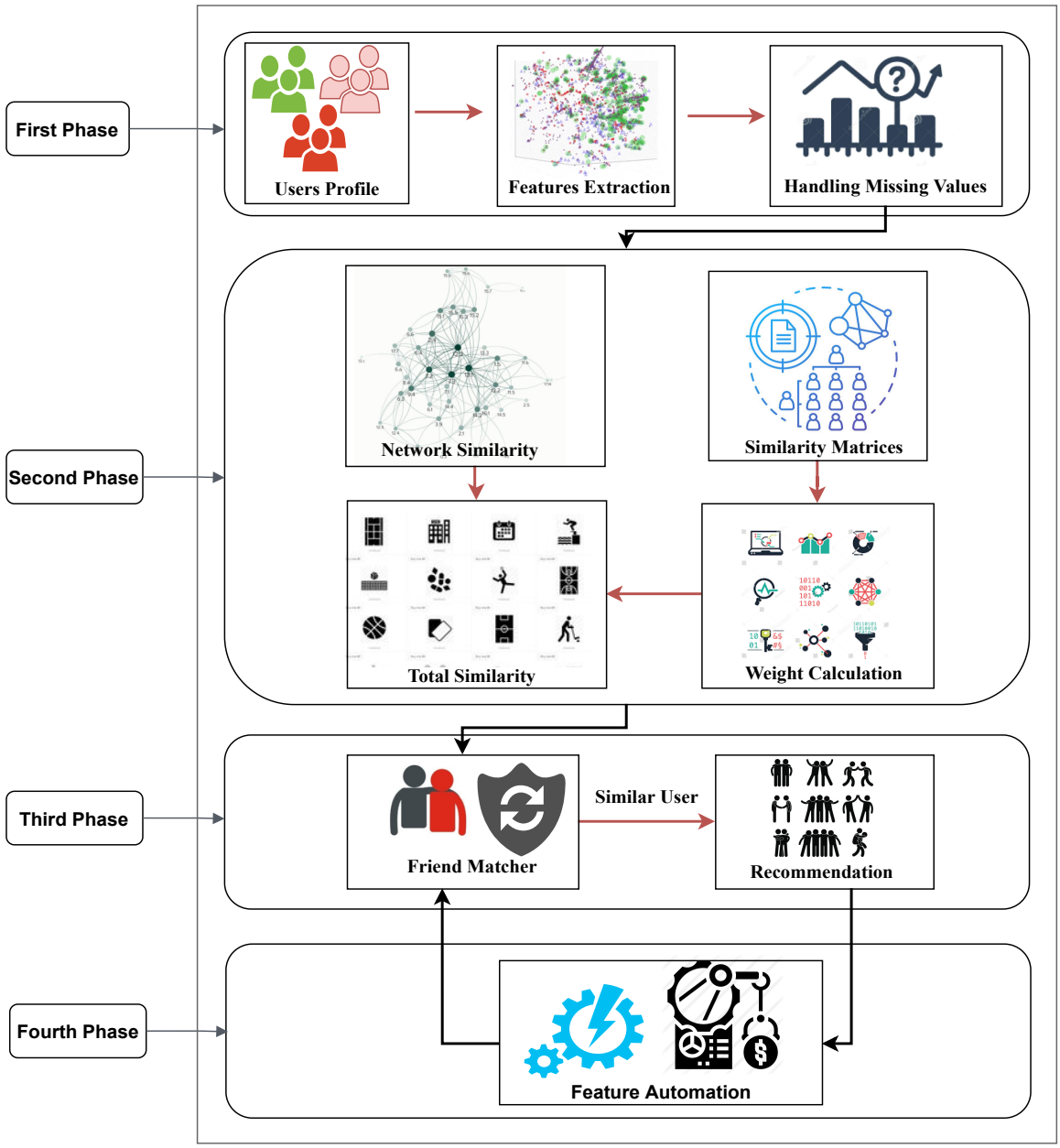


Fig. 1. Proposed Architecture for Recommendation System.

4) User Similarity: To measure similarity between two users u and v, where $(u, v) \in G.V$ in a considered network $G(E, V)$, where E represents friendship link and V represent user, we proposed a modified NS for network similarity and modified weight calculation algorithm for profile similarity, we use network and profile similarity for user recommendation.

We describe user similarity part in three phases: network similarity, profile similarity, friend matching.

A.1 Network Similarity: In a social network, people like to create new friendships with those closest to each other in the social graph. They create new edge and contact with those friends. Here we compute network similarity between two users u and v who are not friends and the closeness between u and v based on their information. Most effortlessly, network similarity can be computed by only using the number of mutual friends of u and v. In this approach, the only used node of friendship graph as a result in some case important information can be loosed, and the performance of network similarity is not good. In [1] used the edge of the social network mutual friendship graph for calculating network similarity between two node distances of two users. In mutual friends, there is no friendship relation between u and v and the mutual friends represent a complete graph called mutual friends graph as a subgraph of our target graph $G = (E, N)$, where N is the set of nodes and represent the social network users, the set of E represent the relationship among the social network user. In the previous measurement, they used node of mutual friendship graph, it cannot provide the relationship among mutual friends, but in mutual friend's edge set among the considered nods, includes all relationship information. Mutual friends graph formally defines as below:

Definition 1. *(Friendship Graph).* In a social network G, u is a node $u \in G.V$, friendship graph of u denoted as $FG(u)$ is a sub-graph of G where, $FG(u).N = \{u\} \cup \{n\}$, where $\forall n \in G.N, n \neq u, \forall e \in FG(u).E, e \in < u, n >$ and $FG(u, v).E = \{< x, x' >, \forall x \in FG(u, v).N, \forall x' \in FG(u, v)$.

Edges represent more information about friendship among considered users in a social network rather than nodes. Edge count provides how strongly the users tie each other. The existing work computes the similarity between u and v; they only compare the mutual friend's graph and the target user $u's$ friendship graph edge count. A friendship graph of u consists of $u's$ entire friend in a graph and all edges among the node. The formal definition of friendship graph is as follows:

Definition 2. *(Mutual Friends Graph).* In a social network G, u and v are two nodes $u, v \in G$, mutual friends graph of u and v, denoted as $MFG(u, v)$ is a sub-graph of G where, $MFG(u, v).N = \{u, v\} \cup \{FG(u).N \cup FG(v).N\}$ and $MFG(u, v).E = \{< u, x > \in G.E \cup < v, x > \in G.E\}, where x \in MFG(u, v).N, x \neq u, x \neq v$.

We use modified network similarity (NS) [1], the network similarity between two users in a social network graph can be computed by the ratio of the number of edges in the mutual friends' graph of u and v and the sum of the number of edge in the friendship graph of u and v.

Definition 3. *(Network Similarity).* Network similarity between two users u and v is defined as:

$$NS(u, v) = \frac{log(|MFG(u, v).E|)}{log(|FG(u).E + FG(v).E|)} \quad (2)$$

Where, $|MFG(u, v).E|$ represent the number of edges in $MFG(u, v)$ and $|FG(u).E + FG(v).E|$ represent the total number of edges in $FG(u)$ and $FG(v)$.

The existing method used only the ratio of $MFG(u, v)$ and $FG(u)$. In the proposed method, we used both u and $v's$ friendship graph edges count. Because of finding out the influence both u and v on mutual friends graph. If u and v have no mutual friends, mutual friends graph remain two nodes but no edge, i.e. $MFG(u, v).V = u, v$ and $MFG(u, v).E = 0$. In this case, the value of network similarity is zero.

A.2 Profile Similarity: Every profile contains some unstructured keywords, and these are associated with the user's details. The profile similarity between u and v can be calculated by measuring the similarity between the same profile items of two profiles. In this paper, we have taken four profile attributes: age, home location, profession, and interest to measure the similarity between profiles. Similarity between two users u and v depends on the similarity value between items $u_{interest}$ and $v_{interest}$, u_{age} and v_{age}, $u_{profession}$ and $v_{profession}$, $u_{location}$ and $v_{location}$. Profile items are heterogeneous, so it is harder to measure the similarity of different items by applying only one similarity measurement formula. However, there have some suitable similarity measures for every specific type of attribute. In this paper, three similarity measures are used to calculate the similarity between the items. Damerau–Levenshtein distance [19], Levenshtein distance [20] and Manhattan Distance [21], which are used to determine similarity of location, profession, age and interest.

A.3 Weight Calculation: Each profile consists of personal information such as age, interest, profession, location etc., and some standard measuring techniques are used to calculate the similarity between profiles attribute. In most existing techniques, only the information of two profiles of the recommended persons is used. Very few research works consider the influential factors of the existing social network to recommend new friends. It is more practical to measure the influence of profile attributes of the existing friends to predict new friend matching. In this case, the authors of [8] use a single set of some attributes to find out the influential factors of the existing network. Nevertheless, we consider both single and multiset attributes to calculate the influential factor of the current network. The multiset of attributes is often responsible for forming new friendship links in real life. For instance, two persons of the same age and interest are more likely to be friends than two persons with only the same age or only the same interest. In that sense, we introduce the concept of a multiset of attribute's comparison to recommend a new friendship link. Only the weight of each attribute is not sufficient for proper friend matching. For efficient fiend matching techniques, we need to compute all sets of attributes. Multiset attribute similarity can be calculated using the following Equation 3.

$$SV_{a_1, a_2 .... a_n} = SV_{\{a_1\}} \times SV_{\{a_2\}} \times ..... \times SV_{\{a_n\}} \quad (3)$$

5) Friend Matching Method: We have used a friend matching method (FMM) that calculates the similarity among two users u and v in two steps. Firstly, we calculate network similarity among two users (u & v) by applying Equation 2

and profile similarity using the following Equation 4. Secondly, it compares both profile and network similarity values with a threshold (TH). If one similarity value is greater than TH, it provides similarity between u and v otherwise dissimilarity. If both network and profile similarities are greater than the TH value, it provides a strong similarity between them. The probability of new link formation increases with the similarity value.

$$PS = \sum_{i=1}^{n} W_i * SV_i \tag{4}$$

In this equation n is the total attributes, W_i = Weight of i^{th} attribute set (e.g. $W_1 = W_{\{age\}}$,$W_2 = W_{\{location\}}$, $W_4 = W_{\{age,location\}}$). $SV_i = i^{th}$ attribute similarity between u and v (e.g. $SV_1 = ageSimilarity(u_{\{age\}}, u_{\{age\}})$ $SV_2 = ageSimilarity(u_{\{location\}}, u_{\{location\}})$). All the similarities among the user will be calculated by using this equation.

In new friendship formation, users in an OSN observe profile attributes or mutual friends or both profile and mutual friends. If there is a similarity in profile attributes, it provides a probability of creating a new friendship link. Besides, a sufficient number of mutual friends provides a possibility to create new friendship links. Moreover, network similarity significantly impacts friendship formation when a user does not update his profile information.

Profile similarity (PS) and network similarity (NS) are independent. Both have an individual influence on user similarity. We use conditional probability To calculate the effect of both NS and PS on user similarity.

Here, total similarity value, $TSV = NS + PS$

$$P(PS|TSV) = \frac{P(PS)}{P(TSV)} \tag{5}$$

$$P(NS|TSV) = \frac{P(NS)}{P(TSV)} \tag{6}$$

$$\neg P(PS|TSV) = 1 - \frac{P(PS)}{P(TSV)} \tag{7}$$

$$\neg P(NS|TSV) = 1 - \frac{P(NS)}{P(TSV)} \tag{8}$$

Here, $P(PS|TSV)$ refers to the conditional probability of how much PS affects TSV, and $\neg P(PS|TSV)$ provides PS not effect to TSV. Calculating user similarity, we rank for each user who does not have a friendship link and a more significant user similarity value than a threshold value. The top of the rank table has the highest similarity. From the rank, table select top k user and recommends to u.

6) Feature Automation: In the case of recommendation of users, there will create a new friendship link. However, in all cases, all recommended users will not be able to create friendship links. For this case, a feature automation technique is introduced here to extract newly friendship link created user information. Firstly, this technique analyzes collected information and calculates each pair of user profile similarity using our three considered similarity measures to calculate attribute set similarity. When several profile pair similarities are calculated, it measures each attribute set's weight using Equation (3) to (8). The effective weight for each attribute set tries to exact the hidden fact that primarily influences creating new friendship links. It compares this calculated weight with the previous weight and which set is more affected and less. According to this decision, it updates attributes sets weight. Friend Matching Method uses the weight information and measures user similarity, and users will be recommended appropriately, and the outcome is better than previous.

B. Proposed Algorithms

In our research worked we approached three algorithms for the recommendation system. The first algorithm computed the missing profile values of the OSN users. Another two algorithms are moderated algorithms of [8]. Algorithm 2 is used to compute the weight of all conceivable sets of contemplated profile attributes. Algorithm 3 is used to estimate the network and profile similarity among the users, then recommended the future friend for the user.

1) Disappeared Value Estimation Technique: Our proposed Algorithm 1 can compute the missing items of the users' profiles. It is mainly a data preprocessing method. To predict the missing profile item of any user, this algorithm firstly discovers the missing profile items. In the algorithm, line number 5 indicates that the user's missing profile item is calculated and a probability computation function $CalculateProbability(p_{ij}, P_uR_i)$ is called in line 6. This function is mainly all friends missing attributes and compute the possibility of items and finally, calculate the largest estimation of item's value for that disappeared item. In our previous work [22], we described more about this algorithm.

Algorithm 1: Prediction of Profile Disappeared Value

```
   Input  : P_u = {p_1, p_2, p_3, .....p_n}      //users profile
   Output: Predicting disappeared values of every profile
1  D =< Location, Interest, Age, Gender >
2  foreach p_i ∈ P_u do
3      P_uR_i =Extracting friends of p_i
4      if P_i[D_k] = NULL then
5          foreach j ∈ P_uR_i do
6              if p_ij = NULL then
7                  pF_i = FriendsofP_i
8                  foreach j ∈ P_uR_i do
9                      X_c = Counting D_k not equal
                         P_uF_i^j inMFG(P_uF_i^j, P_i)
10                     Y_v = Vector of pair number
11                     Discover Rank of X_c      //using Equation (1)
12                     Y_v.Push(Rank)
13                 end
14             end
15         end
16         Extraction Max(Y_v)
17         Discover P_uF_i[D_k] for maximum value
18     end
19 end
```

2) Feature Weight Computation System: Algorithm 2 was utilized to compute the weight of every set of attributes. In algorithm 2, lines 5 to 7 indicate that Manhattan Distance, Levenshtein Distance and Demaru Levenshtein Distance are used to compute the similarity of every user profile attributes with his all friends profile. To calculate the similarity of the profile attributes (profession, interest, & location), Levenshtein Distance and Demaru Levenshtein Distance are used. And calculated the 'age' similarity of the profile with the help of Manhattan Distance. In line 6 was used for calculating user

attributes similarity and took the better value of similarity. In lines, 8 to 19 are used to compute the similarity for every user.

Algorithm 2: Feature Weight Computation System (FWCS)

Input : $P = \{p_1, p_2, p_3, ...p_n\}$
Output: Weight of each feature
1 $p_i =< profession, age, interest, location >$
2 $i \leftarrow 1$
3 **foreach** $p_i \in P$ **do**
4 PR_i =Extract friends of p_i
5 **foreach** $j \in PR_i$ **do**
6 $DLD_{location} +=$ $DemaruLevenshteinDistance(p^i_{location}, PR^{ij}_{location})$
$DLD_{interest} +=$ $DemaruLevenshteinDistance(p^i_{interest}, PR^{ij}_{interest})$
$DL_{profession} +=$ $LevenshteinDistance(p^i_{profession}, PR^{ij}_{profession})$
$LD_{location} +=$ $LevenshteinDistance(p^i_{location}, PR^{ij}_{location})$
$LD_{interest} +=$ $LevenshteinDistance(p^i_{interest}, PR^{ij}_{interest})$
$DM_{age} += ManhattanDistance(p^i_{age}, PR^{ij}_{age})$
$DL_{profession} +=$ $LevenshteinDistance(p^i_{profession}, PR^{ij}_{profession})$
7 **end**
8 **if** $DLD_{location} > LD_{location}$ **then**
9 $D^i_{location} = \frac{DLD_{location}}{|PR_i|}$
10 **else**
11 $D^i_{location} = \frac{LD_{location}}{|PR_i|}$
12 **end**
13 **if** $DLD_{interest} > LD_{interest}$ **then**
14 $D^i_{interest} = \frac{DLD_{interest}}{|PR_i|}$
15 **else**
16 $D^i_{location} = \frac{LD_{location}}{|PR_i|}$
17 **end**
18 $D^i_{age} = \frac{DM_{age}}{|PR_i|}$
19 $D^i_{profession} = \frac{DL_{profession}}{|PR_i|}$
20 **end**
21 $W_{\{s\}} = \frac{\sum D^i_{\{s\}}}{|P|}$ /* Calculated weight for each attribute set, here, s represent attribute set*/

We consider all probable profile attributes sets to calculate the weight in our suggested algorithm, but in the paper [8], authors are only considered profile attributes to calculate the weight. On the friend matching, this weight of profile attributes creates various impacts. The presence of attributes is called an attribute set. The multiplication process is used to calculate the set value of attributes. Example: multiplication of $similarity_{\{profession\}}$ and $similarity_{\{gender\}}$ is produce $similarity_{\{profession,gender\}}$. As we know that, if we multiply two positive numbers, the result will be too smaller if both numbers are less than one. So, the attribute similarity set value becomes small if any attribute similarity is small in the set, and also, the profile attribute similarity value will be too smaller. The average value of all set similarities values is called profile similarity value. In that circumstance, dissimilar profile values will be nearest to 0, and similar profiles will be nearest to 1. By this technique, the proposed algorithm can easily discover the dissimilar and similar profiles.

3) Friend Matching System: We calculated user similarity among pairs of users (u & v) using Algorithm 3 after measuring every attribute set's weight.

In this proposed algorithm, lines 3 to 6 indicates that every profile attributes similarity is calculated. LocationSimilarity mentions the most suitable similarity value among Levenshtein and Demaru-Levenshtein similarity. NS and PS are calculated in lines 7 and 8. It generates a rank of similarity based on avg. of PS and NS; if the TH (threshold) is lowest than both NS and PS or one of them. From every user rank table, the topmost k users are recommended.

Algorithm 3: Friend Matching System (FMS)

Input : $P = \{p_1, p_2, p_3, ...p_n\}$ //from Algorithm 2
Output: Rank of the similar user
1 $p_i =< F^i_{age}, F^i_{gender}, F^i_{location}, F^i_{interest} >$
2 **foreach** $i \notin P$ *and* $j \notin PR_i$ **do**
3 $SV_{age} = ManhattanDistance(F^i_{age}, F^j_{age})$ //SV refers similarity value
4 $SV_{profession} =$ $LevenshteinDistance(F^i_{profession}, F^j_{profession})$
5 $SV_{location} = LocationSimilarity(F^i_{location}, F^j_{location})$
6 $SV_{interest} = InterestSimilarity(F^i_{interest}, F^j_{interest})$
7 $NS = NetworkSimilarity(p_i, p_j)$ // using Equation (2)
8 $SV\{s\} \leftarrow \{profession, interest, location, age\}$ // using Equation (3)
9 $PS = \sum W_{\{s\}} * SV_{\{s\}}$ // using Equation (4)
10 $TSV = NS + PS$ //Total similarity value
11 $P(TSV|PS, NS) =$ $\left[\frac{P(PS|TSV)*P(NS|TSV)*P(TSV)}{P(PS|TSV)*P(NS|TSV)*P(TSV) - \neg P(PS|TSV)*\neg P(NS|TSV)*\neg P(TSV)}\right]$ /* using Equation (5) to (8) */
12 **if** $P(TSV|PS, NS) > TH$ **then**
13 $Profile^i_j = P(TSV|PS, NS)$ //TH is the threshold value
14 **end**
15 **end**
16 $Rank_i \leftarrow sort(Profile_i)$ /* Users recommended based on rank of similarity */

IV. Experimental Results and Discussion

Here, we represent our experimental results and discuss the performance evaluation of our proposed algorithms. We used a real OSN dataset to evaluate the performance and examine the experimental results from different angles. Proposed algorithms are developed in the C++ language.

A. Dataset Collection and Description

In our experiment, we used the Facebook dataset and collected it from [23]. In this dataset, we discovered an accurate OSN undirected friendship graph. Four thousand (4K) different users and eighty-eight thousand (88K) edges are available in our dataset. Friendship is defined through searching for an edge among the pair of users. Many to many relationships for every user has been calculated on the whole dataset. Moreover, a vast number of similarity values are created from the dataset. We can compare it to an identical matrix where the matrix's upper and lower parts have the equivalent value. So, we calculated only one part from these two parts. As a result, complexity is decreased. Our experimental results on this dataset illustrated in the Table II.

B. Factor Coefficient

It is necessary to understand which items are more important for users, so considering the characteristics of similarity among all users of friends and the average similarity of each set of features has calculated from our considered dataset shown Table III with their coefficient. Each feature's coefficient is essential for friend matching. The Friend Matching Method

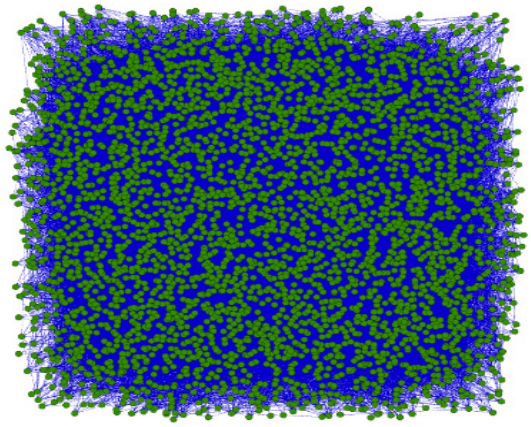
TABLE II. STATISTICS OF SOCIAL NETWORK

Hyper Parameter	Weight
Nodes	4039
Edges	88235
Nodes in largest WCC	4039 (1.00)
Edges in largest WCC	88235 (1.00)
Number of triangles	1612010
Fraction of closed triangles	0.2647

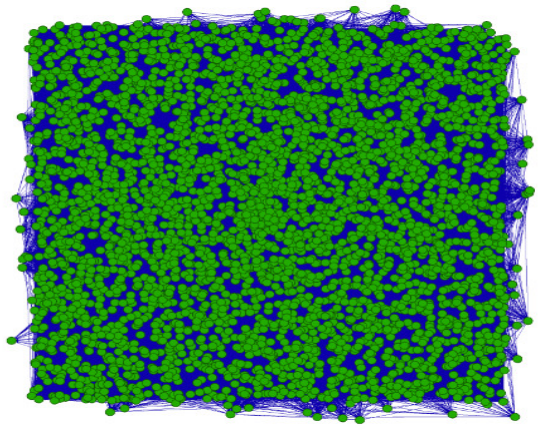
calculates the similarity between two users multiplying with this corresponding item set weight in friend matching.

C. Result and Analysis

Consequently, more than 55% of new similar edges have been established by using our proposed method, which is shown in Table IV. The experimental result is also inspiring when we applied it to the custom dataset. In Fig. 2(a), we show the actual social network friendship graph, and in Fig. 2(b) we additionally showed the resulting graph obtained after processing the dataset by our recommended method.



(a) Real Dataset Social Network Friendship Graph.



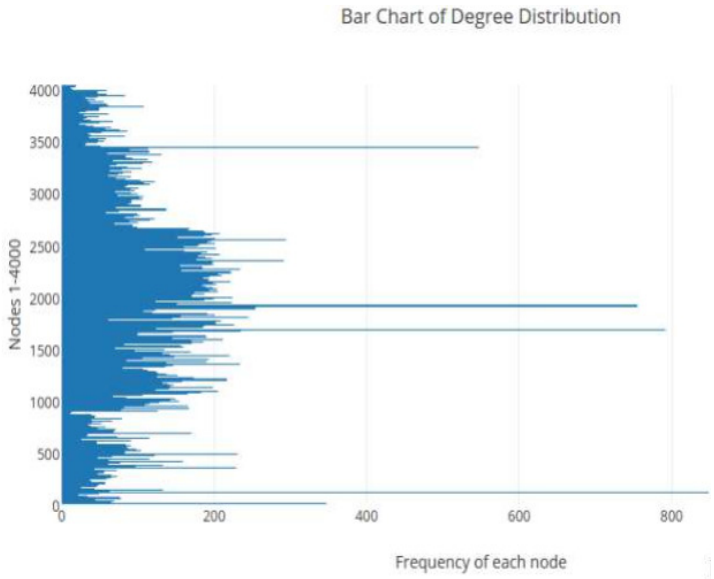
(b) Social Network Friendship Graph after Applying Proposed Formula.

Fig. 2. Friendship Graph.

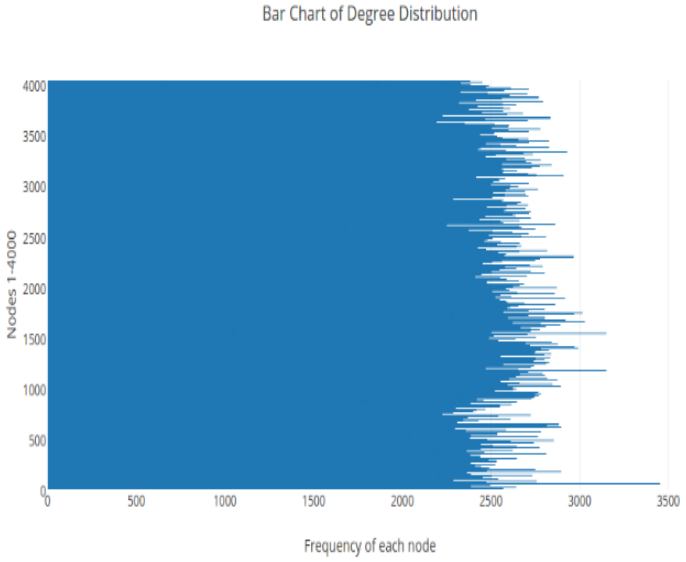
D. Comparison of Degree

When we compare two distinct graphs, it is necessary to consider degree comparison. The average degree in the two undirected graphs is shown in Table V. on an average, only 43 edges are discovered before the network processing but using our method, 1953 edges are discovered after the processing.

For every complex network showing degree, the comparison is very fundamental and must be given case. To explain more, we have included the histogram of the frequency distribution of each node to describe the degree. From that evidence, we can be able to compare more efficiently. The Fig. 3(a) show real case of dataset and Fig. 3(b) shows the processed case of our dataset.



(a) Degree Distribution of Real Dataset.



(b) Degree Distribution of Processed Dataset.

Fig. 3. Degree Distribution.

From the above evidence, we can normalize the idea that

TABLE III. FACTORS COEFFICIENTS

Factor	{l}	{i}	{a}	{p}	{l,i}	{l,a}	{l,p}	{i,a }	{i,p}	{a,p }	{l,i,a}	{l,i,p}	{i,a,p}	{l,a,p}	{l,i,a,p }
Weight	0.791	0.440	0.741	0.588	0.351	0.588	0.471	0.338	0.286	0.457	0.271	0.231	0.191	0.345	0.152
Coefficient	13%	7%	12%	10%	6%	9%	8%	5%	4%	7%	4%	4%	3%	6%	2%

TABLE IV. STATISTICS OF SOCIAL NETWORK EDGE

	Before	After	Increasing %
Number of Edges	88,235	4,809,284	55.50

TABLE V. COMPARISON OF DEGREE

	Before	After
Networks Type	Undirected	Undirected
Degree(Avg.)	43.691	1953.98

the small network number of groups can be suggested before processing. After processing, each group may have a higher number of profiles that will give more accurate friendship results or, in other words, a more efficient friendship network will produce. According to the presented tables and figures in the previous section, we can say that the proposed method has significantly impacted the friend matching system for this dataset.

As a further study of this experimental result, firstly, we divided it into three classes (A, B, C) based on the range of similarities for the friendships set and representing the ranges 0.0–0.49, 0.50–0.70, 0.71–1.00, respectively. Recommended friendships between the A, B, C classes are shown in Fig. 4. According to the calculation of similarities proposed recommended system recommended 39% future friends in class A, 50% in class B, and 11% in class C for the user. So, it's proof that the proposed approach effectively recommended more new relationships for the user. Furthermore, it delivers satisfying independence for other recommended friendships, which are not in the actual friendship graph.

E. Calculation of Accuracy

To estimate the accuracy, we used the probabilistic technique. Firstly, we randomly took few edges whose similarities measured and friends in an existing social network. Finally, from the existing network, we erased those edges. In the first scenario, 0.7157017 is the average similarity value, and 0.67801 is the average similarity value after erased the edges. So, after calculation, we see that our proposed algorithm accuracy is 97%. That indicated 97% users in real friendships graph and our proposed recommendation system can be capable of recommending the future friends for the users and 3% of real friendship which our proposed system has not recommended. Calculation of error rate is shown in Table VI.

F. Clustering Coefficient Comparison

The clustering coefficient is a prevalent measure for such real networks, especially for the social network, which shows

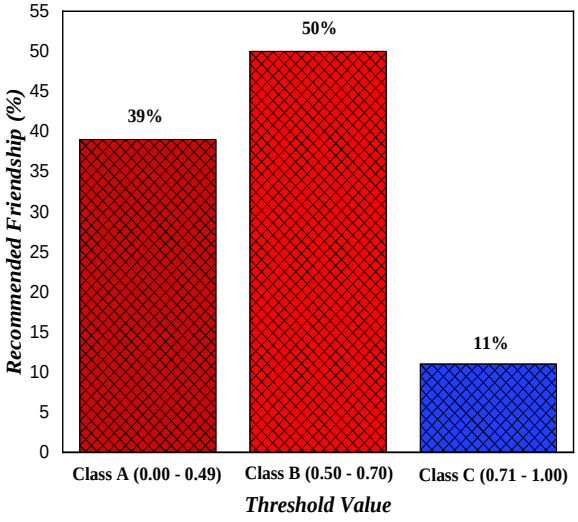


Fig. 4. User Friendships Distribution between the Three Classes.

TABLE VI. CALCULATION OF ERROR

Similarity Value	Before	After	Difference	Error %
Avg.	0.7157017	0.67801	0.0376917	3%(almost)

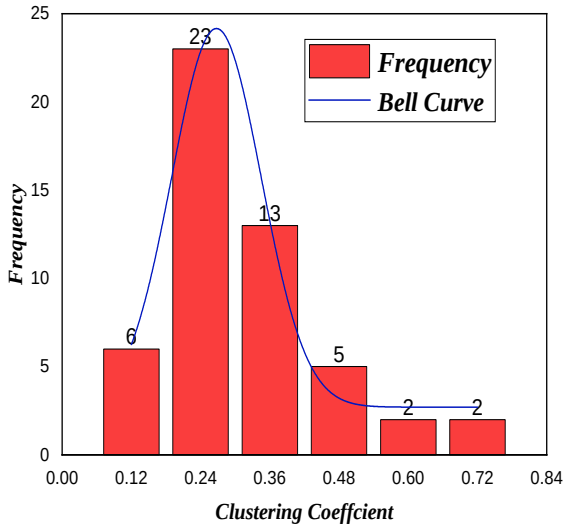
how tightly bond groups exist in the network. Here in our approach, we use a small dataset to examine the clustering coefficient. The result is also promising. Average clustering coefficient comparison is shown in Table VII. For the local version of the clustering coefficient where it is calculated for each node and the values are shown by the following Fig. 5(a) where shows the scenario of before processing and the Fig. 5(b) shows the scenario of after processing and adding more edges.

TABLE VII. CLUSTERING COEFFICIENT COMPARISON.

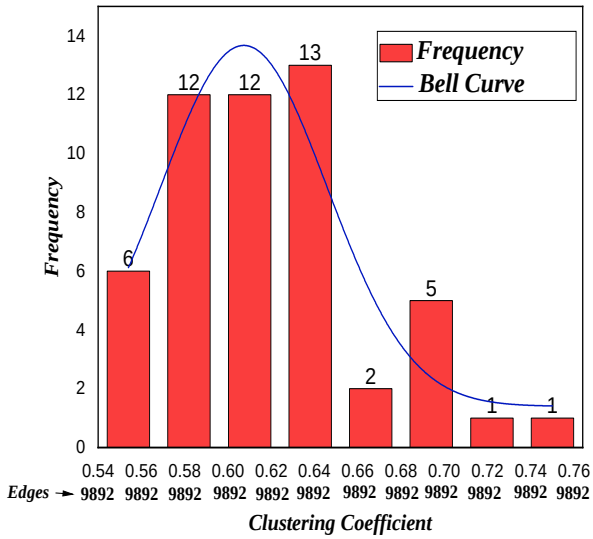
	Before	After
Clustering Coefficient Avg.	0.247242735478	0.632683188902

G. Comparison Assessment

In addition, to evaluate the results more accurately, we calculated the number of true positives (nTP) and the number of false-positive (nFP). More formally in Equation 9.



(a) Clustering Coefficient Frequency for Each Node at First Case (Before Processing).



(b) Clustering Coefficient Frequency for Each Node at Second Case (After Processing).

Fig. 5. Clustering Coefficient Frequency for Each Node.

$$\textbf{Precision value:}\frac{nTP}{(nTP+nFP)} \tag{9}$$

Thus the higher value gives better precision. Moreover, the true positive (TP) rate refers to how many real friendships have been accurately suggested by the recommended system. Similarly, the false positive (FP) rate refers to the number of actual friendships the system has not suggested. The confusion matrix is used to calculate these rates. In Table VIII shows the confusion matrix.

We have calculated multiple comparisons for our dataset. Considering the threshold value of the real dataset, the false positive rate is 0.00260085, and the true positive rate is 0.0165362. Moreover, 0.96566 is the precision value. That indicates 97% users on our dataset are recommended accurately by our proposed system.

Moreover, to compare the proposed model with other existing methods, we have re-implemented the models used in [8, 10, 11, 12, 13, 14, 16] as accordance with the description in the paper to make a fair comparison. All the models are evaluated on the same data set to ensure the fairness of the comparison. In Table IX shows a comparison of our system with some existing works that we have re-implemented on the same data set that we have used.

Graphical comparisons of proposed model with [8], [10], [11], [12], [13], [14], and [16] has shown in Fig. 6.

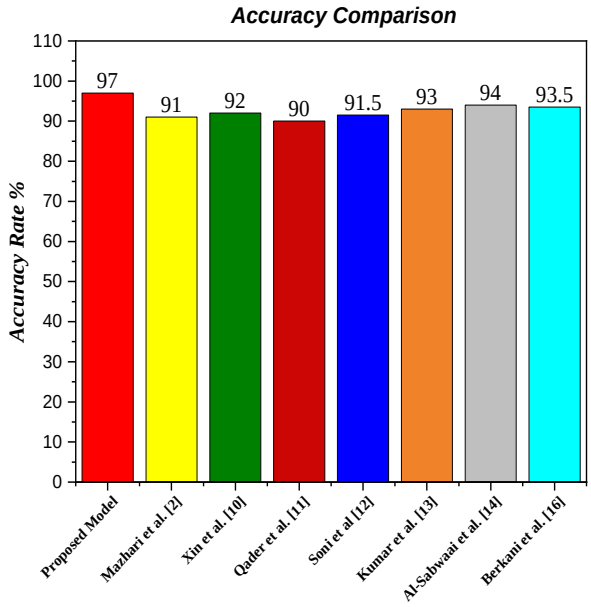


Fig. 6. Graphical Comparison of Proposed Model with [8], [10] [11], [12], [13], [14], and [16] which is Re-implemented on our Data Sets.

The ROC curve of Fig. 7 shows the true positive and the false positive rate for different threshold value.

V. Conclusion

This paper's fundamental achievement is designing and developing a user recommendation system consequent to profile attributes and network connection. This paper proposes an effective method to calculate the maximum number of similar users from a social network. Sometimes profile attributes of the user are hidden or missed, but this hidden or missed attributes data can affect profile similarity calculation. Most of the existing methods are failed to solve this problem, but our proposed approach solves it and gives better performance. Our proposed recommendation method achieved 97% accuracy and 96.566% precision which means the system properly recommended future friends for the user in the social network.

In the future, we will spread this work by combining attributes, for example, shares, comments, likes, pictures, status,

TABLE VIII. CONFUSION MATRIX

		Predicted Class		
		Positive	Negative	Total
Actual Class	Positive	True Positive (TP)	False Positive (FN)	$P = TP + FN$
	Negative	False Negative(FP)	True Negative(TN)	$N = FP + TN$
	Total	$P' = TP + FP$	$N' = FN + TN$	$S' = P + N = P' + N'$

TABLE IX. ACCURACY COMPARISON

Authors	Year	Accuracy Rate %
Mazhari et al. [8]	2015	91%
Xin et al. [10]	2020	92%
Qader et al. [11]	2020	90%
Soni et al. [12]	2020	91.5%
Kumar et al. [13]	2018	93%
Al-Sabaawi et al. [14]	2018	94%
Berkani et al. [16]	2021	93.5%
Proposed Method	-	**97%**

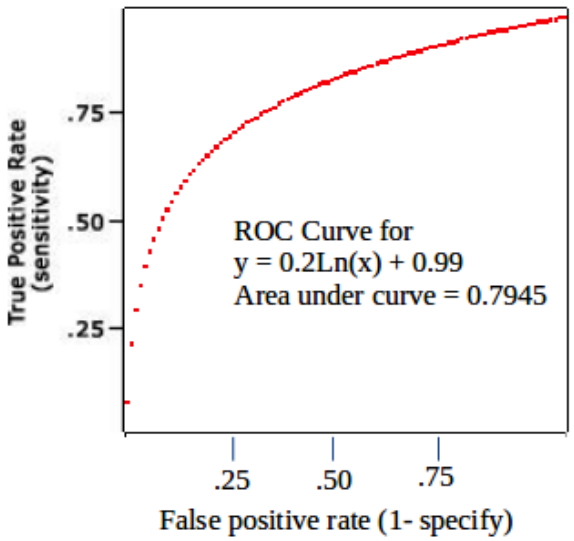


Fig. 7. ROC Curve of our Proposed FMS Algorithm.

etc. By combining those kinds of attributes with our method, we will be capable of calculating the sentiment analysis of OSN users and quickly identify and observe the illegal activities in the online social network.

REFERENCES

[1] C.G. Akcora, B. Carminati, E. Ferrari, User similarities on social networks, Social Network Analysis and Mining 3 (3) (2013) 475–495.

[2] M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a feather: Homophily in social networks, Annual review of sociology 27 (1) (2001) 415–444.

[3] T. Kaya, H. Bicen, The effects of social media on students' behaviors; facebook as a case study, Computers in Human Behavior 59 (2016) 374–379.

[4] P. Bhattacharyya, A. Garg, S.F. Wu, Analysis of user keyword similarity in online social networks, Social network analysis and mining 1 (3) (2011) 143–158.

[5] D. Yang, C. Huang, M. Wang, A social recommender system by combining social network and sentiment similarity: A case study of healthcare, Journal of Information Science (2016) 0165551516657712.

[6] R. Buettner, Predicting user behavior in electronic markets based on personality-mining in large online social networks, Electronic Markets (2016) 1–19.

[7] C. Cai, H. Xu, A topic sentiment based method for friend recommendation in online social networks via matrix factorization, Journal of Visual Communication and Image Representation 65 (2019) 102657.

[8] S. Mazhari, S.M. Fakhrahmad, H. Sadeghbeygi, A user-profile-based friendship recommendation solution in social networks, Journal of Information Science 41 (3) (2015) 284–295.

[9] Y. Yang, J. Pei, A. Al-Barakati, Measuring in-network node similarity based on neighborhoods: a unified parametric approach, Knowledge and Information Systems (2017) 1–28.

[10] M. Xin, L. Wu, Using multi-features to partition users for friends recommendation in location based social network, Information Processing & Management 57 (1) (2020) 102125.

[11] S.A. Qader, A.R. Abbas, Dual-stage social friend recommendation system based on user interests, Iraqi Journal of Science (2020) 1759–1772.

[12] M.T.S.S. Soni, P. Singhai, Friend recommendation system using machine learning method, Journal of Scientific Research & Engineering Trends, vol. 6, Issue 5, 2020.

[13] P. Kumar, G.R.M. Reddy, Friendship recommendation system using topological structure of social networks, in: Progress in Intelligent Computing Techniques: Theory, Practice, and Applications, Springer, 2018, pp. 237–246.

[14] A.M.A. Al-Sabaawi, H. Karacan, Y.E. Yenice, Exploiting implicit social relationships via dimension reduction to improve recommendation system performance, PloS one 15 (4) (2020) e0231457.

[15] M. Shabaz, U. Garg, Shabaz–urvashi link prediction (sulp): A novel approach to predict future friends in a social network, Journal of Creative Communications 16 (1) (2021) 27–44.

[16] L. Berkani, S. Belkacem, M. Ouafi, A. Guessoum, Recommendation of users in social networks: A semantic and social based classification approach, Expert Systems 38 (2) (2021) e12634.

[17] G. Razis, I. Anagnostopoulos, Discovering similar twitter accounts using semantics, Engineering Applications of Artificial Intelligence 51 (2016) 37–49.

[18] J. Lindamood, R. Heatherly, M. Kantarcioglu, B. Thuraisingham, Inferring private information using social network data, in: Proceedings of the 18th international

conference on World wide web, ACM, 2009, pp. 1145–1146.

[19] E. Brill, R.C. Moore, An improved error model for noisy channel spelling correction, in: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2000, pp. 286–293.

[20] B. Cao, Y. Li, J. Yin, Measuring similarity between graphs based on the levenshtein distance, Appl. Math 7 (1L) (2013) 169–175.

[21] M. Gardner, Taxicab geometry, in: The Last Recreations, Springer, 1997, pp. 159–175.

[22] M.A. Islam, L. Islam, Calculation of client similarities in large-scale on social network using recommendation framework, in: 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), IEEE, 2019, pp. 679–684.

[23] J. Leskovec, A. Krevl, SNAP Datasets: Stanford large network dataset collection, `http://snap.stanford.edu/data` (Nov. 2020).

An Open-source Wireless Platform for Real-time Water Quality Monitoring with Precise Global Positioning

Niel F. Salas-Cueva[1], Jorch Mendoza[2], Juan Carlos Cutipa-Luque[3], Pablo Raul Yanyachi[4]
Electronic Engineering Professional School
Universidad Nacional de San Agustín de Arequipa, 04000, Arequipa, Peru[1,2]
Pedro Paulet Institute for Astronomical and Aerospace Research
Universidad Nacional de San Agustín de Arequipa, 04000, Arequipa, Peru[3,4]

***Abstract*—Sustainable development associated with the agricultural field of Arequipa, a region in economic growth, is vulnerable to contamination of water resources, putting production systems and food security at risk. Therefore, it is necessary to implement an automated system to control, management, and monitor this vital resource. The proposed work proposed a system to measure water quality monitoring in reservoirs and lakes with high accurate related to global positioning. It includes an embedded computer, multiparameter sonde, and an additional dual GNSS/INS in hardware architecture. The software architecture is fully open-source with compatibility, modularity, and interoperability features between Python and MySQL, allowing data management for real-time data in visual interface on a platform that stores unlimited data logging, monitors and analyzes. The proposed system is validated in an experimental test that measures the water quality of a huge agricultural reservoir, where certified instrumentation is mandatory, as compared to other methods used locally for this action.**

Keywords*—Open-source; water quality monitoring; real-time; python; visual interface; MySQL; dual Global Navigation Satellite System (GNSS); Inertial Navigation System (INS); multiparameter sonde

I. Introduction

Arequipa (latitude -16.3988, longitude -71.535, and altitude 2335 m.a.s.l.) is one of the main economic regions in Peru with approximately 1,497,438 inhabitants. It is also one of the fastest growing regions in the country, stands out for mining and agricultural activities, which generate some conflicts in the management of water resources and those related to water quality [1]. This entails many risks and losses in all aspects of sustainable development in the urban-rural area that surrounds the region; either due to the misuse of chemical products related to these industrial activities and others [2].

The objective of this work is to develop an open-source integrated system for measuring water quality with accurate positioning. A dual GNSS device is used in conjunction with the depth meter embedded in the sonde. Electronic systems condition the signals from the sensors and the positioning system to send them to a high-performance computer. Measurement data is stored in memory in structured query language (SQL) database, to be later transmitted remotely through the 802.11a Wi-Fi network. In addition, a multiplatform interface is developed for data visualization in real-time and in offline mode.

The paper is organized as follows: Section 1 presents a brief resume of related works in integration system for remote water quality measurement; Section 2 presents the system description in hardware and software; Section 3 presents the validation of integrated system when measured a huge agriculture water reservoir; and Section 4 gives the conclusion.

II. Related Work

The study of water quality is carried out by conventional methods such as taking samples in appropriate containers for subsequent analysis in the laboratory, methods that require high time, cost and human resources [3]. Another method includes specialized sensors that measure parameters such as pH, conductivity, salinity, turbidity, etc. in-situ [4], [5]. These sensors, embedded in a multiparameter sonde, serve to measure large volumes of water (lagoons, reservoirs, coastal waters) using commonly manned vessels to collect data in real-time and to save in a data-logger or in a computer. Measurement position is commonly provided by additional global positioning system (GPS)[6]. Some manufactures have developed their smartphone application in order to provide the positioning based in the smartphone global positioning system. Another research combines inertial navigation system with precise global positioning system [7]. High computational hardware is increasingly used to improve the accuracy of these systems and to use big data analysis [8]. Researchers have been developed systems with low-cost sensor and ZigBee low range wireless communication to measure water quality in aquaculture [9]. Another solution to measure wide areas of water uses multiple set of sensors replicated in different location to form a wireless sensor network [10].

Water quality monitoring systems use information technologies, such as human-machine interfaces, databases, structured programming, facilitating the visualization and alert of the measured parameters. In [11], the authors present the application of these techniques in reservoirs that feed large hectares of agricultural land. Among these information technologies, the use of free software, such as Python, MySQL and Grafana stands out, which have shown promise in similar applications [12], [13]. Regarding the database and wireless remote sensing, in [14], the authors present a remote monitoring system applied to the management of a bridge with measurements of parameters such as voltage, current, positioning, images, etc. The open-source hardware

and software tackle many problems related to security, data management, flexibility, analysis, and low power. They have shown successfully implementing process and configuration related monitoring wireless sensors in similar application [15]. There are several researchers including low-cost sensors for water quality measurement systems [16], [17]. However, the environmental agencies and industry prefer the use of certified instrumentation and precise positioning to validate their analyses [18], [19].

III. Materials and Methods

The laboratory for complex control process and unmanned vehicles at the 'Universidad Nacional de San Agustín de Arequipa' is leading research projects related to the development of enhanced systems for ocean and nature supervision [20], [21]. This paper describes a modular system to be employed in measuring water quality with high accuracy, both in unmanned and manned vehicles.

The proposed system shows the interaction between user and water as a liquid element for its remote monitoring; aided by dual GNSS, inertial sensors, database management system and other resources as shown in Fig. 1. The goal of the system is to quantify water quality through a multiparameter sonde with data metrics involving static measurement with high precision and dynamic measurement with additional data of position and velocity. Display and format metric data should be available for all measurements, both offline and real-time. High performance and great coverage are required in order to measure reservoirs, lakes and coastal waters. All these requirements are correlated to achieve the proposed goal.

One of the requirements is safe and fast connectivity, for which we will have a point-to-point wireless link with a coverage range greater than 15 km away. The wireless connection obtains continuous and synchronous data to study and analyze the water quality. Unwanted chemical parameters can be diagnosed in real-time in a certain area. Commonly, certified instrumentation uses industrial communication protocols and additional equipment for monitoring and control.

Fig. 2 presents the block diagram of the proposed system where the embedded computer, the GNSS/INS device and the multiparameter sonde are hardware communicating each other trough wired protocols, the wireless protocols enable a communication from a ground station that generally is a laptop or a smarthphone. Another software are related to each component in this block diagram and will be explained in subsequent sections. Fig. 3 presents the system description of the monitoring system that consists of the Aquatroll 600 model multiparameter sonde, a Jetson TX2 high performance computer, a VN-300 as the dual GNSS/INS (global positioning system with inertial measurement unit), a TP-Link router CPE510, and a platform interface for data visualization running on a smartphone laptop or similar to a ground station. The sonde communicates with the embedded computer through the Modbus industrial protocol. The embedded computer receives data from the GNSS/INS synchronizing it with the data received from the sonde. The sonde and the dual GNSS/INS data are stored in the embedded computer memory. Using a laptop or a smartphone, and through a wireless link, the user can remotely access to the embedded computer in order to set and read the data. A graphical user interface is created for data visualization with indicators and alerts, both offline and real-time.

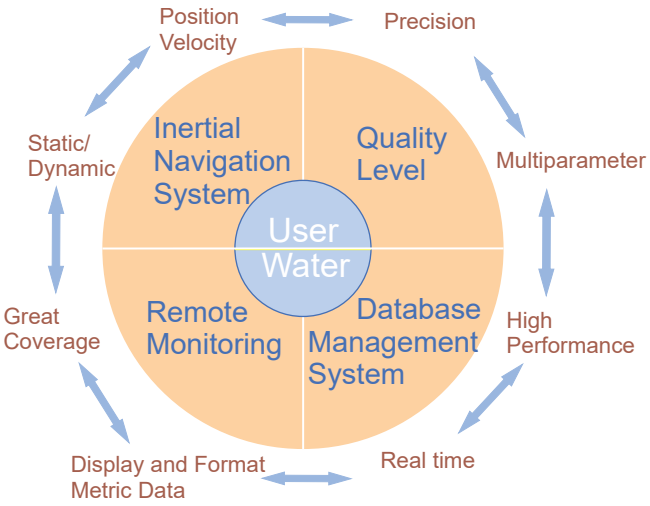


Fig. 1. Requirements of the Integrated System to Measure Water Quality.

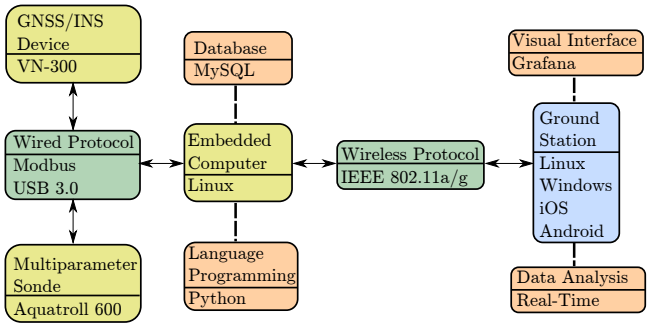


Fig. 2. Block Diagram of the Water Quality Measurement System.

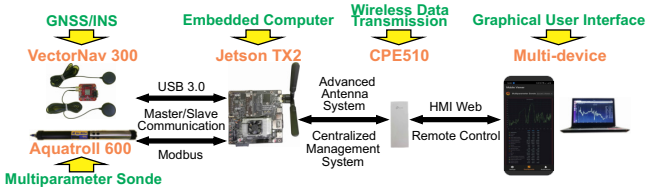


Fig. 3. System Integration of the Water Quality Measurement.

A. Software Architecture

The main software is based on Python open-source language programs with free libraries and modules, such as 'PyMySQL' and 'PyModbus' to connect databases and remote sensors in the well known industrial Modbus communication protocol. The Python codes communicate, configure, convert data types, compute mathematical equations, administrate and save data in a computer memory. Fig. 4 describes the drivers of this software architecture, the remote access is carried out using secure shell protocol (SSH) to enable local devices running on Windows, Linux, iOS, Android, etc. The communication is set with permissions according to the user

profile, from reading to editing the existing Python codes. The 'PyMySQL' module is a Python library for MySQL clients based on Python enhancement proposal (PEP) 249, that uses a high level application programming interface (API) named 'mysql.connector' for interaction with SQL databases.

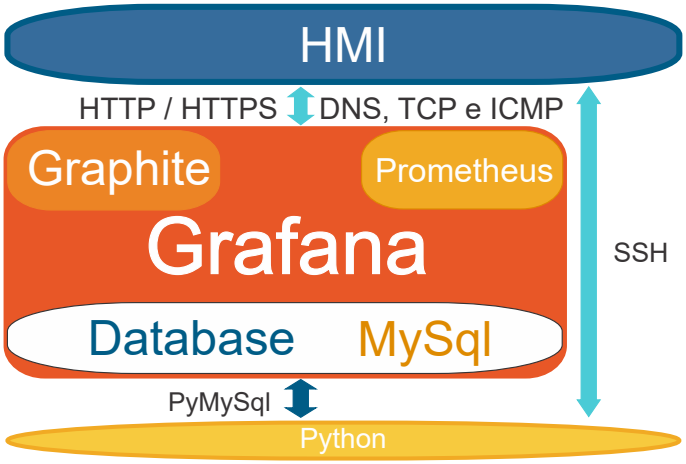


Fig. 4. Software Architecture of the Proposed System.

The data displays in a graphical user interface (GUI) and is developed on Grafana, a software tool that allows to connect with many databases, including MySQL. This software includes Prometheus and Graphite modules to request information and to display it in a modern dashboard, respectively. Grafana uses the hypertext transfer protocol (HTTP) to read the MySQL database and the dashboard runs on a web browser on multi-platform devices, from smartphone to advanced computer. It is mandatory to enable the port 3000 for access as a client, other necessary configurations are related to transmission control protocol (TCP), domain name system (DNS) and internet control message protocol (ICMP).

The software platform allows the user to easy and efficient interaction through key performance indicators in the dashboard for data visualization, using Prometheus and Graphite metrics. Through Python code and MySQL, the sensor data and GNSS/INS data are saved in the remote computer. Then, the GUI reads and displays the data, updating constantly every second. This means that the user consults the parameter values to be displayed by communicating with the MySQL database. These consulted parameters are registered from the sensors attached to the sonde and to the dual GNSS/INS devices, and are properly organized in SQL format to be transmitted in real-time. The user consults the values in the GUI using text indicators, graphical indicators, alert indicators, time series indicators, and others provided by Grafana.

All protocol and technological standards used for the interoperability of the measurement system are represented in Fig. 5. The system has compatibility with communication protocols, such as Modbus, serial UART, IEEE 802.11a/n, TCP/IP, SSH, etc. For the communication with the multiparameter sonde, a module named 'PyModbus' is used, enabling the two wire synchronous communication to set, read, write and save the collected data. Using the 'PyMySQL' module, another Python code manages the data in an SQL format and saves it in the memory of the main embedded remote computer. The multiparameter sonde and the dual GNSS/INS devices have different sampling rates and the data are saved in different tables of the database. However, the GUI updates the data independently every second and keeps the wild sampling rate for all parameters to be displayed. The open-source software files are available in [22].

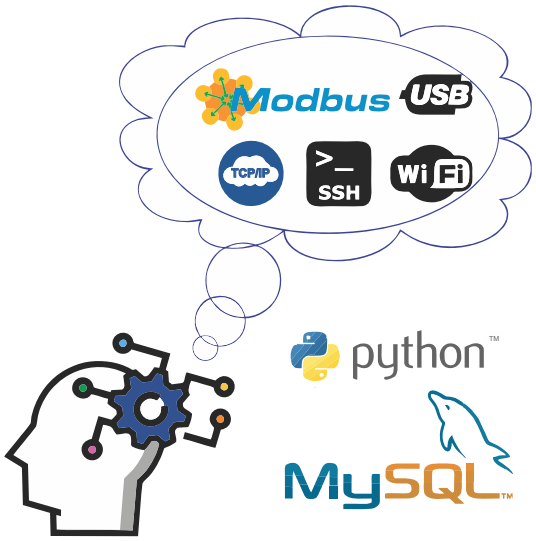


Fig. 5. Interoperability of the System with Multiple Protocols.

B. Hardware Architecture

This section presents the hardware architecture shown in Fig. 6, it is a centralized architecture building on an embedded computer connected to the multiparameter sonde and to the GNSS/INS devices through wire communication protocol. For the remote access, it uses IEEE 8011.a/n long range access point devices.

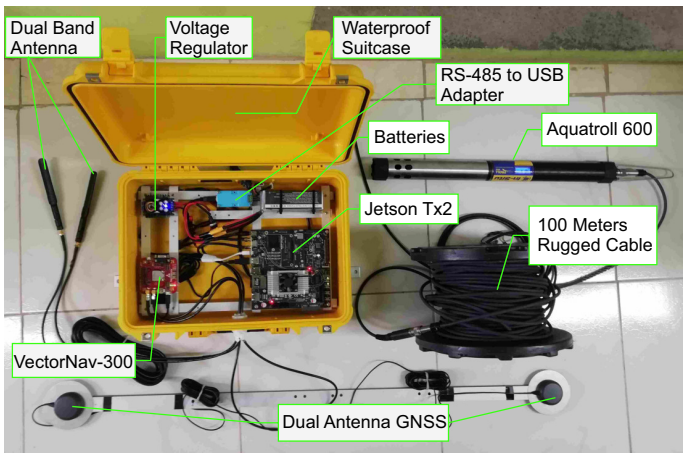


Fig. 6. Hardware Architecture of the Water Quality Measurement System.

The embedded computer is a Jetson TX2, a 64-bit ARM A57 quad-core, 1.33 TFLOPS fast embedded Artificial Intelligence, and higher energy efficiency 7.5 W - 15 W. It interconnects peripheral devices and processes all information regarding data writing, reading, storing, management, transmitting, etc. The power source is fed with two HRB, 5200

mAh, 50 C, 11.1 V, Li-Po batteries and a 19 V tension regulator required for optimal work, giving an autonomy of approximately six hours.

The sonde is a multiparameter in situ water quality instrument, series Aqua Troll 600, equipped a RS485/Modbus communication protocol, provides laboratory level sampling through optical sensors, can be submerged up to 100 meters deep through a rugged cable, has 3 seconds of sampling rate, reads 20 parameters and expandable to more (Table I). The sonde is connected to the embedded computer through the RS485 to USB adapter, the electrical connection follows the standard color code: RS485 (+) or blue to A pin, RS485 (-) or green to B pin, and GND/RETURN or black to GND pin, respectively. The embedded computer runs on Ubuntu 16.04 and where the USB adapter port is identified by ttyUSB0.

TABLE I. PARAMETERS OF THE MULTIPARAMETER SONDE ACCORDING TO THE AVAILABLE SENSORS

Parameter	Range	Resolution	Units
Temperature	268.15 °K to 323.15 °K	0.16 °K	°K
Pressure	0 m to 200 m	% full scale	PSI
Depth	0 m to 200 m	0.01 % full scale	m
Level-Depth to Water	0 m to 200 m	0.01 % full scale	m
Level-Surface Elevation	0 m to 200 m	0.01 % full scale	m
Actual Conductivity	0 to 35 S/m	10 μS/m	μS/m
Specific Conductivity	0 to 35 S/m	10 μS/m	μS/m
Resistivity	0 to 35 S/m	0.1 ohm-cm	ohm-cm
Salinity	0 to 350 PSU	0.1 PSU	PSU
Total Dissolved Solids	0 to 350 ppt	0.1 ppt	ppt
Density of Water	0 to 35 S/m	0.1 g/cm^3	g/cm^3
Barometric Pressure	300 to 1,100 mbar	0.1 mbar	mmHg
pH	0 to 14 pH units	0.01 pH	pH
pH mV	0 to 14 pH units	0.01 pH	mV
ORP	±1,400 mV	0.1 mV	mV
Dissolved Oxygen Concentration	0 to 20 mg/L 20 to 50 mg/L	0.01 mg/L	mg/L
Dissolved Oxygen % Saturation	0 to 20 mg/L 20 to 50 mg/L	0.01 mg/L	% Sat
Oxygen Partial Pressure	300 to 1,100 mbar	0.1 mbar	torr
External voltage	8 to 36 VDC	0.1 V	Volts
Battery Capacity	0 to 100 %	1 %	%

The dual GNSS/INS device is a VN-300 series, VectorNav manufacturer, combines inertial navigation system (INS), attitude heading reference system (AHRS), a global navigation system, measures angular velocities (roll, pitch and yaw) with a dynamic heading accuracy of 0.2 ° and static accuracy of 0.15 ° (see Table II for whole parameters). The embedded computer runs on Ubuntu 16 and where the USB port for the dual GNSS/INS device is identified by ttyUSB1. The remote communication wireless device is a Tp-link Pharos CPE510, provides a point-to-point wireless link with a coverage range of more than 15 km, with a speed of up to 300 Mbps (40MHz, dynamic), with the IEEE 802.11a/n wireless standard. It is fed with two HRB 5200 mAh, 50C, 11.1 V, Li-Po batteries, connected in series to obtain 22.2 V. Fig. 7 presents a detailed description of a waterproof case that contains the embedded computer and electronics, has a resistant design to protect systems against dust, water, impacts, and corrosion for chemical agents, commonly found in the harshest environment. The case dimensions are 425 mm (length), 284 mm (width) and 155 mm (height). A structural grid supports the electronics and impermeable Bulgin class connectors for the multiparameter sonde and Wi-Fi and GNSS antennas located at the external. The open-source hardware files are available in [22].

TABLE II. PARAMETERS OF THE DUAL GNSS/INS DEVICE

Parameter	Format	Accuracy	Units
Yaw	float	0.2	deg
Pitch	float	0.3	deg
Roll	float	0.3	deg
Latitude	double	9e-6	deg
Longitude	double	9e-6	deg
Altitude	double	1.5	m
Velocity X	float	0.05	m/s
Velocity Y	float	0.05	m/s
Velocity Z	float	0.05	m/s
Accel X	float	0.004	m/s^2
Accel Y	float	0.004	m/s^2
Accel Z	float	0.004	m/s^2
Angular Rate X	float	$10^\circ/hr$	rad/s
Angular Rate Y	float	$10^\circ/hr$	rad/s
Angular Rate Z	float	$10^\circ/hr$	rad/s

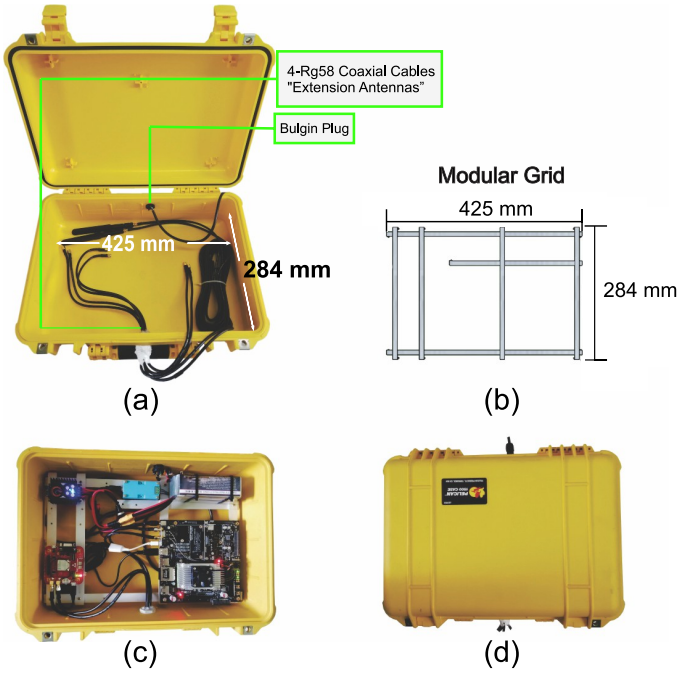


Fig. 7. Case Dimensions with Modular Grid for Supporting Embedded Computer and Electronics.

IV. RESULTS

To validate it through a real experiment, an unmanned surface vehicle (USV) (Fig. 8), whose most positive aspect is that it is scalable in terms of measurement instrumentation, with an autonomy of 5 hours and payload of 100 kg. Transporting the proposed system on a mission to measure the water quality of a huge reservoir of approximately 58000 m^3 capacity, 150 m x 150 m length, and 6 m depth, used for agriculture irrigation, and located in a desert area of Majes (longitude -72.1908, latitude -16.3586, and altitude 1.402 m.a.s.l.). The full mission took 2 hours and the data serve to ensure the water quality of the typical crops, such as chili, paprika, vegetables, potato, onion, corn, alfalfa, garlic, tomato, etc. The concern is the water is transported from the Andes through ducts and may contain unwanted concentrations that may compromise the crop and its subsequent harvest. Fig. 9 shows the measurement area of the cited agricultural water reservoir with dimensions and a segment of a circular trajectory recorded by the system.

Fig. 10 shows the global positioning of this circular trajectory, with the main parameters relative to water quality. It is observed, despite the vehicle execute a circular maneuver,the umbilical of the multiparameter sonde is subjected to external hydrodynamic forces that can vary slightly its depth location.

Fig. 8. System Carried on an Unmanned Surface Vehicle in a Water Quality Measurement Mission of a Huge Agriculture Reservoir.

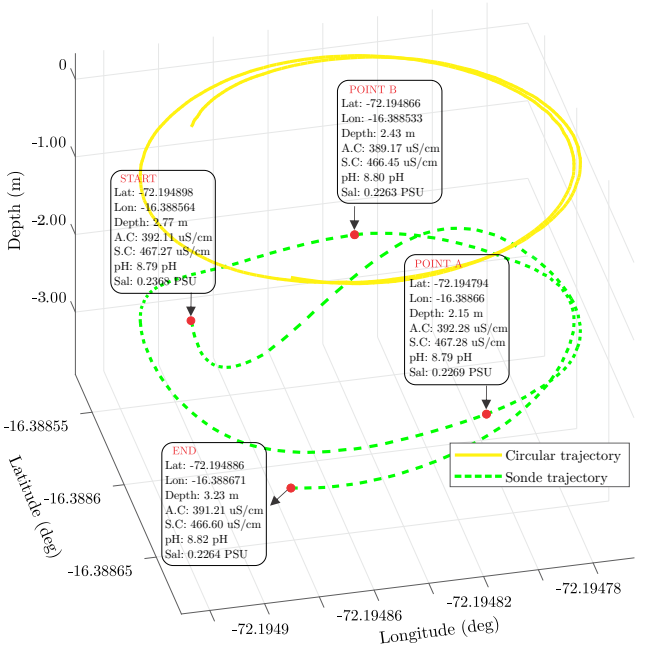


Fig. 10. Water Quality Measurement with Accurate Global Positioning in Underwater Space Environment (Lat: Latitude; Lon: Longitude; Depth; A.C: Actual Conductivity; S.C: Specific Conductivity; pH: Potential of Hydrogen; Sal: Salinity).

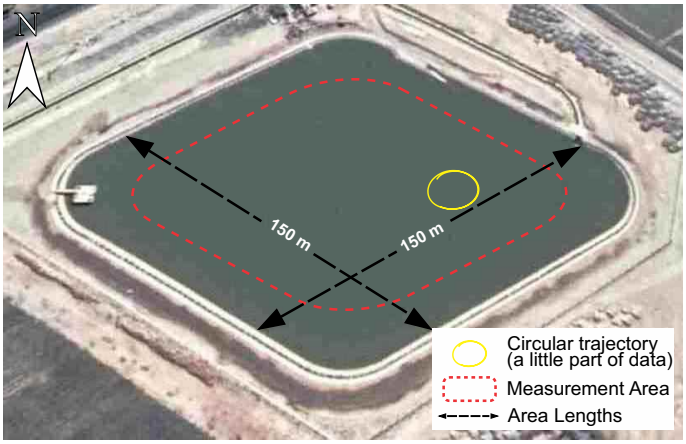


Fig. 9. Measurement Area of the Agricultural Water Reservoir with Dimensions and Circular Trajectory of the Unmanned Vehicle.

Fig. 11 and Fig. 12 present the real-time monitoring interface during the experimental test, the updated time was set to one second for both the sonde and the dual GNSS/INS, a total 1137 samples are related to the sonde parameters, and a total of 23999 samples are related to the location of these measures according to the dual GNSS/INS device. The time interval of this mission runs from 12:00 to 14:00 hs. The sonde is put up to 4 m depth, necessary to measures water quality. Fig. 11 shows a general description in the visual interface with the 20 parameters regarding water quality, including global positioning. Fig. 12 shows by separated the 15 parameters related to accurate location of the measurement using the dual GNSS/INS device. Due to the specific mission task, the vehicle is moving, and the INS enables additional data, such angular and linear velocities, that can also be visualized in the interface and in real-time.

Fig. 13 shows an offline analysis using Matlab plotting the

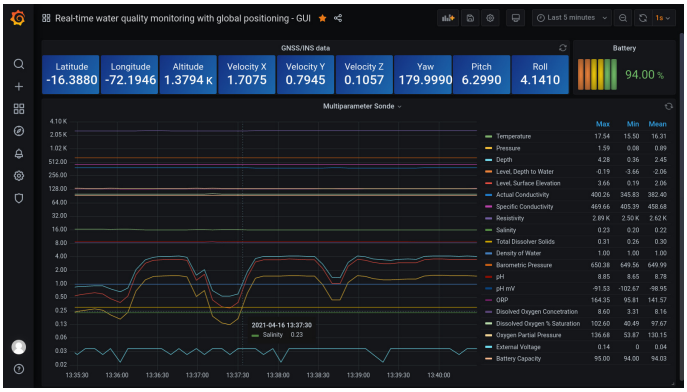


Fig. 11. Result of Experimental Test in the GUI of the Proposed System, Focused the Sonde and the GNSS/INS Parameters.

main water quality parameter. In this case, there are only four variables selected from the MySQL database, real electric conductivity, specific electric conductivity, pH level, salinity. Fig. 14 shows an offline analysis using Matlab plotting the main global positioning parameters. The software architecture enables the possibility to use any other third part software for a detailed off-line analysis. Table III shows the measured parameters in details, both when dynamic and static. The dynamic computes mean values of the whole data, and the static is the specific measurement point without vessel motion. These results have shown there are slightly variation in

measurements, the big data acquisition is necessary, can serve for better estimation and to observe specific region of the water where conditions may be critical to irrigate agricultural crops. Regarding alerts, the visual interface enables to set minimum and maximum indicators for all the parameters in order to detect inadmissible water conditions for the agriculture purposes. The user access safely protected and will be able to observe minimum, maximum, average values, etc. As an example, the alert is activated when the pH value reaches a maximum of 8.85, considering that this value may alter the life cycle of the crop and further economic losses. The proposed visual interface also may notify the alert online using JSON message protocol. Despite the updating time is one seconds, the whole sampling data are received relative to the sampling frequencies, 1 sample per three seconds by the sonde and 50 samples per second by the dual GNSS/INS device.

Fig. 12. Result of Experimental Test in the GUI of the Proposed System, Focused the GNSS/INS Parameters.

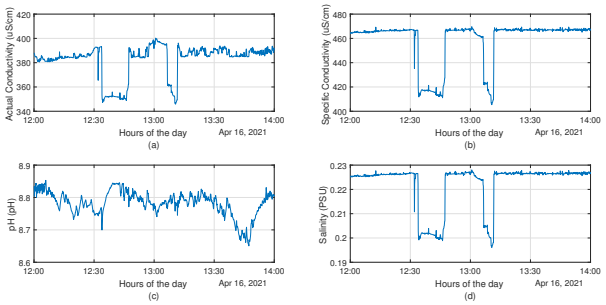


Fig. 13. Offline Analysis of Selected Data using Matlab Plot: (a) Actual Conductivity; (b) Specific Conductivity; (c) pH; and (d) Salinity.

In huge water reservoirs, it is important to observe different point of water quality, and the proposed alternative tackled this inconvenient. Another further application is for marine water quality monitoring in Peru, because these values may change dramatically relative to the depth, and it is a key factor to ensure the habitat of marine species. It is known that GNSS devices have a higher accuracy and refresh rate than GPS, but these are more susceptible to obstacles such as trees, buildings, tunnels or electromagnetic fields. Considering tasks away from these obstacles, the GNSS is an attractive solution. In [23], the GNSS devices with dual antennas presents an immunity in their measurements for a reference system of constant move-

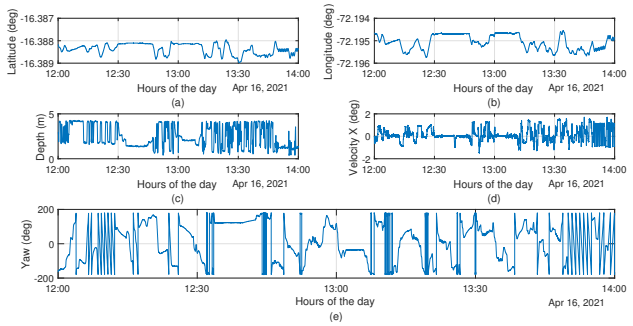


Fig. 14. Offline Analysis of selected data using Matlab plot: (a) Longitude; (b) Latitude; (c) Depth; (d) Velocity X; and (e) Yaw.

TABLE III. MEAN VALUES OF THE MEASUREMENT

Parameter	Dynamic (mean)	Static	Unit
Temperature	289.455	289.678	$^{\circ}$K
Pressure	0.893	0.416	PSI
Level-Depth to Water	-0.627	-0.292	m
Level-Surface Elevation	0.627	0.292	m
Actual Conductivity	38240.454	39214.309	μS/m
Specific Conductivity	45868.145	46784.503	μS/m
Resistivity	2618.354	2550.089	ohm-cm
Salinity	0.222	0.227	PSU
Total Dissolved Solids	0.298	0.304	ppt
Density of Water	0.999	0.999	g/cm^3
Barometric Pressure	649.987	649.777	mmHg
pH	8.782	8.807	pH
pH mV	-98.949	-100.402	mV
ORP	141.574	133.967	mV
Dissolved Oxygen Concentration	8.160	8.384	mg/L
Dissolved Oxygen Saturation	97.673	100.921	% Sat
Oxygen Partial Pressure	130.151	134.389	torr
External voltage	0.035	0.027	Volts
Battery Capacity	94.028	94.000	%
Latitude	-16.388	-16.388	deg
Longitude	-72.194	-72.194	deg
Depth	0.745	0.396	m
Velocity X	0.954	0	m/s
Velocity Y	0.001	0	m/s
Yaw	10.952	1.301	deg

ment. Moreover, it presents a greater accuracy and reliability when implementing an inertial navigation system (INS) to help in situations of signal loss. Based on this, the GNSS/INS VN-300 device used here provides a better horizontal accuracy of 1 meter and better vertical accuracy of 1.5 meters with operating frequencies up to 400 Hz in contrast with alternatives that use GPS technologies as a global positioning system [16], [17]. The proposed system is also prepared with serial ports to receive other certificated sensors and instrumentation, that fill with the high standard requirements for industry and governmental agencies. In particular, the Aquatroll 600 sonde is very common in Peru for agriculture and mining activities and the insertion of new sondes should be pass with rigorous certification procedures before to get their practical acceptance.

V. CONCLUSION

Monitoring water conditions is essential in agricultural activities and in countries where mineral exploration is so close to these activities. The integration of hardware and software is necessary to attend more requirements in terms of precision, accuracy, and human-machine interface for data visualization. This work validates the integration of a certified

multiparameter sonde with a dual GNSS/INS device in an embedded computer system in order to increase the accuracy in global position of measurement data. Moreover, the work presents open-sources technologies, a detailed description of the visual interface using MySQL and completely Python code. The system is validated when it is attached as a payload to an unmanned vehicle during the mission to measure the state of the water in a huge reservoir destined for agricultural activities in the Majes-Arequipa region, a desert area irrigated with water that flows from the Andes. Due to limited access and budget, this project collects 20 water quality parameters, based on the availability of sensors and the study area. This platform has a great potential for scalability leading to future work that aims to add more instrumentation, programming, communication and incorporation of IoT. These can contribute to research work with more features and applications of monitoring control of the aquatic environment.

ACKNOWLEDGMENT

The authors thank the 'Universidad Nacional de San Agustín de Arequipa' for the financial support given to the construction of an unmanned aquatic vehicle, under contract number IBAIB-08-2018-UNSA project. The acknowledges go also to the administrative staff of the 'Junta de Usuarios Pampa de Majes' for supporting the experimental tests in the VR-4 agriculture reservoir.

REFERENCES

[1] G. Salmoral, E. Zegarra, I. Vázquez-Rowe, F. González, L. del Castillo, G. R. Saravia, A. Graves, D. Rey, and J. W. Knox, "Water-related challenges in nexus governance for sustainable development: Insights from the city of arequipa, peru," *Science of The Total Environment*, vol. 747, p. 141114, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S004896972034643X

[2] C. S. Santana, D. M. Montalván Olivares, V. H. Silva, F. H. Luzardo, F. G. Velasco, and R. M. de Jesus, "Assessment of water resources pollution associated with mining activity in a semi-arid region," *Journal of Environmental Management*, vol. 273, p. 111148, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301479720310744

[3] J. Wright, J. Liu, R. Bain, A. Perez, J. Crocker, J. Bartram, and S. Gundry, "Water quality laboratories in colombia: A gis-based study of urban and rural accessibility," *Science of The Total Environment*, vol. 485-486, pp. 643–652, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969714004793

[4] M. S. U. Chowdury, T. B. Emran, S. Ghosh, A. Pathak, M. M. Alam, N. Absar, K. Andersson, and M. S. Hossain, "IoT based real-time river water quality monitoring system," *Procedia Computer Science*, vol. 155, pp. 161–168, 2019, the 16th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2019),The 14th International Conference on Future Networks and Communications (FNC-2019),The 9th International Conference on Sustainable Energy Information Technology. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050919309391

[5] S. Pasika and S. T. Gandla, "Smart water quality monitoring system with cost-effective using IoT," *Heliyon*, vol. 6, no. 7, p. e04096, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844020309403

[6] Y. Li, L. Tian, W. Li, J. Li, A. Wei, S. Li, and R. Tong, "Design and experiments of a water color remote sensing-oriented unmanned surface vehicle," *Sensors*, vol. 20, no. 8, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/8/2183

[7] J. Quintas, F. C. Teixeira, and A. Pascoal, "An integrated system for geophysical navigation of autonomous underwater vehicles." *IFAC-PapersOnLine*, vol. 51, no. 29, pp. 293–298, 2018, 11th IFAC Conference on Control Applications in Marine Systems, Robotics, and Vehicles CAMS 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405896318322079

[8] V. Partel, S. Charan Kakarla, and Y. Ampatzidis, "Development and evaluation of a low-cost and smart technology for precision weed management utilizing artificial intelligence," *Computers and Electronics in Agriculture*, vol. 157, pp. 339–350, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169918316612

[9] R. A. Bórquez López, L. R. Martinez Cordova, J. C. Gil Nuñez, J. R. Gonzalez Galaviz, J. C. Ibarra Gamez, and R. Casillas Hernandez, "Implementation and evaluation of open-source hardware to monitor water quality in precision aquaculture," *Sensors*, vol. 20, no. 21, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/21/6112

[10] W.-Y. Chung and J.-H. Yoo, "Remote water quality monitoring in wide area," *Sensors and Actuators B: Chemical*, vol. 217, pp. 51–57, 2015, selected Papers from the 15th International Meeting on Chemical Sensors, 16-19 March 2014, Buenos Aires, Argentina. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925400515000982

[11] J. Walker, "Remote monitoring of stock water reservoirs," *Rangelands*, vol. 43, no. 2, pp. 65–71, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0190052820301188

[12] J. S. Horsburgh, S. L. Reeder, A. S. Jones, and J. Meline, "Open source software for visualization and quality control of continuous hydrologic and water quality sensor data," *Environmental Modelling & Software*, vol. 70, pp. 32–44, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364815215001115

[13] W. L. Schulz, B. G. Nelson, D. K. Felker, T. J. Durant, and R. Torres, "Evaluation of relational and nosql database architectures to manage genomic annotations," *Journal of Biomedical Informatics*, vol. 64, pp. 288–295, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046416301526

[14] Z. li Yang, "Design on remote sensing monitoring system of navigation pharos in bridge area for inland waterway," *Procedia Computer Science*, vol. 131, pp. 409–415, 2018, recent Advancement in Information and Communication Technology:. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050918306033

[15] A. S. Ali, C. Coté, M. Heidarinejad, and B. Stephens, "Elemental: An open-source wireless hardware and software platform for building energy and indoor environmental monitoring and control," *Sensors*, vol. 19, no. 18, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/18/4017

[16] W. Jo, Y. Hoashi, L. L. Paredes Aguilar, M. Postigo-Malaga, J. M. Garcia-Bravo, and B.-C. Min, "A low-cost and small usv platform for water quality monitoring," *HardwareX*, vol. 6, p. e00076, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468067219300367

[17] K. Rajalashmi, N. Yugathian, S. Monisha, and N. Jeevitha, "Prevention of mixing of contaminated water with potable water using internet of things based water quality management system," *Materials Today: Proceedings*, vol. 45, pp. 1008–1011, 2021, international Conference on Advances in Materials Research - 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214785320318484

[18] C. Qin, S.-L. Li, S. Waldron, F.-J. Yue, Z.-J. Wang, J. Zhong, H. Ding, and C.-Q. Liu, "High-frequency monitoring reveals how hydrochemistry and dissolved carbon respond to rainstorms at a karstic critical zone, southwestern china," *Science of The Total Environment*, vol. 714, p. 136833, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969720303430

[19] G. Guillet, J. L. Knapp, S. Merel, O. A. Cirpka, P. Grathwohl, C. Zwiener, and M. Schwientek, "Fate of wastewater contaminants in rivers: Using conservative-tracer based transfer functions to assess reactive transport," *Science of The Total Environment*, vol. 656, pp. 1250–1260, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969718347351

[20] M. M. Valdivia-Fernandez, B. A. Monroy-Ochoa, D. D. Yanyachi, and J. C. Cutipa-Luque, "Parameter estimation of the alba autonomous surface craft," *International Journal of Advanced Computer Science*

and Applications, vol. 11, no. 9, 2020. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2020.0110982

[21] E. S. Rodriguez-Canales and J. C. Cutipa-Luque, "Lpv/$\mathcal{H}_\infty$ control of a twin hull-based unmanned surface vehicle," *Journal of Control, Automation and Electrical Systems*, vol. 32, no. 2, pp. 245–255, 2021. [Online]. Available: https://doi.org/10.1007/s40313-020-00669-7

[22] "An open-source wireless platform for real-time waterquality monitoring with precise global positioning," https://github.com/NielSalas/An-open-source-wireless-platform-for-real-time-water-quality-monitoring-with-precise-global-position, accessed: 2021-06-14.

[23] M. Specht, C. Specht, P. Dabrowski, K. Czaplewski, L. Smolarek, and O. Lewicka, "Road tests of the positioning accuracy of ins/gnss systems based on mems technology for navigating railway vehicles," *Energies*, vol. 13, no. 17, 2020. [Online]. Available: https://www.mdpi.com/1996-1073/13/17/4463

Structured and Unstructured Robust Control for an Induction Motor

Jhoel F. Espinoza-Quispe[1], Juan C. Cutipa-Luque[2], German A. Echaiz Espinoza[3], Andres O. Salazar[4]
Electronic Engineering Professional School
Universidad Nacional de San Agustín de Arequipa, Arequipa, Peru[1,2,3]
Computer and Automation Engineering Department
Universidade Federal de Rio Grande do Norte, Natal, Brazil[4]

***Abstract*—The indirect field-oriented control approaches for induction motors have recently gained more attention due to its use in trend areas, such as electromobility, electric vehicles, electric ships, and unmanned vehicles. This work studies the performance of two advanced control synthesized by the $\mathcal{H}_\infty$ norm as an alternative to the classical Proportional-Integral and Derivative controller. It will be assessed in terms of the performance against disturbance variations in the reference speed in the nominal conditions. The tuning of the parameters of these controllers must be defined of the stability and performance of the system and to increase their operation range frequency. An algorithm is proposed to reach a better shape of weighting functions. A numerical simulation will be shown where the advances in structured advanced controller synthesis with unstructured $\mathcal{H}_\infty$ controller is still the good election for the induction motor control. Unstructured controller approach shows still good robustness in performance and stability compared with the structured controller. Constraints imposed in structured controller is the main disadvantage to improve its robustness properties. However, compared with a conventional PID approach, the structured controller has shown quite good performance and can become in one of the most attractive approaches for practitioners.**

Keywords*—$\mathcal{H}_\infty$ *Robust control; induction motor; indirect field oriented control

I. Introduction

Nowadays, advances in research and technology seek to tackle the problem of air pollution. The study of control approaches for induction motor has gained attention due its massive use in electric vehicles and other solution involving electromobility. There are two main strategies used in induction motor control, the first is the direct torque control (DTC) and the second is the indirect field-oriented control (IFOC) [1]. The trend in innovation is the use of advanced controllers, the latest published works highlight that the IFOC is a high performance system, but it does not guarantee robustness in performance and in stability. In [1], the authors evaluate the performance between conventional Proportional-Integrative and advanced controllers with DTC and IFOC strategies. The main disadvantage in PI controller is the presence of high overshoot peak that may be reduced with modified PI structures, such as used in [2]. The advanced controllers presented quite good performance in energy efficiency relative to fuzzy logic and conventional PI controllers. The sliding mode control approach presented good speed tracking and energy efficiency, but the chattering effect may cause high frequency vibration and damage the electromechanical pieces, such as bearings and transmission gears in the powertrain [3], [4]. The unstructured $\mathcal{H}_\infty$ robust control applied to an induction motor is presented in [5] and further works considered gain scheduling and current controller [6], [7]. Recent approaches seeks to improve the performance of structured controller using the linear matrix inequalities for optimization [8], [9]. These approaches present a better tuning procedures to achieve the robustness in performance and stability and guaranteed the user requirements. In [10], the authors present emerging concept related to control of mechatronic systems, putting the loop shaping design as a good tool for synthesis of industrial controllers and remarking the $\mathcal{H}_\infty$ as a good alternative in advanced control due to the robustness. In [11], the authors present a $\mathcal{H}_\infty$ controller synthesis including a model reference adaptive estimator to avoid the use of encoders. Other authors propose a robust control approach applied to ship propulsion electric motor [12]. Another advanced control approaches [13], [14], [15], [16] are also applied to induction motors, but this work is regarding the comparison of the two types which have in common the $\mathcal{H}_\infty$ norm for the controller synthesis.

The organization of this paper follows: Section I presents the introduction over the induction motor control approaches, Section II presents the mathematical model of the induction motor, Section III presents the robust control approaches (structured and unstructured), Section IV presents the numerical results and Section V provides the conclusions.

II. Motor Model

Using the park transform, there are three general models to represent an induction motor. The first is based on an arbitrary rotating reference frame, the second is based on a synchronous rotating reference frame, and the third is based on a stationary reference frame. In this work, following the nomenclature in [17], the $qd0$ stationary reference frame is used to model the induction motor where the equations relative to the stator are:

$$v_{qs}^s = \frac{N}{\omega_b}\psi_{qs}^s + r_s i_{qs}^s \tag{1}$$

$$v_{ds}^s = \frac{N}{\omega_b}\psi_{ds}^s + r_s i_{ds}^s \tag{2}$$

$$v_{0s} = \frac{N}{\omega_b}\psi_{0s} + r_s i_{0s} \tag{3}$$

where v_{qs}^s, v_{ds}^s, and v_{0s} are the stator voltages; ψ_{qs}^s, ψ_{ds}^s, and ψ_{0s} are their magnetic flux; i_{qs}^s, i_{ds}^s, and i_{0s} are their electric currents; N is the number of poles; ω_b is the base electrical

frequency; and r_s is the stator resistance.
Similarly, the equations relative to the rotor are:

$$v'^s_{qr} = \frac{N}{\omega_b}\psi'^s_{qr} - \frac{\omega_r}{\omega_b}\psi'^s_{dr} + r'_r i'^s_{qr} \tag{4}$$

$$v'^s_{dr} = \frac{N}{\omega_b}\psi'^s_{dr} + \frac{\omega_r}{\omega_b}\psi'^s_{qr} + r'_r i'^s_{dr} \tag{5}$$

$$v'_{0r} = \frac{N}{\omega_b}\psi'_{0r} + r'_r i'_{0r} \tag{6}$$

where v'^s_{qr}, v'^s_{dr}, and v'_{0r} are the rotor voltages; ψ'^s_{qr}, ψ'^s_{dr}, and ψ'_{0s} are their magnetic flux; i'^s_{qr}, i'^s_{dr}, and i'_{0s} are their electric currents; N is the number of poles; ω_b is the base electrical frequency; r'_r is the rotor resistance; and ω_r is the rotor angular frequency.
Considering $\Psi = [\psi^s_{qs}, \psi^s_{ds}, \psi_{0s}, \psi'^s_{qr}, \psi'^s_{dr}, \psi'_{0r}]^T$ as the magnetic flux vector and $I = [i^s_{qs}, i^s_{ds}, i_{0s}, i'^s_{qr}, i'^s_{dr}, i'_{0r}]^T$ as the current vector, the following relation can be expressed:

$$\Psi = X_{sr}.I, \tag{7}$$

where X_{sr} is the stator-rotor reactance matrix equivalent to:

$$X_{sr} = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \tag{8}$$

with:

$$X_{11} = \begin{bmatrix} x_{ls}+x_m & 0 & 0 \\ 0 & x_{ls}+x_m & 0 \\ 0 & 0 & x_{ls} \end{bmatrix}, \tag{9}$$

$$X_{22} = \begin{bmatrix} x'_{lr}+x_m & 0 & 0 \\ 0 & x'_{lr}+x_m & 0 \\ 0 & 0 & x'_{lr} \end{bmatrix}, \tag{10}$$

$$X_{12} = X_{21} = \begin{bmatrix} x_m & 0 & 0 \\ 0 & x_m & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{11}$$

where x_{ls} is the stator leakage reactance, x_m is the stator magnetizing reactance, and x'_{lr} is the rotor leakage reactance. All parameter value can be found in Table I.
The motor electromagnetic torque T_{em} is defined as:

$$T_{em} = \frac{3}{2}\frac{N}{2\omega_b}(\psi^s_{ds}.i^s_{qs} - \psi^s_{qs}.i^s_{ds}), \tag{12}$$

and the 'per unit of speed relation' with the externally-applied mechanical torque T_{mech} (in the same direction of rotor speed) and the damping torque T_{damp} (in the opposite direction of motor speed) is expressed as follows:

$$T_{em} + T_{mech} - T_{damp} = 2H\frac{d}{dt}\left(\frac{\omega_r}{\omega_b}\right), \tag{13}$$

where $H = J\omega^2_{bm}/2S_b$ is the inertia constant, J is the rotor inertia, ω_{bm} is the base mechanical frequency, and S_b is rated volt-ampere. Table I shows the main parameter values of the 20 HP induction motor used in this work for numerical validation.

The whole above relations and differential equations should be linearized for their use in synthesis of proposed control approaches. In this work, the nonlinear model is represented in Simulink to be used numerical approximation for linearization

TABLE I. PARAMETER VALUES OF THE INDUCTION MOTOR

Parameters	Symbols	Values	Units
Stator winding resistance	r_s	0.1062	Ω
Stator leakage reactance	x_{ls}	0.2145	Ω
Stator magnetizing reactance	x_m	5.8339	Ω
Rotor leakage reactance	x'_{lr}	0.2145	Ω
Rotor winding resistance	r'_r	0.0764	Ω
Number of poles	N	4	$-$
Moment of inertia	J	2.8	$Kg.m^2$
Rated voltage	V_{rated}	220	v
Rated frequency	f_{rated}	60	Hz
Rated speed	N_{rated}	1748.3	rpm

through the Matlab 'linmod' command in order to obtain the following linear time invariant representation:

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx + Du, \end{aligned} \tag{14}$$

where x is the state vector, u is the control vector, y is the measurement, A is the state matrix, B is the control matrix, C is the output matrix and D is the feedthrough matrix. The relation between output and input in the s Laplace domain results in the transfer function of the system:

$$G(s) = \frac{y(s)}{u(s)} = C(sI - A)^{-1}B + D \tag{15}$$

Table II presents the poles and zeros of three reduced-order models of the system $G(s)$, commonly found in the literature. The reduced-order models have been be obtained using Hankel singular values that indicate the respective state energy of the system. The 5th reduced-order model is used for controller syntheses and the 2nd reduced-order model is used to find the natural frequency $\omega_0 = 0.21$ rad/s to shape properly the sensitivity functions to improve the performance and robustness of the $\mathcal{H}_\infty$ controllers.

TABLE II. POLES AND ZEROS OF THE REDUCED-ORDER MODELS FOR $G(s)$

	9th order	5th order	2nd order
Poles	Value	Value	Value
p1	0	0	0.2121
p2	-2850766.0028	$-4.7652 + 3.7675j$	-0.2043
p3	-1451235.2842	$-4.7652 - 3.7675j$	
p4	-1451235.2313	0.2121	
p5	-113.0944	-0.2052	
p6	$-4.7652 + 3.7675j$		
p7	$-4.7652 - 3.7675j$		
p8	0.2121		
p9	-0.2052		
Zeros	Value	Value	Value
z1	0	0	-2.6353
z2	-2850766.0028	-4.7619	
z3	-1451235.2579	$-3.5928 + 2.0923j$	
z4	-113.0944	$-3.5928 - 2.0923j$	
z5	-4.7619		
z6	$-3.5928 + 2.0923j$		
z7	$-3.5928 - 2.0923j$		

III. CONTROL APPROACH

This section presents two control approaches based in loop shaping design, or well known as the $\mathcal{H}_\infty$ mixed sensitivity approach. The first approach leads to get an $\mathcal{H}_\infty$ controller

with no defined structure, named unstructured controller. The second approach leads to get an $\mathcal{H}_\infty$ controller with a conventional PID structure, named structured controller. Both approaches use sensitivity weighting functions for loop shaping in order to get a robustness in performance and stability [18], [19].

A. Unstructured Controller

Fig. 1 shows G, the 5th order induction motor system represented by (15), inserted into the augmented plant P with two port representation commonly used in the $\mathcal{H}_\infty$ controller synthesis. The diagram also contains the weighting functions W_*, the exogenous output z, the exogenous input w, the control variable u and the controlled variable y. According to this representation, w is equivalent to the reference rotor speed reference signal $\omega_r(ref)$, u is equivalent to the electromagnetic torque signal T^*_{em}, y is the measured rotor speed signal ω_r. And z contains the error between reference and measured rotor speed, the electromagnetic torque control, and the measured rotor speed, weighting respectively by functions W_P, W_u, and W_T.

The problem of unstructured control is defined here as to find a rotor speed controller K that minimizes the $\mathcal{H}_\infty$ norm of the transfer function between exogenous output z and exogenous input w as follows [18]:

$$\|T_{zw}\| = \|W_P S \;\; W_u K S \;\; W_T T\|^T_\infty \tag{16}$$

where the sensitivity function S, the complementary sensitivity function T and the control sensitivity function KS are defined respectively as:

$$S = (I - GK)^{-1} \tag{17}$$

$$T = (I - GK)^{-1} GK \tag{18}$$

$$KS = (I - GK)^{-1} K \tag{19}$$

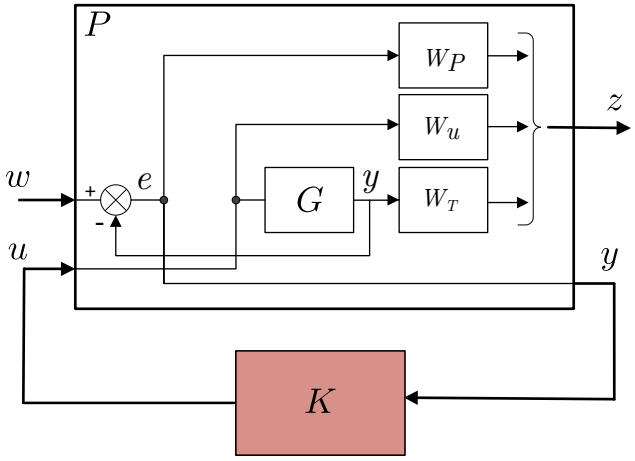


Fig. 1. Unstructured $\mathcal{H}_\infty$ Controller.

B. Structured Controller

Fig. 2 shows G, the 5th order induction motor system represented by (15), inserted also into the augmented plant P with two port representation. The diagram contains the weighting functions W_*, the exogenous output z, the exogenous input w, the control variable u and the controlled variable y. And contains all variables at same representation in the last section with the only difference that the controller K has now a defined PID structure:

$$K = k_p + k_i \frac{1}{s} + k_d s, \tag{20}$$

The problem of structured control is defined here as to find a rotor speed controller K that minimizes the $\mathcal{H}_\infty$ norm of the transfer function between exogenous output z and exogenous input w as follows [19]:

$$\|T_{zw}\| = \|W_P S \;\; W_T T\|^T_\infty \tag{21}$$

subject to structure constraints given in (20). It is important to note the electromagnetic torque control signal is not included in the above matrix to avoid degradation in performance due to limited constraints given to K structure.

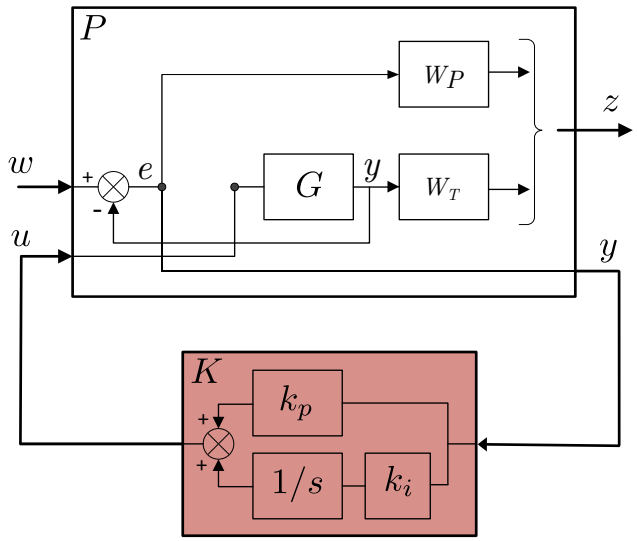


Fig. 2. Structured $\mathcal{H}_\infty$ Controller.

C. Computation Resources

The problem of minimization of $\mathcal{H}_\infty$ norm of equations (16) and (21) will be solved using computational resources available in commercial and non-commercial software, such as Matlab, Gnu-Octave, and Scilab. This solution is based in the algebraic Riccati equation, or alternatively putting the problem in the linear and bi-linear matrix inequalities for respective optimization [20]. Assuming that the involved matrices in 15 satisfy detectability and stabilizability conditions, there is a suboptimal controller K that the closed function T_{zw} achieves:

$$\|T_{zw}\| = \gamma, \tag{22}$$

where γ is a real number related with the suboptimal control problem. The design requirements for control of an induction motor is given in terms of closed loop sensitivities:

1) Closed loop stability,
2) $\overline{\sigma}(S) < 1$ for $\omega < \omega_P$,
3) $\overline{\sigma}(T) < 1$ for $\omega > \omega_T$,
4) $\overline{\sigma}(KS) < 1$ for $\omega > \omega_u$,

where the crossover frequencies of the sensitivity functions are ω_P, ω_T and ω_u, respectively. These frequencies are commonly set according to good tracking, good disturbance and noise rejection, and model uncertainties. In this work, the goal is to improve still more these requirements and these crossover specification frequencies are also tuned using a frequency sweep algorithm as shown below.

D. Frequency Sweep

The design requirements given through sensitivity functions S, T and KS are weighting with W_P, W_T and W_u according to:

$$W_P(s) = \frac{s/M_P + \omega_P}{s + \omega_P A_P} \tag{23}$$

$$W_T(s) = \frac{s + \omega_T A_T}{s/M_T + \omega_T} \tag{24}$$

$$W_u(s) = \frac{s + \omega_u A_u}{s/M_u + \omega_u} \tag{25}$$

where M_P, M_T, M_u, A_P, A_T and A_u are constants. In this work, both the unstructured and structured control synthesis use a sweeping in frequency in its cutoff frequencies in order to improve the operation bandwidth.

Algorithm 1 Frequency sweep

```
Define the system G(s), the weighting function constants M_P, M_T, M_u, A_P, A_T and A_u,
Set the settling time t_s; the initial cutoff frequencies ω_P, ω_T and ω_u; and the frequency steps Δ_P, Δ_T and Δ_u,
while ω_P < ω_0 < ω_T and ω_u > ω_0 do:
    Sweep ω_P = ω_P + Δ_P,
          ω_T = ω_T − Δ_T,
          ω_u = ω_u − Δ_u,
    Select the H_∞ unstructured or structured controller,
    Configure the augmented plant P,
    Compute the controller K and γ,
    Plot the sensitivity functions S, T and KS,
    Plot time responses of the feedback system and
    compute the ‖i_as‖_∞ and ‖T*_em‖_∞,
    Save the data,
end while
Generate the report with the available results: K, γ, ‖i_as‖_∞ and ‖T*_em‖_∞.
```

The sweeping in frequency starts with initial cutoff frequencies given around the natural frequency ω_0, the final report identifies the absolute maximum values of the stator current i_{as} and the electromagnetic torque T^*_{em}. This algorithm enables to observe the water bed effect between performance and stability in robust control approaches, aids to select a better controller K between a set, ensures the control variable do not exceed physical limits, and guarantees good performance according to the settling time t_s and the zero steady state error.

IV. Results

This section presents the results of the two control approaches used in this work. Fig. 3 presents the simulator of an induction motor on Simulink, where the $qd0$ transform is explicit, the block of robust speed controller contains the controller (unstructured or structured), the speed rotor reference signal is denoted by $\omega_r(ref)$, and other related components of the dynamic model [17].

A. Unstructured Robust Control

The initial cutoff frequencies were set to $\omega_P = 0.01$ rad/s, $\omega_T = 1$ rad/s, $\omega_u = 2000$ rad/s. After a number of 100 iterations running into the sweep algorithm, the report shows a better achieved gamma value of $\gamma = 0.9549$ with cutoff frequencies of $\omega_P = 0.01$ rad/s, $\omega_T = 0.9$ rad/s, $\omega_u = 1000$ rad/s. The obtained suboptimal unstructured controller K is of 7th order. Fig. 4 presents the sensitivities (S, T, KS) and their respective weightings (W_P, W_T, W_u) in the frequency domain response, there is no crossing between each sensitivity and its respective weighting that confirms the good robustness in performance and stability.

B. Structured Robust Control

The initial cutoff frequencies were set to $\omega_P = 0.01$ rad/s, $\omega_T = 1$ rad/s. After a number of 100 iterations running into the sweep algorithm, the report shows an achieved gamma value of $\gamma = 23.7674$ with cutoff frequencies of $\omega_P = 0.01$ rad/s, $\omega_T = 0.1$ rad/s. The obtained suboptimal unstructured controller K has a PID structure (20) with $k_p = 163.3002$, $k_i = 0.0016$ and $k_d = 0$. Fig. 5 presents the sensitivities (S, T) and their respective weightings (W_P, W_T) in the frequency domain response. Despite the good loop shaping in S and T, there are crossings between each sensitivity and its respective weighting that confirms the poor robustness in performance and stability. This means that the unstructured controller has better properties of disturbance rejections, and the constraints given in this structured controller have limited the suboptimal solution to high gamma values.

C. Time Domain Responses

This section presents numerical results using a simulator when the system is subjected to the reference rotor speed $\omega_r(ref)$ and the required mechanical torque T_{mech}. The $\omega_r(ref)$ starts in zero and increases in ramp mode to reach a nominal value of the motor 188.5 rad/s. The required mechanical torque is a pulsating signal given in function of the machine nominal torque 81.49 Nm as shown in Fig. 6.

Fig. 7 presents the time domain responses of v_{ag}, one of the three-phase voltages applied to the stator. Considering the structured controller (top side), the $\|v_{ag}\|_\infty$ appears in the first mechanical torque variation, reaching 552.0 V. Considering the unstructured controller (bottom side), the $\|v_{ag}\|_\infty$ appears in the first mechanical torque variation, reaching 286.1 V. It is observed high step variation in mechanical torque produces a peak in the voltage to compensate this demand.

Fig. 8 presents the time domain responses of i_{as}, one of the three-phase current applied to the stator. Considering the structured controller (top side), the $\|i_{as}\|_\infty$ appears in the

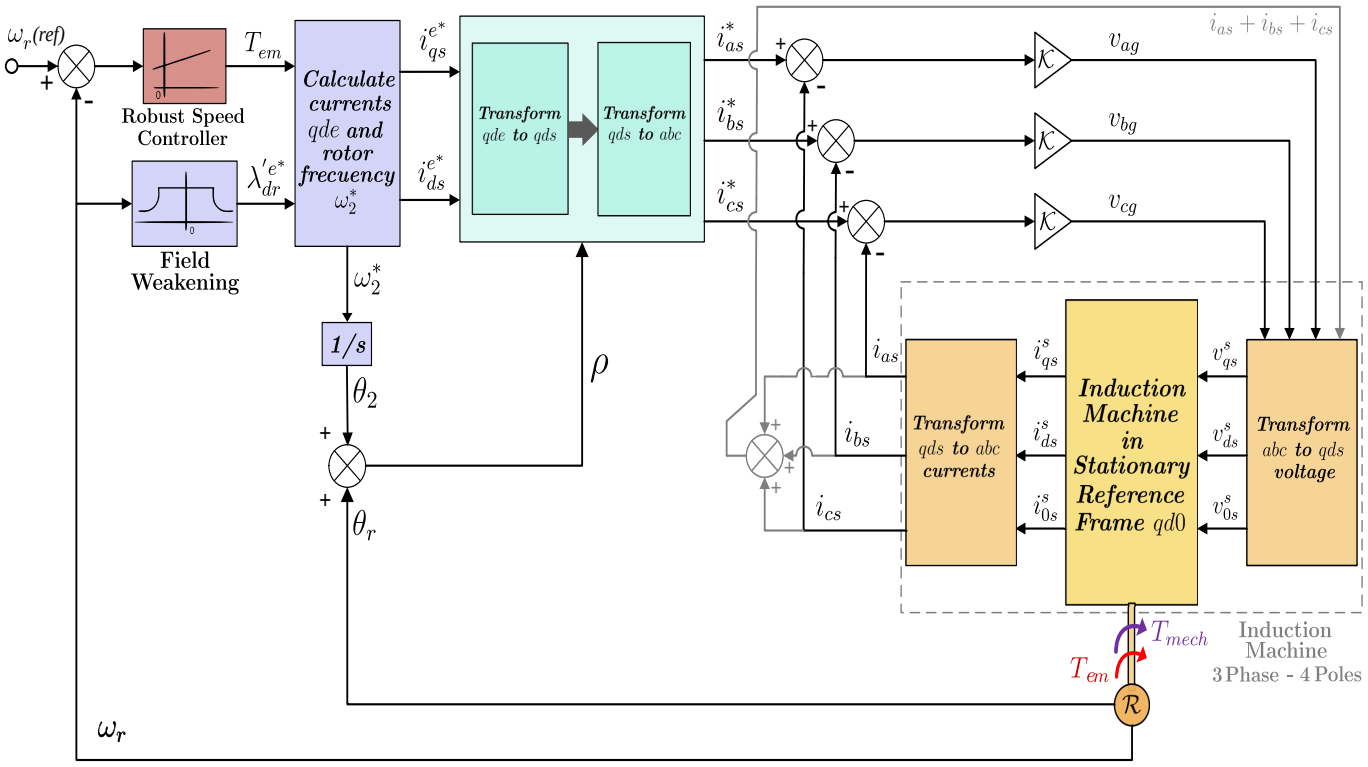


Fig. 3. Simulator of Induction Motor in Simulink.

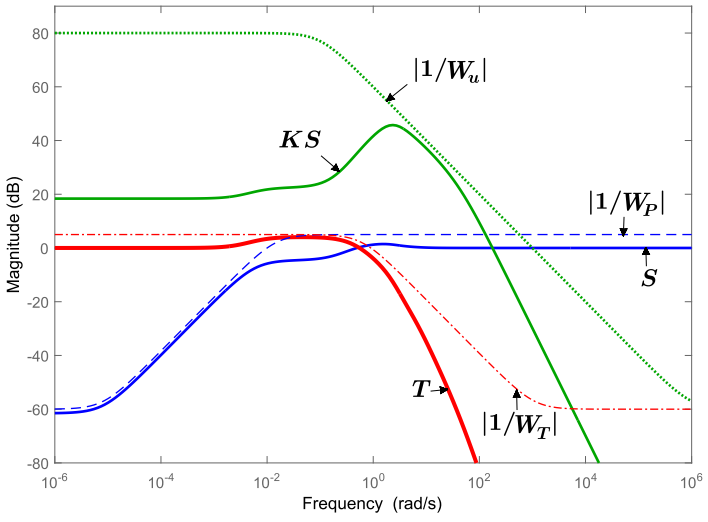


Fig. 4. Sensitivities Unstructured $\mathcal{H}_\infty$ Controller.

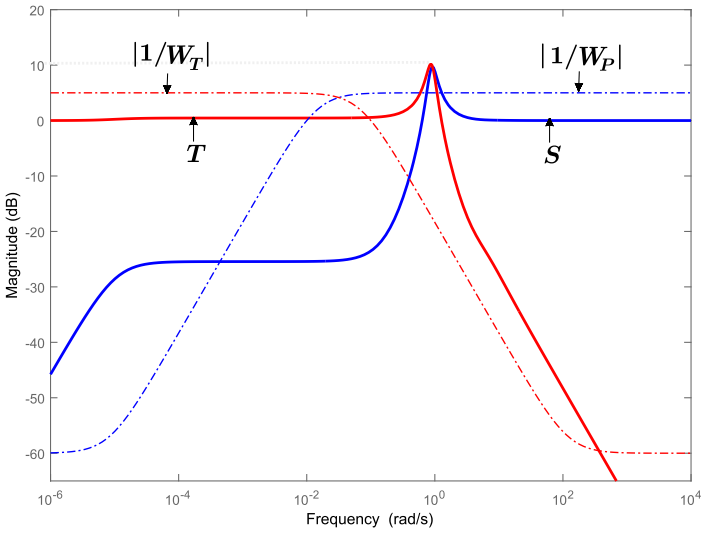


Fig. 5. Sensitivities Structured $\mathcal{H}_\infty$ Controller.

starting motor, reaching 300.4 A. Considering the unstructured controller (medium side), the $\|i_{as}\|_\infty$ appears in the starting motor, reaching 211.0 A. Considering the conventional PI controller, tuned as in [17], (bottom side), the $\|i_{as}\|_\infty$ appears in the starting motor, reaching 147.4 A. It is observed high mechanical torque produces increments in the currents to compensate this demand.

Fig. 9 presents a comparison between controlled variables, speed rotor ω_r, using unstructured, structured and PI controllers. At right side, zoom shows the structured controller has the best tracking despite the variation in mechanical torque, following by the unstructured controller. The poor performance in these results is getting by the conventional PI controller, included the significant error in steady state.

Fig. 10 presents a comparison of control variable responses, electromagnetic torque T_{em}, using unstructured, structured and PI conventional controllers. The structured controller has the maximum $\|T_{em}\|_\infty = 288$ Nm between the three controllers. The lowest $\|T_{em}\|_\infty = 141$ Nm is related with the con-

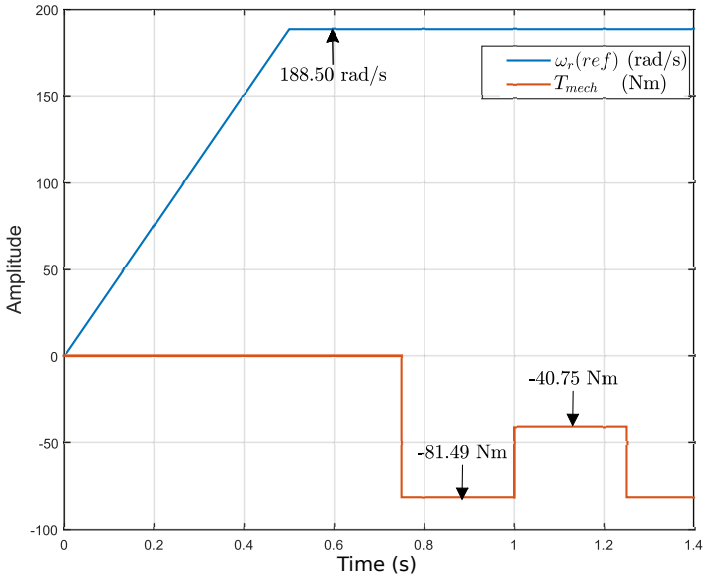


Fig. 6. Reference Speed $\omega_r(ref)$ and Mechanical Torque T_{mech}.

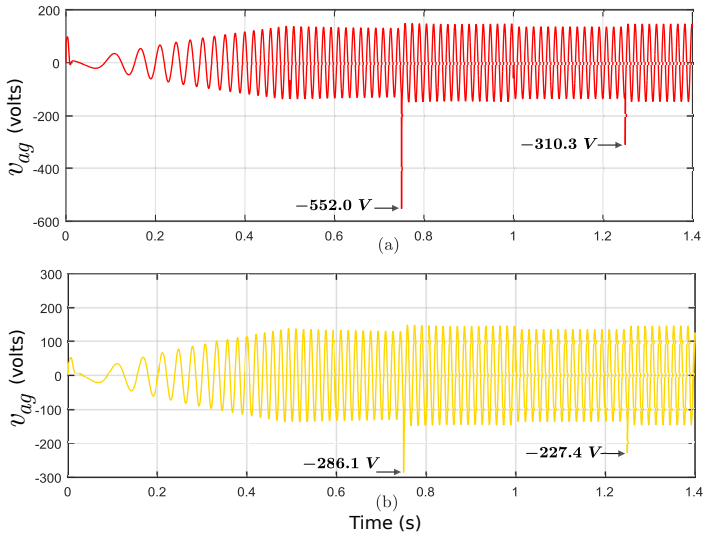


Fig. 7. The Stator Voltage v_{ag} Responses with $\mathcal{H}_\infty$ Controllers: (a) Structured Controller, (b) Unstructured Controller.

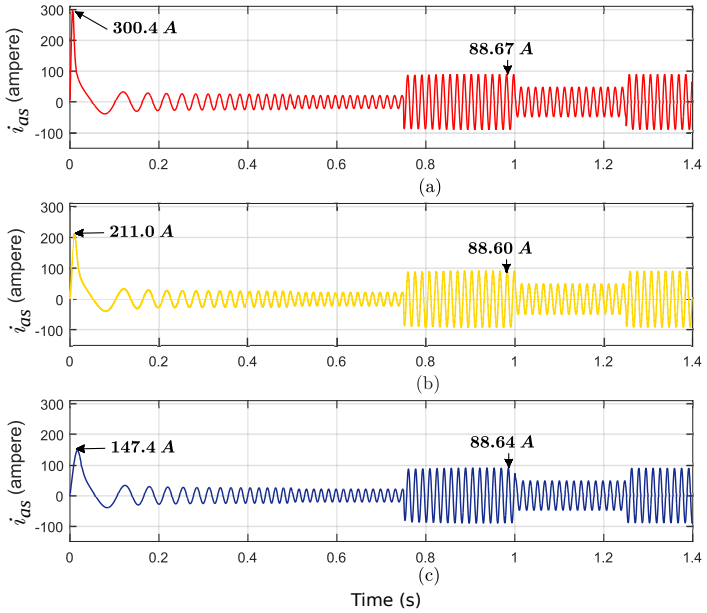


Fig. 8. The Stator Current i_{as} Responses with $\mathcal{H}_\infty$ Controllers: (a) Structured Controller, (b) Unstructured Controller, (c) PI Conventional Controller.

ventional PI controller. The moderate results are getting by the unstructured controller. However, there are no significant difference in the steady state.

Despite the use of $\mathcal{H}_\infty$ approach to synthesis structured controller [18], [7], there is no information regarding the synthesis of structured controller using the same norm. Relative to the two last cited works, the proposed frequency sweeping algorithm enables to seek a better performance ensuring that the control efforts remain between the physical limits, the reader may observe the current and voltages responses.

V. Conclusion

This paper deals with the study of the two robust control approaches based in $\mathcal{H}_\infty$ norm and applied to improve the performance and stability of an induction motor. The first approach, the unstructured $\mathcal{H}_\infty$ control presents quite good responses, included robustness properties. The second approach, the structured $\mathcal{H}_\infty$ control approach presents better responses relative to the conventional PI controller. However, its constraints in its defined structure limits a suboptimal solution regarding robustness and may compromise the disturbance rejection. The proposed sweep frequency algorithm, used in the two approaches, seeks a gamma suboptimal value close to 1, iterating the solution relative to the natural frequency. The structured controller is more attractive for practitioners due to its straightforward implementation; but its robustness properties should be studied in a further work considering more parameters and using a testing bench. Results indicate the good election for robustness design is still the unstructured controller.

Acknowledgment

The authors thank the *Universidad Nacional de San Agustín de Arequipa* for supporting this research under incentives for research publication.

References

[1] M. Aktas, K. Awaili, M. Ehsani, and A. Arisoy, "Direct torque control versus indirect field-oriented control of induction motors for electric vehicle applications," *Engineering Science and Technology, an International Journal*, vol. 23, no. 5, pp. 1134–1143, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2215098619326734

[2] H. Guessoum, C.-E. Feraga, L. Mehennaoui, M. Sedraoui, and A. Lachouri, "A robust performance enhancement of primary h∞ controller based on auto-selection of adjustable fractional weights: Application on a permanent magnet synchronous motor," *Transactions of the Institute of Measurement and Control*, vol. 41, no. 11, pp. 3248–3263, 2019. [Online]. Available: https://doi.org/10.1177/0142331218823861

[3] A. Devanshu, M. Singh, and N. Kumar, "Sliding mode control of induction motor drive based on feedback linearization," *IETE Journal of Research*, vol. 66, no. 2, pp. 256–269, 2020. [Online]. Available: https://doi.org/10.1080/03772063.2018.1486743

[4] D. C. Happyanto, A. Aditya, and B. Sumantri, "Boundary–layer effect in robust sliding mode control for indirect field oriented control of 3-phase induction motor," *International Journal on Electrical Engineering and Informatics*, vol. 12, pp. 188–204, 2020.

[5] C. Paterson, I. Postlethwaite, D. Walker, and A. Bashagha, "The development and evaluation of an h-infinity induction motor controller," *IFAC Proceedings Volumes*, vol. 29, no. 1, pp. 3368–3373, 1996, 13th World Congress of IFAC, 1996, San Francisco USA, 30 June - 5 July. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474667017581975

[6] E. Prempain and I. Postlethwaite, "H∞, design for an induction motor," *IFAC Proceedings Volumes*, vol. 35, no. 1, pp. 211–216, 2002, 15th IFAC World Congress. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474667015387905

[7] C.-E. Feraga, M. Sedraoui, and R. Bachir Bouiadjra, "Enhanced indirect field-oriented control of single-phase induction motor drive using h∞ current controller," *Arabian Journal for Science and Engineering*, vol. 44, no. 8, pp. 7187–7202, 2019. [Online]. Available: https://doi.org/10.1007/s13369-019-03850-6

[8] J. K. Jain, S. Ghosh, S. Maity, and P. Dworak, "Pi controller design for indirect vector controlled induction motor: A decoupling approach," *ISA Transactions*, vol. 70, pp. 378–388, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S001905781730438X

[9] J. Jain, S. Ghosh, and S. Maity, "Concurrent pi controller design for indirect vector controlled induction motor: Concurrent pi controller design," *Asian Journal of Control*, vol. 22, 09 2018.

[10] C. M. Ionescu, E. H. Dulf, M. Ghita, and C. I. Muresan, "Robust controller design: Recent emerging concepts for control of mechatronic systems," *Journal of the Franklin Institute*, vol. 357, no. 12, pp. 7818–7844, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0016003220303999

[11] A. A. Z. Diab, A.-H. M. El-Sayed, H. H. Abbas, and M. A. E. Sattar, "Robust speed controller design using h infinity theory for high-performance sensorless induction motor drives," *Energies*, vol. 12, no. 5, 2019. [Online]. Available: https://www.mdpi.com/1996-1073/12/5/961

[12] G. Rigatos, P. Siano, and M. Abbaszadeh, "Nonlinear optimal control for ship propulsion systems comprising an induction motor and a drivetrain," *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment*, vol. 234, no. 2, pp. 409–425, 2020. [Online]. Available: https://doi.org/10.1177/1475090219885213

[13] E. Zhao, J. Yu, J. Liu, and Y. Ma, "Neuroadaptive dynamic surface control for induction motors stochastic system based on reduced-order observer," *ISA Transactions*, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0019057821004900

[14] M. A. Mossa and H. Echeikh, "A novel fault tolerant control approach based on backstepping controller for a five phase induction motor drive: Experimental investigation," *ISA Transactions*, vol. 112, pp. 373–385, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0019057820305188

[15] P. Alkorta, J. A. Cortajarena, O. Barambones, and F. J. Maseda, "Effective generalized predictive control of induction motor," *ISA Transactions*, vol. 103, pp. 295–305, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0019057820301610

[16] Z. Yang, Q. Ding, X. Sun, H. Zhu, and C. Lu, "Fractional-order sliding mode control for a bearingless induction motor based on improved load torque observer," *Journal of the Franklin Institute*, vol. 358, no. 7, pp. 3701–3725, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0016003221001733

[17] C.-M. Ong *et al.*, *Dynamic simulation of electric machinery: using MATLAB/SIMULINK*. Prentice hall PTR Upper Saddle River, NJ, 1998, vol. 5.

[18] S. Skogestad and I. Postlethwaite, *Multivariable feedback control: analysis and design*. Citeseer, 2007, vol. 2.

[19] P. Apkarian and D. Noll, "Nonsmooth h∞ synthesis," *IEEE Transactions on Automatic Control*, vol. 51, no. 1, pp. 71–86, 2006.

[20] P. Gahinet and P. Apkarian, "Decentralized and fixed-structure h∞ control in matlab," in *2011 50th IEEE Conference on Decision and Control and European Control Conference*, 2011, pp. 8205–8210.

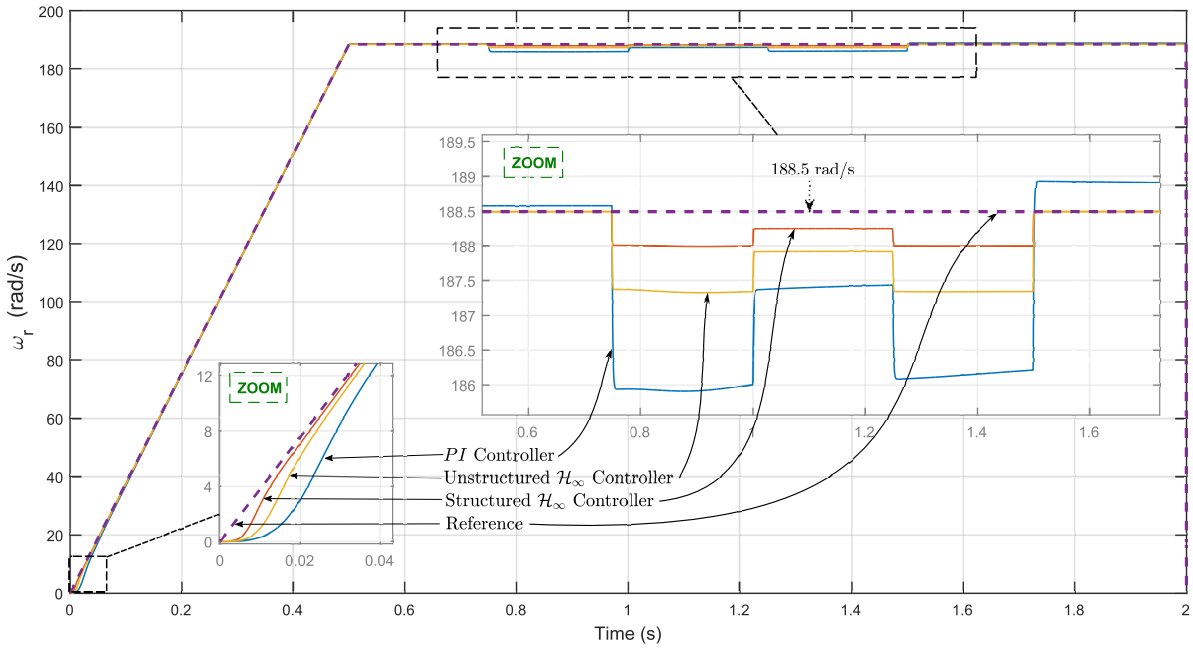


Fig. 9. Controlled Variable Comparison, Rotor Speed ω_r.

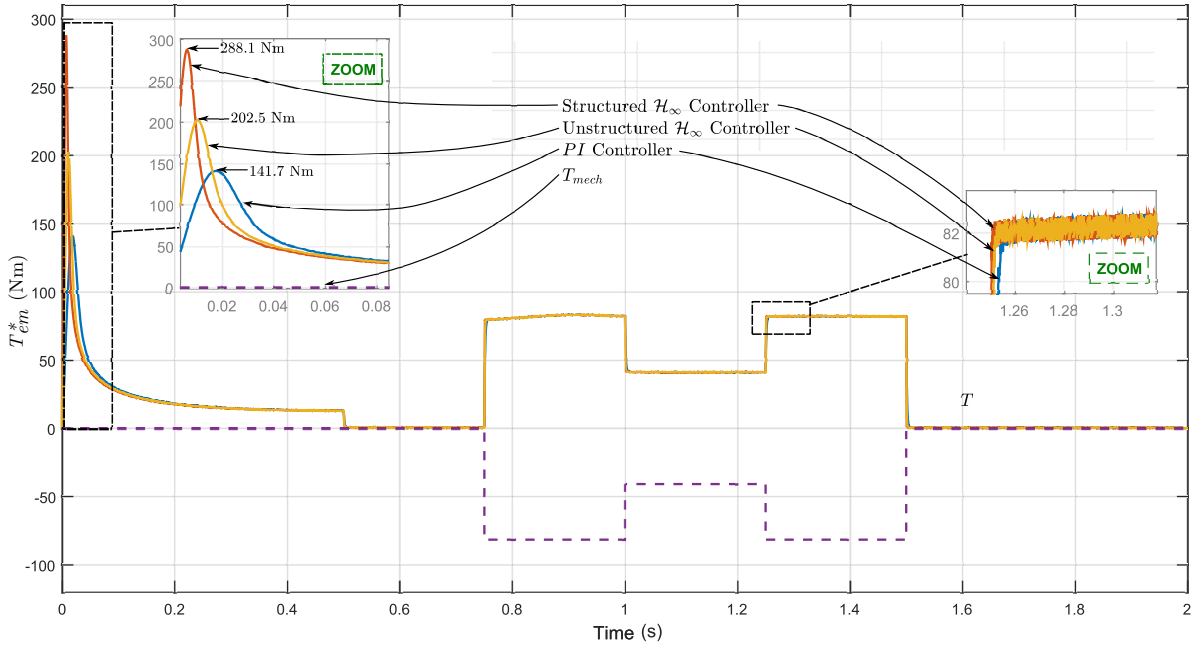


Fig. 10. Control Variable Comparison, Electromagnetic Torque T^*_{em} Compensates the Required Mechanical Torque T_{mech}.

Employing DDR to Design and Develop a Flipped Classroom and Project based Learning Module to Applying Design Thinking in Design and Technology

Mohd Ridzuan Padzil, Aidah Abd Karim, Hazrati Husnin
Faculty of Education, Universiti Kebangsaan Malaysia
43600 Bangi, Selangor, Malaysia

***Abstract*—The purpose of this study is to discuss the Design and Development (DDR) research approach that was used to develop a Flipped Classroom and project-based learning modules for students of Design and Technology (D&T). The module's fundamental theory is based on 21st-century teaching and learning models, as well as design thinking. The DDR process is divided into three phases: analysis of needs, design and development, and evaluation. The phase of needs analysis is used to ascertain the necessity of module development and the application of design thinking. Three distinct data collection methods were used in this phase: semi-structured interviews, survey studies, and document analysis. The findings from this phase serve as a backup for the next phase. The Isman Instructional Design Model (2011) is adapted for use in this phase as a guide for module design and development. Additionally, the Fuzzy Delphi Method is used to obtain expert consensus on module material design, teaching and learning strategies, software development and hardware development requirements, and module prototype evaluation. The final phase is implementation and evaluation, which focuses on determining the module's effectiveness in the actual teaching and learning process. Each finding is organised and documented more systematically and orderly in accordance with the DDR phase in order to produce more meaningful research results. The conclusion of this article proposes a conceptual framework for the research.**

Keywords*—Flipped classroom; project-based learning; design and development (DDR); Isman instructional design model

I. INTRODUCTION

The establishment of quality students with a high level of skills and thinking ability to face the wave of global revolution of the 21st century can only be achieved by implementing a quality education system and in line with technological developments [1]. Thinking skill is a skill or ability of a person to use the intellect to understand, explore various ideas, make judgments, make decisions and solve problems [2]. These skills include analytical skills, synthesis skills, evaluative skills, and the ability to generate new and noble concepts to solve problems [3], [4]. In fact, mastery of a wide range of cognitive knowledge and skills, including problem-solving skills, reasoning and creative and innovative thinking, should be inculcated and instilled in each student's identity. These skills are needed to prepare students who are always ready to face increasingly complex, fast-paced, and highly challenging challenges now and in the future. Statistics show that 72.8% of the country's labor market for low-skilled and semi-skilled workers still depends on the skills of foreign workers. This is due to the existence of a 21st-century skills gap among students regarding skills and knowledge. Trilling and Fadel [5] found that seven essential skills of the 21st century are poorly applied among students at secondary and tertiary levels, namely, oral and written communication, critical thinking, problem-solving skills, professionalism, teamwork, use of technology, leadership and project management.

The adaptation of design thinking skills becomes a part of the demand in the industrial and commercial areas. These fields need workers with problem-solving skills that can solve problems within multiple domains and can predict what the new issues will be [6], [7]. Design thinking began to emerge in the late 20, in which the idea was explored by developed countries, such as Singapore and Korea to face obstacles and difficulties of managing the economy. Another country, such as Denmark combined design thinking with a social science approach to create a novel solution for society [8]. Additionally, countries in the Asia Pacific region, such as China, South Korea and India promoted design thinking in university education through programs that concentrated on cultivating design thinking [6].

A. STEM Education

The demands of the 21st-century industry in Industrial Revolution 4.0 are now more focused on students whose theoretical mastery and practical skills alone are no longer sufficient to thrive as workers in the 21st century [9], [10]. The UNESCO agenda, which places creativity and innovation as the key to sustainable development requires significant changes in economics, technology and society due to the emergence of a digital economy that encompasses innovation skills, creativity and inventors. needs to be handled efficiently [11]. Theoretical knowledge alone is not enough to make students competent and competitive. Students also need to have high skills, especially in practical or hands-on skills and skills in problem-solving, creative thinking, written and oral communication, and teamwork [10].

Education-based in Science, Technology, Engineering, and Mathematics, known as STEM is introduced and seen as a catalyzer for education transform to boost the quality of education in many countries. STEM education aims to build the workforce in the STEM domain and served as a critical

vehicle for addressing the primary challenges of the 21st century [12]. The report of STEM Committee of the United States (US) National Council for Education, Science and Technology (2013) shows that to maintain American excellence in design and innovation, training and skills in STEM must be integrated in excellent working order. Even global companies in America promote the implementation of the STEM education system to ensure those job requirements in STEM are satisfied [13].

Design thinking skills and practices are quintessential ingredient in STEM education as it can benefit students' creativity and innovation in solving a problem. Such skills include analyzing the problems, creating and developing prototypes or models, gathering feedback, and redesigning creatively to solve the problems identified [14]–[16]. In other words, the advantage of one who practices design thinking proves that he or she is a competent individual and is always ready to meet the government's aim of developing human capital and the ability to become a globally competitive designer. Design thinking is also regarded as a particular skill that the students must master to match the needs of 21st-century student-centred learning in problem-solving [7].

B. Design and Technology

In Malaysia, design thinking starts to be deemed inevitable in the education system as a value-added to students to embrace challenges in the future. However, knowledge of theory is solely insufficient because other skills are pertinent to make the students competent, competitive, as well as creative and critical thinkers. Therefore, advancements in learning strategies, such as blended or hybrid learning, must intensify the students' competencies in the era of Industrial Revolution 4.0 [17].

Design and Technology (D&T), also known as Reka Bentuk dan Teknologi (RBT) is a transformation in the education system in Malaysia from Secondary School Integrated Curriculum (KBSM) to Secondary School Standard Curriculum (KSSM). The course was offered in the high school syllabus in 2017 to provide students with creative thinking and the ability to integrate current technology in problem-solving.

Furthermore, the implementation of D&T emphasises on how to design and produce quality products that are marketable [18]. Nevertheless, according to a report by Curriculum Development Division (BPK), the Ministry of Education Malaysia (MOE) addresses some issues and challenges that need to be overcome to enhance the quality and effectiveness of D&T learning in schools [19] as shown in Table I.

The 21st-century teaching and learning strategies are an approach that can be used to enhance teachers' pedagogical skills and student competency. In this regard, teachers' pedagogical approaches should be diversified and related to the use of the latest technology. These approaches encompass planning activities, conducting research, analysing data, and communicating the information obtained [13]. Furthermore, this approach can help to address problems related to the time and teaching resources limitations faced by the D&T teachers.

The flipped classroom is a teaching strategy in 21st-century learning that uses blended learning methods based on online technology to improve students' knowledge and performance for engineering-related subjects, such as D&T. Besides that, with the flipped classroom approach, problems related to time constraints, achievement, interest, knowledge, and students' motivation towards learning are resolved [20], [21]. Therefore, this approach is also suitable for D&T subjects, especially the topic of the Design Process based on time constraints and others, as stated in the previous paragraph.

The main concept of flipped classroom approach according to Bergmann & Sams [22] emphasizes six main features, namely: (i) shift teaching time; (ii) lectures can be accessed anywhere; (iii) active student involvement in the classroom; (iv) class workshops; (v) teachers as facilitators; and (vi) adapting learning using technology. These characteristics require teachers to think and apply new ideas that lead to learning objectives, learning processes and activities, assessment and future learning.

Project-based learning is one of the innovations introduced in the education system, which is an approach in the teaching and learning process that is student -centered. This learning approach encourages students to be active in the process of collaboration, communicating with each other in a small group to perform project assignments [23]–[27].

Project-based learning approach is an activity planned using problem -based learning (PBM) and inquiry learning approaches (BPK 2017). This method refers to activities that require students to identify methods to solve the problems presented and subsequently plan the entire project. Students take full responsibility for their project assignments, produce projects or artifacts as learning outcomes and present their assignments in front of peers and teachers for feedback.

TABLE I. D&T ISSUES AND CHALLENGES

Issues and Challenges	**Description**
Student Readiness	• Students are still weak in the development and production of design ideas. • Student readiness is low due to a lack of exposure to basic practical practices such as soldering, hammering, etc. • Student readiness - Poor fundamental knowledge displays a lack of basic knowledge, for example, less skilled in using hand tools.
Teaching and Learning Resources (Teaching aids, Module, Handbooks)	• Lack of teaching and learning resources is discovered. • Training received during the course is not sufficient • Modules and handbooks should be available for reference or trigger ideas for teachers.
Teacher Readiness.	• Teachers need specialized courses. • A more detailed, more comprehensive guide should be given. • Teachers do not receive proper training and exposure.

Project-based learning also can provide an active learning environment. A productive learning environment will be able to enhance students' knowledge through their experiences while engaging with learning activities and design project development in the classroom [2], [27]–[31]. Therefore, in this study, students' understanding of topics is crucial to determine the effect of applying design thinking skills on the development of students' knowledge through the implementation of flipped classroom module activities. Furthermore, even the integration of technology in project-based learning in teaching and learning and the role of teachers as facilitators help the students develop their creative thinking to build the concepts learned using Internet technology as a medium of information delivery and obtaining various sources of learning materials.

II. Design and Development of Flipped Classroom and Project Based Learning Module

The conceptual framework is a description of the idea of the entire study conducted [32]. It is also a backbone of the research that aims to explain how an idea formed will drive the planning and implementation of a study. This module development is based on the Design and Development Research by Richey and Klein [33], which is a systematic technique used to develop teaching modules. It involves a process that includes needs analysis activities, determining what must be mastered, creating educational goals, material design to achieve objectives, and implementing and evaluating the teaching materials' effectiveness [33][34]. This approach contains three main systematic phases: the needs analysis, development design, and evaluation phase, as mentioned in Fig. 1.

A. Phase 1- Need Analysis

A needs analysis was the first phase in DDR. A needs analysis was a critical stage in developing a product, in which information could be obtained through customers directly or indirectly[35]–[37]. It was intended to look at the problems that arose to predict solutions to future customer needs. Environmental information among the selected population was collected and analysed to identify the matter's needs. This phase also focused on what should be doing compared to what had been done in a study that identified the need to develop flipped classroom and project-based learning modules on knowledge, skills and design thinking of D&T students.

Discrepancy Model by McKillip [38] will used as a model of the needs analysis phase. The Discrepancy Model was a model used in the field of educational research. This model emphasised several expectations, namely the process of setting goals, the method of measuring performance that involved identifying what should be done and identifying discrepancy (discrepancy identification) that should have happened (what ought to be) and what exactly a problem was (*what was*). In the context of this study, needs analysis helped to obtain information about the need to develop flipped-classroom and project-based learning modules from the perspective of users, namely teachers and students based on the following research questions, namely, (i) Exploring the need for the application of design thinking for D&T subjects based on the teacher's perspective; (ii) Exploring teachers views on the need for flipped classroom and project-based learning module to apply design thinking among D&T students; and (iii) What is the level of readiness of the D&T students towards the implementation of flipped classroom and project based learning module.

In this study, mixed data collection methods were employed in this needs analysis phase. The methods comprised semi-structured interviews, surveys and document analysis. This data collection method was used to obtain evidence or information from various sources to answer research questions[39]. The qualitative data collected provided an in-depth overview of the phenomena studied [42]. The overview consisted of requirements required in the process of design and development of modules. The interview method was an effective way to obtain information on opinions, thoughts, views, and experiences. The interview was also carried out to understand what was experienced and thought by the informants [43]. In addition, document analysis, such as data analysis on Sistem Pentaksiran Tingkatan Tiga (PT3) examinations for Form 3 students in 2019, Dokumen Standard Kurikulum dan Pentaksiran (DSKP) D&T subjects and textbooks were also used to obtain justification and information to support the data collected. The survey method was chosen because it was a method of study that was planned following standard practices and exhibited a high level of trustworthiness. Apart from that, this method is sufficient to obtain views and needs about the issue stated and resolved [36]–[38]. The findings of this needs analysis justify that there is a need to develop teaching modules that focus on improving knowledge, skills and application of design thinking among lower secondary school students in D&T subjects.

B. Phase 2- Design and Development

The second phase of design and development was an essential part of this research. Ven Den Akker, Gravemeijer, McKenney and Nievee [32] explained that this phase was crucial and should be emphasised because the developed products, whether modules, models, or curriculum, were relevant and needed to undergo detail to ensure it benefited the actual target audience.

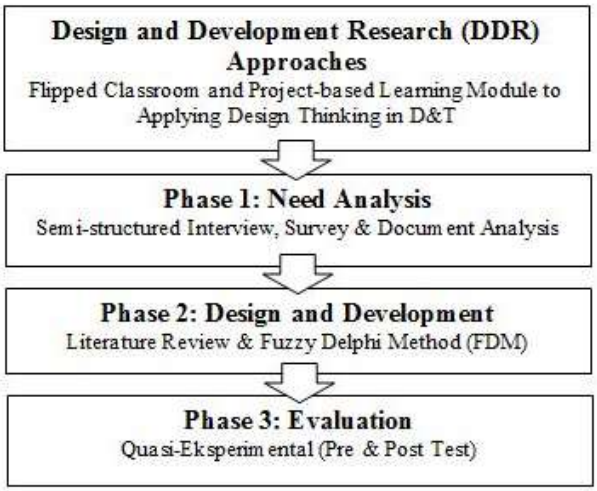


Fig. 1. Research Flow Chart.

In this study, Isman instructional design model [40] and Gagne's Nine Events of Instruction [43] was employed to develop the module. This model contained several processes suitable for use in module development. Isman's model in Fig. 2 was an instructional design model that emphasised ways of planning, developing, implementing, evaluating, and organizing productive learning activities in ensuring student achievement [40]. The theoretical foundation of this study came from Behaviorism, Cognitivism and Constructivism views. The adaptation of this model was centred on the relationship between stimulus and student response to strengthen the students' knowledge of the environment in Behaviorism theory. Next, it was related to the mental learning process and the students' experience to enhance the students' motivation for a more effective learning process. As such, teaching activities in the model gave emphasis to the students on how to retain the knowledge gained over the long term.

The adaptation of these models in previous studies as a framework in the design and development of teaching modules had been proven through the successful development of teaching modules for Physics subjects [41] and graphic design learning modules for students with hearing problems [42]. Therefore, the selection of these models in this study was appropriate and in line with the primary objective of developing this module to improve the knowledge, skills and design thinking of students who took D&T subjects. Table II illustrates the adaptation of the Isman instructional design model used in this study for the design and development phase. This relationship explained the need to implement to ensure that the project-based module development process used a flipped-classroom approach to student design thinking in the planned RBT subjects.

The first part of the design phase is the formation of key components and elements carried out through literature review and preliminary validation with several experts [43] to determine the list of components and elements involved in the design and development of module.

Learning theory and teaching models in designing and developing a teaching material must be given careful emphasis so that the modules produced will meet the learning objectives and can be implemented well. Such approaches and models are Constructivism Theory, Cognitive Theory with Multimedia Learning (Mayer 2009), Isman Teaching Model (Isman 2011), Gagne's 9 Step Teaching Model (1985), Reverse Classroom Model (Enfield & State 2013), Project Based Learning (Katz & Chad 2000) as well as the Design Thinking Model (Institute of Design. Stanford 2009) as in Table III.

To see to what extent components and elements are appropriate, components and elements will go through a content validation process [44]. The elements that have been formed are next through a content validation process that can see the extent to which the elements of design thinking that have been developed have been successfully defined. To obtain content validity, researchers have used an approach through expert evaluation as suggested by Creswell (2012) and Johnson and Christensen (2020) [45], [46].

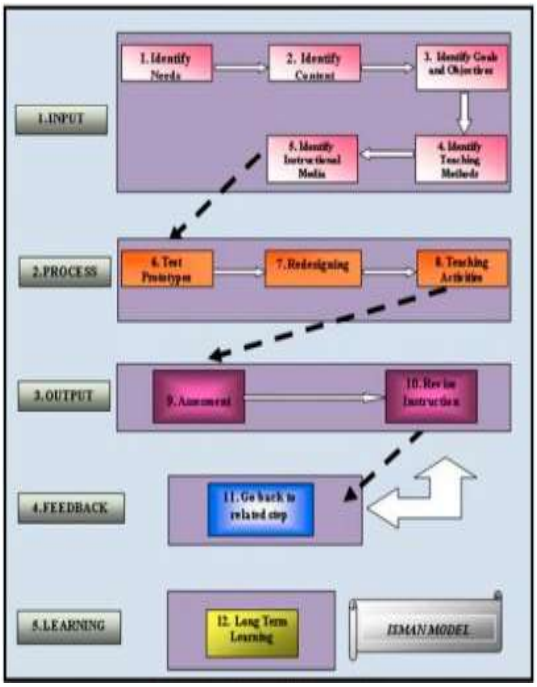


Fig. 2. Isman Instructional Design Model

TABLE II. ADAPTATION OF ISMAN INSTRUCTIONAL MODEL IN DESIGN AND DEVELOPMENT PHASE

Step	Work Log	Description
Input	• Identify needs • Identify content • Identify learning objectives • Identify teaching methods • Identify assessment materials • Identify teaching media	• Design flipped classroom and project-based learning module to enhance knowledge, skills, and design thinking based on expert opinion. • Obtain expert agreement on the design of the developed module.
Process	• Prototype test • Redesign instructions • Teaching activities	• Development of module prototype based on expert consensus through Fuzzy Delphi Method (FDM).
Output	• Testing • Result analysis	• Perform a pilot test
Feedback	• Check instructions	• Review the comments given by students and teachers
Learning	• Learning	• Conduct the quasi-experimental to evaluate the effectiveness.

TABLE III. ADAPTATION OF TEACHING AND LEARNING MODEL IN DESIGN AND DEVELOPMENT PHASE

Learning Theory	Model
Social Constructivism Theory	Isman Instructional Design Model (Isman 2011)
Cognitive Theory with Multimedia Learning (Mayer 2009)	Gagne's 9 Events of Instruction (1985)
	Flipped Classroom Model (Enfield & State 2013)
	Project-based Learning Model (Katz & Chad 2000).
	Design Thinking Model (d. school Stanford 2009)

Therefore, a total of three field experts were selected consisting of design thinking experts, pedagogics and curriculum experts who acted to evaluate, examine the measurement of constructs, content or scale and then see how much a construct is relevant or related to the concept being measured [46]. Then, these already formed elements will be carried to the next phase.

Next, the Fuzzy Delphi Method (FDM) was used in the module development phase to obtain the agreement of a group of experts to confirm, evaluate, reject, and add each component and element received in the previous stage before developing the module prototype. The strength of this method involved the diversity of experts in determining and validating the components and elements selected in the development of the module, whether it was appropriate in the context of the study conducted.

This method was a method and instrument of measurement that improved the Delphi Technique that Murray, Pipino and Gigch introduced in 1985. Fuzzy Delphi Method or FDM was a combination of fuzzy set theory and the traditional Delphi method added by Kaufman and Gupota in 1998 [20], [47]. This improvement made FDM a measurement tool to be more productive and solve problems that had precision and uncertainty for a study [36]. This method was chosen because it was proven to obtain expert agreement to decide components and elements needed in the module development.

The selection of selected expert panels was based on the expertise and experience of the expert panel and focused on the areas studied. Berliner [36] stated that the selected specialist must have more than five years of consistent experience, and this would provide in-depth results related to the issues being studied. In this study, the panel of experts appointed comprised lecturers and teachers from various fields, such as Educational Technology, Curriculum, and D&T.

Before that, a content validation process will be conducted to see the suitability of the components and elements was proposed [44]. To that end, researchers will use expert assessment as proposed by Creswell [45] and a total of three field experts consisting of design thinking experts, pedagogy and curriculum experts will be selected to evaluate, examine the construct measurements. Fig. 3 displays a flow chart using the Fuzzy Delphi (FDM) method to obtain empirical findings.

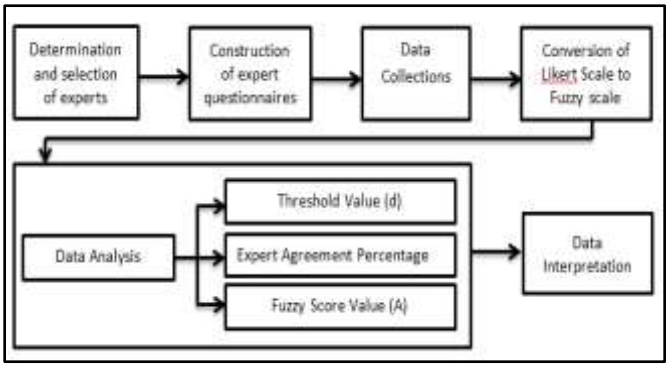


Fig. 3. Fuzzy Delphi Method (FDM) Approach.

In order to run the Delphi Fuzzy Method, several steps need to be followed to ensure that this study is considered an empirical study. There are:

Step 1: Formation of questions for the Fuzzy Delphi questionnaire was adaption from the previous research. In this step, the researchers have developed a questionnaire based on the literatur review and discussions with experts. The Likert scale of the questionnaire is based on the requirements of the research questions required to develop a flipped-classroom and project-based learning module to apply design thinking among RBT students. The developed questionnaires were submitted to the experts through two methods [37], namely, (i) meet face to face each expert and (ii) make online dissemination such as via email that has been identified as knowledgeable in the field studied.

Step 2: All linguistic variables are converted into fuzzy triangular numbers. The fuzzy rij number is a variable for each criterion representing the K expert.

$i= 1\ldots m, j=1,\ldots n, k=1\ldots.,k$ and $r_{ij} = 1/K\ (r^1_{ij} \pm r^2_{ij} \pm r^K_{ij})$ (1)

Step 3: Data has been converted to a fuzzy scale using a Delphi Fuzzy Analysis template developed from Microsoft Excel software. The threshold value, (d), will be calculated based on the following formula:

$$d(\tilde{m}\tilde{n}) = \sqrt{\frac{1}{k}[(m_1 - n_1)^2 + (m_2 - n_2)^2 + (m_3 - m_3)^2]}$$

Step 4: According to Cheng and Lin [48], if the expert evaluation data is less than the threshold values of 0.2, all experts consider that all experts have reached a consensus. Apart from that, if the group consensus percentage is more than 75%, then the following data analysis is to use the Defuzzification Process to obtain the fuzzy score value (A). The value of the fuzzy score (A) must be greater than or equal to the median value (α - cut value) of 0.5[49]. If the data findings are equal to or greater than 0.5, this means that the item is accepted by expert agreement.

A= (1/3)*(m1+m2+m3)

Through of the consensus experts obtained, prototypes of flipped-classroom and project-based module learning were developed. This phase involved restructuring the content program, organisation chart, storyboard, flowchart program, screen design, and evaluation process and repetition. Before the actual group thoroughly utilised module application, a pilot study on a group of students was conducted to identify problems that arose in the developed module.

C. Phase 3- Evaluation

The final phase in development design research is to evaluate the effectiveness of the module. Russell [48] stated that the evaluation of a developed module can be determined through activities or questions. There are three types of assessment that can be used to evaluate the entire module, namely, formative, summative and validation assessment [50] which is done during the teaching process, at the end of the teaching process and after the teaching and learning process.

There are three main research questions in this phase, namely, (i) does the module help teachers apply design thinking from the aspect of teacher teaching and learning strategies?; (ii) does the module help to increase the knowledge of the Design Process among students?; (iii) does the module help to improve Design Process skills among students? and; (iv) Does the activities of the module help to inculcate design thinking among students? Evaluation focuses on through user evaluation of module activities implemented in actual teaching and learning process.

In this study, Quasi-experimental studies were used to determine the effect of pre-and post-test interventions between the two test groups. The implementation of this approach is done in the classroom i.e., real situations, to determine whether changes made in small group assessment can effectively be applied in real contexts [45]. To ensure that treatment had an effect on student performance, pre-tests were given to the two groups as a baseline and to show that both groups had similar cognitive development. Then a post-test will be given after treatment, and the difference between the pre-test and post-test scores will indicate whether there is a measurable effect after treatment. Fig. 4 shows the flow chart of the module effectiveness evaluation study procedure.

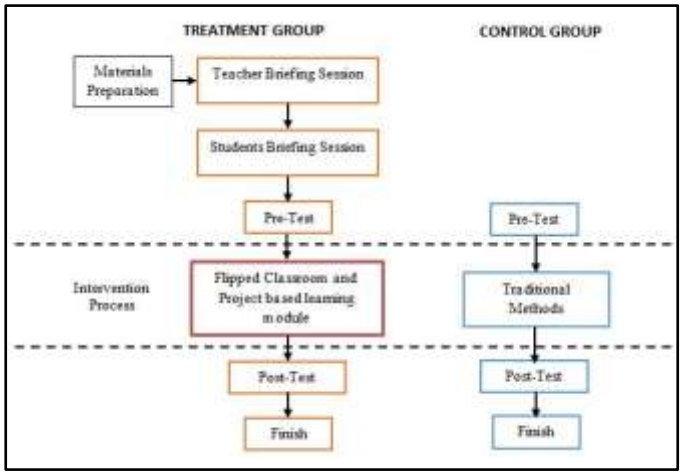


Fig. 4. Evaluation Phase Procedures.

The main difference between quasi-experimental and pure experimental studies is that the control and treatment student groups were not randomly selected due to several factors such as cost, time, logistics and evaluation of school administrators [47]. These differences cause the experimental study to be exposed to various threats or disruptions during the implementation process. Threat or interference of validity refers to interference that may threaten the study process being conducted and may cause the results obtained to possibly reflect false conditions of cause and effect between the treatment and control groups [45]. There are two types of experimental validity namely internal validity and also external validity.

Internal validity refers to the presence of differences on dependent variables as a result of manipulation of independent variables [47]. Internal effects need to be performed to ensure that the study findings are truly from the treatment of independent variables that are not influenced by any external variables.

External validity refers to the fact that the results of a study can be adopted by other groups with other environments and at different times as long as the characteristics of the study remain the same. It involves the results of the study findings can be generalized to other samples and study sites [47]. External threats refer to problems that interfere with researchers making correct inferences from sample data on others, past and future settings and situations [45].

The study sample in this phase is purposive sampling, that is, the study sample consists of two groups of students who take the subject of Design and Technology in a school. The minimum study sample of each group involved was a total of 35 students. Characteristics such as age, the number of students in the class, learning environment, study time and also teacher qualifications were determined in this study. Once the equivalent characteristics are set for both groups, random assignments were given to determine which class would act as the experimental group and the control group. The two selected groups had similar characteristics at the beginning of the quasi-experimental procedure, and this step was important to reduce bias (Kim & Steiner, 2016) in the quasi-experimental procedure.

III. Research Conceptual Framework

In designing and developing quality flipped classroom modules, various aspects needed to be a concern. The combination of several theories and learning models in the design and development study process had provided detailed and systematic guidelines according to several vital phases, namely Needs Analysis, Design and Development, and Evaluation. Fig. 5 shows this study's conceptual framework, which eventually formed a flipped classroom and project-based learning module to enhance the students' knowledge, skills, and design thinking in RBT subjects. The basis of the flipped classroom approach was to expose and cultivate the students to be prepared with learning materials outside of study time before attending the class[21], [51], [52]. Thus, the students were ready with the fundamental knowledge through this approach and applied the knowledge gained in discussion sessions and group activities.

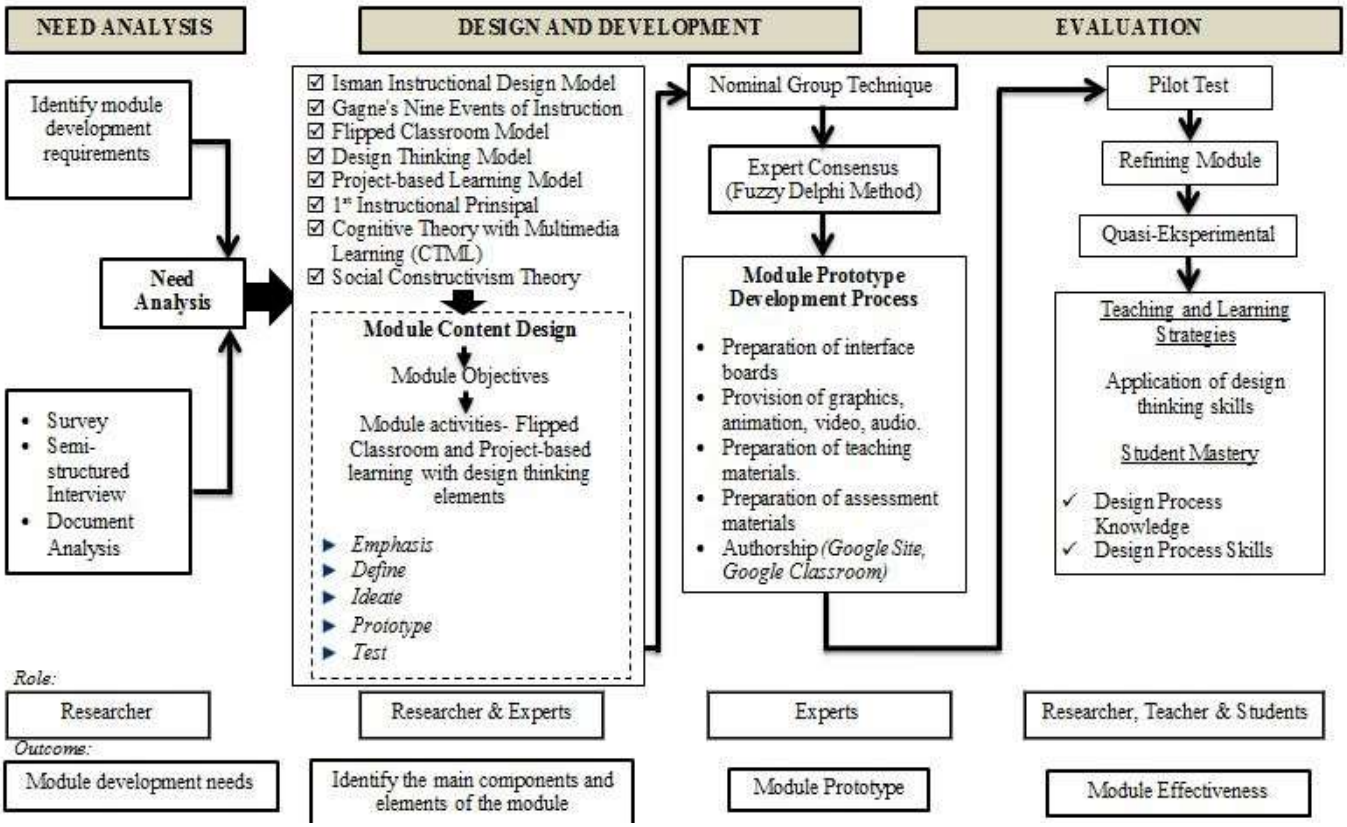


Fig. 5. Research Conceptual Framework (Adaptation Morisson II Model).

IV. Conclusion and Recommendation

Implementing the flipped classroom and project-based learning approach in the teaching process will help develop and cultivate the students' thinking and behaviour in a more positive and creative direction. Apart from that, the production of this module will make the classroom environment more active with the use of imaginative methods that actively involve the students in learning activities despite their different backgrounds. Moreover, through technological advances and implementing new policies in education, teachers are encouraged to redefine strategies and practical approaches in the teaching process. As a result, they are more competitive, as well as have creative and critical thinking.

This study is related to the process of design and development of flipped classroom and project-based learning module for the D&T subjects. This adaptation involves three main phases in Design and Development Research (DDR): needs analysis, design and development, and evaluation. This developed module will present different insights from the existing flipped classroom approach. The application of design thinking and the emphasis on skills and knowledge of the design process will make the students better prepared to face the challenges of today's world. In conjunction with this awareness, it is hoped that this study will be able to contribute to the formation of future students to be more competitive and have creative and critical thinking.

Acknowledgment

We would like to thank all those who have supported and made this paper a success. Special thanks to the Ministry of Education Malaysia for the Hadiah Latihan Persekutuan (HLP) and University Kebangsaan Malaysia to carry out the study.

References

[1] M. F. Mohd Rosdi, S. Abdul Kadir, and M. I. Nazri, "Tahap Kemahiran dan Kecenderungan Pemikiran Kritis Murid Tingkatan Dua Dalam Mata Pelajaran Kemahiran Hidup Bersepadu (KHB) di Negeri Kedah dan Hubungannya," Asia Pacific J. Educ. Educ., vol. 32, pp. 45–60, 2018.

[2] K. A. Kadir, N. M. Nordin, and Z. Iksan, "Reka Bentuk dan Pembangunan Modul Sains (e-SMART) Berasaskan Pendekatan Konstruktivisme 5E dan Analogi : Satu Alternatif Strategi Pengajaran Bagi Menerapkan Kemahiran Berfikir Kreatif Murid," vol. 1, p. 80, 2017.

[3] S. Z. Salleh and A. S. Shaari, "Pelaksanaaan Kemahiran Berfikir, Sikap Pelajar dan Masalah Untuk Mengaplikasikan Kemahiran Berfikir dalam Pengajarandan Pembelajaran Bahasa Melayu," vol. 9, no. November, pp. 1–10, 2019.

[4] R. E. Mayer, "2003_Cognitive Theory of Multimedia Learning," Educational Psychologist, vol. 38, no. 1. pp. 43–52, 2003.

[5] B. Triling and C. Fadel, "Century Skills," 21St Century Ski., no. Book, p. 48, 2009.

[6] Toshiaki Kurokawa, "Design Thinking Education at Universities and Graduate Schools," Sci. Technol. Trends, vol. 46, pp. 50–63, 2013.

[7] K. L. Cook and S. B. Bush, "Design thinking in integrated STEAM learning: Surveying the landscape and exploring exemplars in elementary grades," Sch. Sci. Math., vol. 118, no. 3–4, pp. 93–103, 2018.

[8] K. K. Hewitt, W. Journell, and R. Zilonka, "What the flip: impact of flipped instruction on self-regulated learning," Int. J. Soc. Media Interact. Learn. Environ., vol. 2, no. 4, p. 303, 2014.

[9] S. Ismail, "Kesediaan guru terhadap pelaksanaan mata pelajaran Reka Bentuk dan Teknologi (RBT) sekolah rendah di Malaysia," Universiti Tun Hussein Onn Malaysia, 2012.

[10] M. Lynch, U. Kamovich, K. K. Longva, and M. Steinert, "Combining technology and entrepreneurial education through design thinking: Students' reflections on the learning process," Technol. Forecast. Soc. Change, no. June, p. 119689, 2019.

[11] A. Athanassios and B. Vasiliki, "education sciences Developing and Piloting a Pedagogy for Teaching Innovation , Collaboration , and Co-Creation in Secondary Education Based on Design Thinking ," Educ. Sci., no. i, pp. 1–11, 2019.

[12] R. W. Bybee, "What is STEM education?," Science (80-.)., vol. 329, no. 5995, p. 996, 2010.

[13] C. Yata, T. Ohtani, and M. Isobe, "Conceptual framework of STEM based on Japanese subject principles," Int. J. STEM Educ., vol. 7, no. 1, 2020.

[14] Y. Li et al., "Design and Design Thinking in STEM Education," J. STEM Educ. Res., vol. 2, no. 2, pp. 93–104, 2019.

[15] R. Razzouk and V. Shute, "What Is Design Thinking and Why Is It Important?," Rev. Educ. Res., vol. 82, no. 3, pp. 330–348, 2012.

[16] B. Hokanson, A. Gibbons, D. Thinking, and D. Process, Design in Educational Technology. 2014.

[17] Rusli, A. Rahman, A. S. Ahmar, and Hastuty, "The design of digital teaching material of higher education in industrial revolution 4.0," J. Crit. Rev., vol. 7, no. 1, pp. 298–299, 2020.

[18] M. Malaysia Education Blueprint, "Malaysia Education Blueprint 2013 - 2025," Education, vol. 27, no. 1, pp. 1–268, 2013.

[19] K. Unit Penyelidikan dan Pentaksiran Sekolah, "Kajian Tahap Pelaksanaan Kurikulum Standard Sekolah Menengah (KSSM)," 2019.

[20] Zanariah Ahmad, "Pembangunan Modul Pedagogi Kelas Berbalik Berasaskan Pembelajaran Reflektif Untuk Politeknik Premier," 2017.

[21] M. Alias, Z. H. Iksan, A. A. Karim, A. M. H. M. Nawawi, and S. R. M. Nawawi, "A Novel Approach in Problem-Solving Skills Using Flipped Classroom Technique," Creat. Educ., vol. 11, no. 01, pp. 38–53, 2020.

[22] G. S. Mason et al., Flipped Classroom As Innovative Practice in the Higher Education System: Awareness and Attitude, vol. 3, no. SGEM2016 Conference Proceedings, ISBN 978-619-7105-72-8 / ISSN 2367-5659. 2012.

[23] J. D. Badia and V. M. Soria, Creative Project-based learning to boost technology innovation, vol. 29, no. 2. Springer Netherlands, 2019.

[24] S. H. Halili and S. A. Ramas, "Penerimaan Guru Sekolah Rendah Terhadap Pendekatan Kelas Berbalik Dalam Mata Pelajaran Bahasa Tamil," no. April, pp. 10–17, 2019.

[25] K. Missildine, R. Fountain, L. Summers, and K. Gosselin, "Flipping the classroom to improve student performance and satisfaction," J. Nurs. Educ., vol. 52, no. 10, pp. 597–599, 2013.

[26] S. J. DeLozier and M. G. Rhodes, "Flipped Classrooms: a Review of Key Ideas and Recommendations for Practice," Educational Psychology Review. 2017.

[27] A. S. Shaarani and N. Bakar, "A New Flipped Learning Engagement Model to Teach Programming Course," vol. 12, no. 9, pp. 57–65, 2021.

[28] N. Isa, M. Machmudi, M. A. I. Shihah, and H. J. Abdullah, "Pembelajaran Berasaskan Projek : Takrifan , Teori dan Perbandingannya dengan Pembelajaran Berasaskan Masalah," CREAM -Current Researc in Malaysia, vol. 2, no. 1, pp. 181–194, 2013.

[29] J. M. Allison, "Project Based Learning to Promote 21st Century Skills : An Action Research Study (Doctoral Dissertation)," no. May, 2018.

[30] A. Nordin, S. N. Salleh, and J. H. Abdullah, "Inovasi Pembelajaran Berasaskan WBL dan Triple Helix dalam Pembelajaran di Politeknik," vol. 3, pp. 111–120, 2018.

[31] A. Aziz Hussin, T. Sarifah Aini Syed Ahmad Akademi Pengajian Bahasa, U. Teknologi MARA, and S. Alam, "Designing a Flipped Classroom Lesson Using the AOCAR Technique," no. October, 2019.

[32] J. Van den Akker, "Educational Design Research," Educ. Des. Res., 2006.

[33] R. C. Richey and J. D. Klein, "Developmental research methods: Creating knowledge from instructional design and development practice," J. Comput. High. Educ., vol. 16, no. 2, pp. 23–38, 2005.

[34] R. C. Richey, J. D. Klein, and W. a Nelson, "Developmental research: Studies of instructional deisgn and development," Handb. Res. Educ. Commun. Technol., pp. 1099–1130, 2004.

[35] S. Siraj, N. Alias, D. DeWitt, H. Zaharah, and M. J. Mohd Ridhuan, Design and Development Research: Emergent Trends in Educational Research. 2013.

[36] M. J. Mohd Ridhuan, "Pembangunan Model Kurikulum Latihan SkiVes Bagi Program Pengajian KejuruteraanPembelajaran Berasaskan Kerja," 2016.

[37] A. Abdul Muqsith, "Pembangunan Model Eni Berasaskan Aktiviti Inkuiri Bagi Program Latihan Kemahiran," 2018.

[38] Ramlan Mustapha, "Reka Bentuk Model Integriti Akademik Berasaskan Penghayatan Rohani," vol. ى, no. ى, p. ى, 2017.

[39] J. Creswell and V. Plano Clark, Designing and Conducting Mixed Methods Research. Sage Publication, 2007.

[40] A. Işman, "Instructional design in education: New model," Turkish Online J. Educ. Technol., vol. 10, no. 1, pp. 136–142, 2011.

[41] N. Alias and S. Siraj, "Effectiveness of Isman Instructional Design Model in Developing Physics Module based on Learning Style and Appropriate Technology," Procedia - Soc. Behav. Sci., vol. 64, pp. 12–17, 2012.

[42] Ibrahim Zainuddin, "Pembangunan Modul Pembelajaran Seni Reka," p. 429, 2017.

[43] C. Okoli and S. D. Pawlowski, "The Delphi method as a research tool: An example, design considerations and applications," Inf. Manag., vol. 42, no. 1, pp. 15–29, 2004.

[44] U. Sekaran and R. Bougie, Research Methods for Business. 2014.

[45] J. Creswell, Educatyional Research : Planning, Conducting and Evaluating Quantitative and Qualitative Research - 4th ed, 4th ed. Pearson, 2013.

[46] R. B. Johnson and L. Christensen, Educational Research Quantitative, Qualitative, and Mixed Approaches Seventh Edition. 2020.

[47] N. M. Noh, S. Siraj, S. H. Halili, M. R. M. Jamil, and Z. Husin, "Application of fuzzy delphi method as a vital element in technology as a tool in design thinking based learning," Asia Pacific J. Educ. Educ., vol. 34, no. d, pp. 129–151, 2019.

[48] C. Cheng and Y. Lin, "Evaluating the best main battle tank using fuzzy decision theory with linguistic criteria evaluation," vol. 142, pp. 174–186, 2002.

[49] J. Cheng and C. Tang, "An Application of Fuzzy Delphi and Fuzzy AHP for Multi-criteria Evaluation on Bicycle Industry Supply Chains," vol. 4, no. 1, pp. 21–34, 2009.

[50] R. C. Richey and J. D. Klein, Design and Development Research Methods, Strategies, and Issues. 2007.

[51] J. Enfield, "Looking at the Impact of the Flipped Classroom Model of Instruction on Undergraduate Multimedia Students at CSUN," TechTrends, vol. 57, no. 6, pp. 14–27, 2013.

[52] A. A. Rahman, N. M. Zaid, Z. Abdullah, H. Mohamed, and B. Aris, "Emerging project based learning in flipped classroom: Technology used to increase students' engagement," 2015 3rd Int. Conf. Inf. Commun. Technol. ICoICT 2015, no. July, pp. 212–215, 2015.

Effective Service Discovery based on Pertinence Probabilities Learning

Mohammed Merzoug, Abdelhak Etchiali, Fethallah Hadjila, Amina Bekkouche
Computer Science Department
Abou Bekr Belkaid University of Tlemcen
B.P. 119 Faculty of Sciences, Tlemcen, Algeria

***Abstract*—Web service discovery is one of the most motivating issues of service-oriented computing field. Several approaches have been proposed to tackle this problem. In general, they leverage similarity measures or logic-based reasoning to perform this task, but they still present some limitations in terms of effectiveness. In this paper, we propose a probabilistic-based approach to merge a set of matching algorithms and boost the global performance. The key idea consists of learning a set of relevance probabilities; thereafter, we use them to produce a combined ranking. The conducted experiments on the real world dataset "OWL-S TC 2" demonstrate the effectiveness of our model in terms of mean averaged precision (MAP); more specifically, our solution, termed "probabilistic fusion", outperforms all the state of the art matchmakers as well as the most prominent similarity measures.**

***Keywords*—Service-oriented computing; web service discovery; rank aggregation; probabilistic fusion**

I. Introduction

The web service technology is actually involved in many applications, such as business processes management and recommendation systems [1]

Thanks to its modularity, composability and loose coupling, this technology is largely utilized in data integration and applications' composition. To ensure these objectives, one has to discover and rank the services that best meet her/ his needs. According to [2], the service discovery can defined as follows:

> Given a web service repository, and a query requesting a service (hereafter service query), finding automatically a service from the repository that matches these requirements is the web service discovery problem. Only those services that: 1) produce at least the requested output parameters that satisfy the post-conditions, 2) use only part of the provided input parameters that satisfy the pre-conditions, and 3) produce the same side effects can be valid solutions to the query.

Several approaches have been proposed in the literature for tackling the web service discovery problem [3]. Based on the works of [4], [5], we distinguish three types of discovery approaches: logic-based reasoning methods, non logic-based techniques (i.e. similarity measures, graph matching, datamining, etc.) and hybrid techniques which merge the logic and the non logic solutions. Despite the progress made in this field, much remains to be done to achieve an acceptable rate of performance. For instance, the logic-based approaches are often characterized by a poor recall rate (Since the underlying semantic of service interfaces can be implicit and not captured by the ontologies) [4]. On the other hand, the similarity measures do not have the same performance; in addition, the choice of the most relevant similarity is not obvious and generally it depends on the actual user's request. Furthermore, a lot of similarity measures may have hyper-parameters (e.g. the fuzzy similarity proposed in [6]) that need to be adjusted for the search; therefore, arbitrary initialization of these parameters is inappropriate and may entail misleading results. Consequently, we must utilize both types of matching algorithms to enhance the discovery performance. In this line of thought, the creation of a hybrid matching algorithm must address the following concerns:

1) How to solve the ordering conflicts entailed by the individual matching algorithms (for instance an algorithm may conclude that service S_1 is better than service S_2, while another may decide that S_2 is better than S_1)?
2) How to infer the most suitable matching algorithm for each user's request, and exploit this knowledge in the fused scheme?
3) How to boost both recall and precision, while preserving a tolerable execution time?

In this paper, we handle the aforementioned difficulties, by adopting machine learning and the theory of probability as a clue for combining the individual matching algorithms.

More specifically, given the m rankings provided by the m matching algorithms (or similarity measures), our machine learning algorithm derives a global ranking by calculating a fusion score for each service S_i; this score is weighted sum of the scores (denoted as $score_{ij}$ where j is the identifier of a matching algorithm) provided by the matching algorithms. Each $score_{ij}$ represents the probability that S_i is relevant to the current request; the more the value of $score_{ij}$ is high, the better the fusion score of S_i. With this fusion scheme, we can answer the abovementioned concerns. In particular, the ordering conflicts are resolved using the weighted sum (which can be considered as weighted vote). Additionally, the most suitable matching algorithms are those that have a higher weight and a higher value of $score_{ij}$ (see equation 22). These heuristic will ensure a good performance in terms of recall and precision. Moreover, if we assume that the m matching algorithms are independent and have a precision equal to p (where $p >= 0.5$), then according to the theorem of jury [7], a majority voting method (or a weighted voting method) will achieve a precision higher than p. In summary, our proposed solution is can be described as follows:

First, we divide each individual ranking into a set of segments. Second, for each segment, we compute its probability of relevance (i.e the probability of having a relevant segment member with respect to the current request). Third, we aggregate the aforementioned probabilities through a linear formula. To choose the ideal number of segments (ns) used in the second step, we perform a cross-validity that evaluate the mean averaged precision of the proposed model.

The remainder of the paper is organized as follows. In Section II, we review the state of the art. We formally define the problem in Section III. Section IV presents the probabilistic fusion algorithm. The results of the experimental study and threads to validity are presented in Section V. Finally, Section VI concludes the paper.

II. State of the Art

The web service discovery has received much attention in the recent years. In general, we discern three types of web service matchmaking approaches: logic matchmaking, non logic matchmaking and hybrid matchmaking [5], [3], [8].

A. Logic-based Matchmaking

The first category of matchning leverages pure logic reasoning, more precisely, the matchmaking utilizes the consistency tests or subsumption mechanisms to decide whether a relationship exists between the user request and the advertised service [9].

The work by [10] presents an automatic location of services (ALS) that allows for discerning five magnitudes of matching degrees (Match, ParMatch, PossMatch, NoMatch and PossParMatch).

In [11], the authors enhance the framework proposed in [10]; in particular, they add additional magnitudes of matching degree such as:

- *RelationMatch*: The advertised service does not meet the required outputs, but it offers outputs having a relation with them.
- *ExcessMatch*: The advertised service meets all the required outputs, but it offers supplementary outputs that are not needed by the user.

A logical matching framework is presented in [9]; this latter architecture takes into account almost all functional properties, including inputs, outputs, preconditions, and effects (IOPE).

-The major weakness of logic-based approaches are the high rate of false positives and false negatives [4].

In addition, the theoretical complexity of subsumption test is Pspace-complete or exp-time complete for certain portions of description logics [12].

B. Non-logic-based Matchmaking

Based on the fact that the aforementioned pitfalls discourage the research in this type of matchmaking, some scholars have developed a new type of solutions. These techniques [13] mainly leverage graph matching, data-mining, combinatorial optimization, and probabilistic matching.

The framework proposed by [8] matches the user's request against the OWL-S using the parameters of service name, service input, and service output. These attributes are first filtered using the part of speech (POS) procedure to eliminate the Stop Words, Special Characters, Numbers, and Uncategorized nouns. Then, the resulting terms are disambiguated using the Wordnet directory. At the end, these terms are matched using a Wordnet based similarity measure.

A new redescription of services is presented in [14]. The main idea consists of using dischlit probability distributions[15] and clustering [16] to provide a latent factor-based specification of services.

The iMatcher1 framework presented in [17] leverages the service profile to perform a syntactic matching of services ; more specifically, it uses four distance functions to match the request and the services (Term Frequency-Inverse Document Frequency [18], the Levenshtein similarity distance [19], the Cosine vector measurement [20], and the divergence measurement of Jensen-Shannon [21]).

In [22], the authors utilize fuzzy sets and rule based systems to tackle the web service discovery and selection problem. More specifically, the proposed work matches both capability attributes (functional aspect) and context attributes (non-functional aspect).

The work by [23] presents a collective dominance function to handle the QoS preferences of a set of users. This function is more flexible and enables the controle of the size of the service skylines.

In [24], the authors tackle the discovery of services. While taking into account the dynamic QoS properties. In particular, they leverage statistical time series to model the QoS fluctuations.

The work in [25] defines a composition framework by means of integration with fine-grained I/O service discovery that enables the generation of a graph-based composition which contains the set of services that are semantically relevant for an input-output request. The proposed framework also includes an optimal composition search algorithm to extract the best composition.

The work of [26] compare the semantic discovery approaches according to several criteria, such as interface type (e.g. OWL-S, WSMO), the scalability, the request expansion, the adopted similarity measure and the use of natural language processing.

The work by [27] proposes a two-stage discovery approach: an offline phase and an online phase. The input of the offline phase is a set of categorized services (most of the existing registries ensure this categorization (e.g. Programmable Web)). Each service is represented as a set of service goals. A service goal is a triple constituted of a verb, a core noun and optional parameters (such as adjectives and non-core nouns). The ensemble of service goals extracted from all services of each category are clustered into groups using K-means algorithm and Wordnet-based similarity measure.

In the online phase, the nearest category (with respect to the request) is retained and thereafter the user's request is expanded using the service goal clusters of the previous

category. At the end, the services of the target category are matched against the expanded query.

C. Hybrid Matchmaking

The third class aggregates the former categories in order to enhance the search quality. There are several ways for merging the aforementioned types: either by using machine learning or heuristics to tune the weights of the matching algorithms, or by using social choice theory to fuse the input rankings, or by leveraging probabilistic / fuzzy relationships to ensure the same purpose.

The most simple heuristic for merging a set of individual matching algorithms is to associate a fixed rank t (or priority) to each matching function.

The OWLS-MX framework [28] matches the inputs/outputs attributes of service profiles. This system proposes seven levels of matching degree (Plug-in, Subsumes and Subsumed-By) and hybrid matching (Logic-based Fail and Nearest-neighbor).

The work by [29] introduces a matchmaker for SAWSDL-based services. The approach leverages both subsumption test and information retrieval models for pairing the request and the advertised services.

The ISEM framework [5] is a hybrid matching approach that combines both the OWLS-MX3 filters and SVM-based learning for discovering services.

Merzoug and al. [3], fuse five matching algorithms (i.e. similarity measures) using a fuzzy dominance relationship [30].

In [31] the authors develop three probabilistic functions for searching and ranking web services. Each function involves multiple matching algorithms (logic, textual similarities, etc.).

In the same work [31], the authors show a comparative evaluation which involves several voting models, such as CombSUM, CombMNZ [32], Borda-fuse model [33], and outranking model [34]. According to the experiments, the CombMNZ system is better than the other voting models, but it is less effective than some individual matching algorithms (such as information loss).

In [35], the authors adapt also the Condorcet fuse model [36] to the service discovery problem. More specifically, they compare the partial scores provided by the individual matching functions through a fuzzified version of the dominance relationship [6]. The preliminary results show that the proposed approach largely outperforms the individual algorithms. However, the results can be largely boosted if a smart parameter tuning is performed.

In [37], the authors introduce a new context-based solution based on QoS (Quality of Service) exploiting both functional and non-functional user's requirements and providing the user ability to control and proceed with the discovery of web services, i.e. the main aim of this work is to locate the appropriate web service correspondence with the context of the user.

In [38], the authors propose a multi-criteria decision method (MCDM) for searching web services based on contextual attributes (e.g. location, language, and size of screen).

TABLE I. EXAMPLES OF PARTIAL MATCHING SCORES OF SERVICES

Services	Parameter	f_1	Services	Parameter	f_2
A	$score_{in}$	0.78	*B*	$score_{in}$	0.86
	$score_{out}$	0.84		$score_{out}$	0.80
	$mean$	0.81		$mean$	0.83
B	$score_{in}$	0.76	*A*	$score_{in}$	0.86
	$score_{out}$	0.80		$score_{out}$	0.78
	$mean$	0.78		$mean$	0.82
C	$score_{in}$	0.78	*C*	$score_{in}$	0.74
	$score_{out}$	0.60		$score_{out}$	0.62
	$mean$	0.69		$mean$	0.68

Since the standard similarity measures (such as Cosine and Extended Jaccard) are not suitable for handling contextual attributes, the authors propose a set of rules and a voting method to compare and rank services.

III. PROBLEM STATEMENT

A. Introduction

In the following, we present a motivating scenario that highlights the major difficulties encountered in web service discovery. We assume that a given user is interested by a service which accepts a set of input concepts $Pin_1, Pin_2...$ and provides a set of output concepts $Pout_1, Pout_2...$, (for the sake of simplicity we disregard for the moment, the other attributes such as preconditions or effects).

To achieve this purpose, the customer may utilize multiple matchmaking algorithms or similarity functions denoted by $f_1 \cdots f_n$. Each function is applied on the request/service parameters (in our case the inputs/ outputs). Let RQ be the request parameter set, i.e. $RQ = RQ_{in} \cup RQ_{out}$, where $RQ_{in} = \{Pin_1, Pin_2, ...\}$, $RQ_{out} = \{Pout_1, Pout_2, ...\}$. Similarly we define the parameter set of the advertised service S as follows: $AS = AS_{in} \cup AS_{out}$.

Each matchmaking function f_j matches the request parameters against the parameters of the advertised services by applying the following equations.

$$score_{in} = f_j(RQ_{in}, AS_{in}) \quad (1)$$

$$score_{out} = f_j(RQ_{out}, AS_{out}) \quad (2)$$

Equations 1, 2 compute the similarity degree between the inputs (resp outputs) of the request and the inputs (resp outputs) of the advertised service.

Table I shows two ranked lists produced by two matching functions f_1 and f_2. Each cell labelled with $score_{in}$ or $score_{out}$ indicates a partial matching score computed through Equations 1 and 2. These matching scores belong to [0,1]. The aforementioned (individual) lists are ranked according to the mean score.

For the sake of simplicity, we suppose that all services have a single input Pin and a single output Pout, the same assumption is considered for the request. By analysing the previous table, we notice the following findings:

First, the two rankings disagree about the ordering of the services A and B. Second, the resolution of the conflict by

computing the mean score over all partial matching scores (see the third line of each service) is not always a relevant heuristic. This solution may be erroneous for some user's requests.

Thus, the creation of an optimal ranking (which provides the highest precision and recall) is not obvious, since we must deal with the specificities of each request as well as the service position within each (individual) list.

As discussed above, each matching function is only effective on a subset of requests, and it may give a poor performance on the remaining requests. Consequently, it will be advantageous to combine a set of matching functions. By doing so, we leverage the advantages of the adopted matching techniques, and we boost the global performances.

To combine the individual matching algorithms, we have to aggregate the partial scores/ ranks of the services. Several aggregating schemes are proposed in the literature [28], [3]. These approaches may leverage voting based models, probability theory, fuzzy set theory, and machine learning.

To determine the most effective mechanism, we have to conduct an exhaustive comparative study and derive the optimal configuration of parameters.

B. Specification of the Discovery Problem

To facilitate the presentation of the problem, we assume the following notations:

let PRL_{ij} be a (partially) ranked list of the i^{th} request under the j^th matching function.

Formally:

$$PRL_{ij} =< (S_1, V_{1ij}), (S_2, V_{2ij}), ...(S_{|dataset|}, V_{|dataset|ij}) >$$

where, $dataset$ is the collection of services (i.e $S_1, .., S_{|dataset|}$) and $V_{kij} \in R^d$, each V_{kij} represents a partial matching score computed through Equation 1 or Equation 2. It measures the similarity between the parameters of the i^{th} request and the parameters of the k^{th} service using the j^{th} matching function. In this case, d is set to 2, since we have two descriptors for inputs and for outputs.

In the following, we specify the discovery problem as follows. Given:

- A set of matching functions $\{f_1, ...f_m\}$.
- A set of user's requests $Q = \{RQ_1, RQ_2, ...\}$; each request is represented by the union of the inputs concepts and the outputs concepts, i.e. $RQ_i = \{Pin_1, Pin_2, ...\} \cup \{Pout_1, Pout_2...\}$.
- A set of (partially) ranked lists, for each request $\{PRL_{11}, ..., PRL_{m1}, .., PRL_{1|Q|} \ldots, PRL_{m|Q|}\}$

We aim to produce a combined ranking (denoted $Combined\_Ranking_i$) for each request RQ_i, such that:

MAP($Combined\_Ranking_1, .., Combined\_Ranking_{|Q|}$) is maximized.

Where:

$Combined\_Ranking_i$: represents the fused list of the i^th request (RQ_i).

MAP: represents the mean average precision criterion. It is defined as follows:

$$MAP(Combined\_Ranking_1, .., Combined\_Ranking_{|Q|}) = \frac{1}{\sum_{i=1}^{|Q|} AveragePrec(Combined\_Ranking_i)} \quad (3)$$

and

$$AveragePrec(Combined\_Ranking_i) = \sum_{k=1}^{|dataset|} precision(Combined\_Ranking_i, k) * rel(k) \quad (4)$$

where $precision(Combined\_Ranking_i, k)$ is the precision at the k^{th} position over the i^{th} combined ranking.

and

$$rel(k) = \begin{cases} 1 & \text{if the service } S_k \text{ is relevant to the } i^{th} \text{ request} \\ 0 & \text{Otherwise} \end{cases} \quad (5)$$

IV. Web Service Discovery and Ranking

In what follows, we present our main contributions to solve this service discovery problem; in particular, we demonstrate the individual matching algorithms (Sections IV-A) as well as the probabilistic fusion scheme (Section IV-B).

A. Individual Matching Functions

In this work, we use the most promising matching functions of the information retrieval field. More specifically, we use five matching functions that are defined below. To match a request R with the service S, we introduce the following notation:

Let RQ be the parameters set of R. let V_{ir} (resp V_{or}) be the vector containing the occurrence numbers of the indexed inputs (resp outputs) of the request R. V_{ir} is derived from RQ_{in}; similarly, V_{or} is derived from RQ_{out}.

In addition, let V_{is} (resp V_{os}) be the vector containing the occurrence numbers of the indexed inputs (resp outputs) of the service S. V_{is} is derived from AS_{in} , similarly V_{os} is derived from AS_{out}. Based on the aforementioned vectors, we define the probability distributions P_{ir} (resp P_{or}) and P_{is} (resp P_{os}) as follows:

$$P_{ir}(k) = \frac{V_{ir}(k)}{\sum_{k=1}^{|V_{ir}|} V_{ir}(k)} \quad (6)$$

$$P_{is}(k) = \frac{V_{is}(k)}{\sum_{k=1}^{|V_{is}|} V_{is}(k)} \quad (7)$$

$$P_{or}(k) = \frac{V_{or}(k)}{\sum_{k=1}^{|V_{or}|} V_{or}(k)} \quad (8)$$

$$P_{os}(k) = \frac{V_{os}(k)}{\sum_{k=1}^{|V_{os}|} V_{os}(k)} \quad (9)$$

The first similarity measure is defined as follows:

$$sim1(R,S) = \frac{1}{2}(cos(V_{ir}, V_{is}) + cos(V_{or}, V_{os})) \quad (10)$$

where cos measures the proportion between the dot product of the compared vectors (or objects) and the product of their length. It is defined as follows :

$$cos(V_{ir}, V_{is}) = \frac{< V_{ir}, V_{is} >}{(||V_{ir}||.||V_{is}||)} \quad (11)$$

and $< X, Y >$ is the dot product operator, $||X||$ is the euclidean norm of X.

Similarly, for V_{or} and V_{os}:

$$sim2(R,S) = \frac{1}{2}(EJ(V_{ir}, V_{is}) + EJ(V_{or}, V_{os})) \quad (12)$$

where EJ (Extended Jaccard) computes the proportion between the size of shared elements and the cardinal of the union. It is defined as follows:

$$EJ(V_{ir}, V_{is}) = \frac{< V_{ir}, V_{is} >}{(||V_{ir}||^2 + ||V_{is}||^2 - < V_{ir}, V_{is} >)} \quad (13)$$

Similarly, for V_{or} and V_{os}:

$$sim3(R,S) = \frac{1}{2}(IL(V_{ir}, V_{is}) + IL(V_{or}, V_{os})) \quad (14)$$

where IL (Information Loss) is based on the percentage of elements that are not shared among the compared objects. The more the percentage is low, the better the similarity degree. It is defined for binary vectors as follows:

$$IL(V_{ir}, V_{is}) = 1 - [\frac{(\sum_{k=1}^{|V_{ir}|} MAX(V_{ir}(k), V_{is}(k)) - < V_{ir}, V_{is} >)}{(\sum_{k=1}^{|V_{ir}|} V_{ir}(k) + \sum_{k=1}^{|V_{is}|} V_{is}(k))}] \quad (15)$$

Similarly, for V_{or} and V_{os}:

$$sim4(R,S) = \frac{1}{2}(JS(P_{ir}, P_{is}) + JS(P_{or}, P_{os})) \quad (16)$$

where JS (Jensen–Shannon based similarity) is based on the estimation of the difference between two probability distributions that represent the compared vectors. The more the difference is low, the better the similarity degree. It is defined as follows:

$$JS(P_{ir}, P_{is}) = (\frac{1}{2}log2) * \sum_{k=1}^{|P_{ir}|} [h(P_{ir}(k)) + h(P_{is}(k)) - h(P_{ir}(k) + P_{is}(k))] \quad (17)$$

where $h(x) = -xlog2(x)$.

Similarly, for P_{or} and P_{os}:

$$sim5(R,S) = \frac{1}{2}(LOG(AS_{in}, RQ_{in}) + LOG(RQ_{out}, AS_{out})) \quad (18)$$

Where LOG (logic matching) is defined as follows :

$$LOG(RQ_{out}, AS_{out}) = MIN_{P_l \in RQ_{out}}(LogMatch1(P_l, As_{out})) \quad (19)$$

In addition:

$$LogMatch1(P_1, AS_{out}) = MAX_{P_k \in AS_{out}}(LogMatch2(P_l, P_k)) \quad (20)$$

In general, the logical comparison of two parameters P_u, P_t is established as follows:

$$LogMatch2(P_u, P_t) = \begin{cases} 1(Exact) & \text{if } P_u \equiv P_t \\ 0.95(plugin) & \text{if } P_u \text{ is parent of } P_t \\ 0.85(Subsume) & \text{if } P_u \sqsubset P_t \\ 0.75(Subsumedby) & \text{if } P_t \text{ is a parent of } P_u \\ 0(Fail) & \text{Otherwise} \end{cases} \quad (21)$$

This is done similarly for AS_{in} and RQ_{in}.

In the following, we present our probabilistic fusion scheme, which is constituted of 3 algorithms. The first one, hereafter referred to as **RPC** (Relevance Probability Computation), computes the knowledge that allows the fusion of the input lists. The second one is termed **PF** (probabilistic fusion), it produces the TopK elements of the combined (fused) ranking. The third one is termed **CVBT** (cross-validation-based tuning). CVBT leverages the cross-validation to select the optimal number of segments.

B. Proposed Algorithms

To build the combined ranking, we adapt the probabilistic approach proposed in [39], to the context of web services. In a nutshell, the basic idea consists of learning a set of probabilities that are involved in the computation of the fused score of each service. The more the fused score is high, the better the rank is. The algorithm performing this task is referred to as RPC. Each learned probability (denoted by $MRelP_{ri}(S_l)$) represents the likelihood that a service S_l returned in segment r is relevant, given that it has been returned by the matching function i.

Algorithm 1 represents the pseudo code of RPC.

RPC algorithm is explained as follows:

Algorithm 1: Algorithm RPC

Input: $dataset : a\ set\ of\ services$,
$SRQ : a\ subset\ of\ requests$,
$m : Integer\ (the\ number\ of\ matching\ functions)$,
$ns :\ Integer\ (the\ number\ of\ segments)$,
$Rel :$
$a\ binary\ matrix\ of\ dimension\ |requests|\ .\ |dataset|$
Output: $MRelP :\ a\ matrix\ of\ dimension\ ns.m$

1 **for** $i = 1\ to\ m$ **do**
2 $\quad ranking_i = EmptyList$
3 $\quad$ **foreach** $Q_j \in SRQ$ **do**
4 $\quad\quad$ **foreach** $S_l \in dataset$ **do**
5 $\quad\quad\quad score = sim_i(Q_j, S_l)$
$\quad\quad\quad\ insertInto(score, ranking_i)$
6 $\quad\quad decreasing\_sort(ranking_i)$
7 $\quad\quad relv\_services = extract(Rel, j)$
8 $\quad\quad$ **for** $r = 1\ to\ ns$ **do**
9 $\quad\quad\quad segment\_members = extract\_seg(ranking_i, r)$
10 $\quad\quad\quad relv[i][j][r] =$
$\quad\quad\quad\quad |relv\_services\_segment\_members|/|segment\_members|$
11 $\quad$ **for** $r = 1\ to\ ns$ **do**
12 $\quad\quad$ **for** $b = 1\ to\ |SRQ|$ **do**
13 $\quad\quad\quad MRelP[i][r] = MRelP[i][r] + relv[i][b][r]$
14 $\quad\quad MrelP[i][r] = MRelP[i][r]/|SRQ|$
15 $return\ (MRelP)$

- (Lines 1 up to 7), for each matching function i and request Q_j, we compute the corresponding ranking termed $ranking_i$.
- (Lines 8-9), we sort the aforementioned ranking and we get the relevant services of the request Q_j.
- (Lines 10-13), for each segment r, we extract its members, thereafter we compute its relevance probability by applying the formula of the precision criterion. This rule calculates the likelihood that a segment r derived from the function i is relevant to the request Q_j.
- (Lines 15-19), for each segment r and each matching function i, we compute their averaged relevance probability (also denoted $MRelP_{ri}$). More specifically, we take the mean of the relevance probabilities related to the requests of the learning set (SRQ).
- (line 22) We return the learned probabilities.

The second algorithm referred to as **PF** (probabilistic fusion) allows to compute a fused score for each service S_l. To this end, **PF** leverages the learned relevance probabilities of the five individual rankings. PF is based on two heuristics (H1 and H2); which are summarized as follows:

- The more the rank (or the segment identifier) of a service S_l is higher within the individual rankings, the more the fused score is better (H1). •
- The more the relevance probability $MRelP_{ri}(S_l)$ is higher, the more the fused score is better (H2).
This rule is explained as follows: if we assume that $MRelP_{ri}(S_l)$ is large, then service S_l is more likely to be relevant and, thus it should be ranked higher in the combined (fused) list. The fused score is summed up as follows:

$$Fscore(S_l) = \sum_{i=1}^{m} \frac{MRelP_{ri}(S_l)}{r} \tag{22}$$

where r is the segment identifier of S_l, i the identifier of the matching function, and m is the number of matching functions.

The pseudo code of **PF** is given in Algorithm 2.

Algorithm 2: Algorithm PF

Input: $Q_j : the\ current\ request\ to\ be\ handled$, $m :$
$Integer\ (the\ number\ of\ matching\ functions)$,
$k :\ Integer\ (size\ of\ the\ returned\ list)$,
$ns :\ Integer\ (the\ number\ of\ segments)$,
$MRelP : a\ matrix\ of\ dimension\ m.ns$
Output: $Top\_k(CombinedList) :\ an\ ordered\ list$

1 **for** $l = 0\ to\ |dataset| - 1$ **do**
2 $\quad Fscore[l] = 0$
3 $CombinedList = EmptyList$
4 **for** $i = 1\ to\ m$ **do**
5 $\quad ranking_i = EmptyList$
6 $\quad$ **for** $each S_l\ in\ dataset$ **do**
7 $\quad\quad score = sim_i(Q_j, S_l)$
8 $\quad\quad insertInto(score, S_l, ranking_i)$
9 $\quad decreasing\_sort(ranking_i)$
10 $\quad$ **for** $each\ S_l\ in\ dataset$ **do**
11 $\quad Sid = Get\_Seg\_ID(S_l, ranking_i)$
12 $\quad Fscore[l] = Fscore[l] + MRelP[i][sid]/sid$
13 **for** $each S_l\ in\ dataset$ **do**
14 $\quad insertInto(Fscore[l], S_l, CombineList)$
15 $decreasing\_sort(CombineList)$
16 $return\ Top\_k(CombinedList)$

PF algorithm is explained as follows:

- (Lines 1-4), we initialize the fused scores with 0, the fused (combined) ranking is also initialized with an empty list.
- (Lines 5-10) for each matching function i and the current request Q_j, we compute the corresponding ranking termed $ranking_i$.
- (Line 11), we sort the aforementioned ranking.
- (Lines 12-14), for each service S_l, we get the identifier of the segment in which he lies (Sid).
- (Line 15), we update the fused score by applying the formula 22 (heuristics $H1\&H2$).
- (Lines 17-20), we create and sort the combined list, according to the decreasing order of the fused score.
- (Line 21): we return the Top_K elements of the combined list.

In what follows, we present the third algorithm referred to as **CVBT** (cross-validation based tuning). This algorithm

aims to select the optimal number of segments (denoted by ns) that ensures the best mean averaged precision of the combined rankings. We notice that $ns \in \{2, 3, ..., round(|dataset|/2)\}$.

To fulfill this goal, we use the cross-validation principle. This means that we firstly initialize ns to a given value, then we divide the requests collection on a set of parts (np parts). Thereafter, we perform the cross-validation section as follows:

- We compute the relevance probabilities (i.e the RPC function) by choosing $(np - 1)$ parts as the set of learning requests.
- We perform the probabilistic fusion (PF) over the entire set of requests.
- We calculate the mean averaged precision (MAP) that corresponds to the actual learning requests.
- We change the set of learning requests, by considering another union of $(np - 1)$ parts, and we redo the previous steps
- We take the average of the calculated MAP and we consider it as the final MAP associated to the actual ns (we denote this result as MAP'). We iterate the previous process (the five steps) for all possible values of ns and then we choose the value that ensures the best MAP'.

In the following, we describe the pseudo code of **CVBT** (see Algorithm 3)

- (Line 1), we divide the entire collection requests on a set of parts, for instance, if $np = 5$, then we have five subsets of requests.
- (Line 2), we initialize the optimal number of segments, as well as the optimal MAP.
- (Lines 3-6), for each candidate ns, we initialize its corresponding (averaged) MAP' with 0.
- (Lines 7-9), for each iteration of the cross-validation, we initialize the learning requests (SRQ). The latter is constituted of $(np - 1)$ parts of the entire collection. Thereafter we use this set (SRQ) to learn the relevance probabilities $(MRelP)$.
- (Lines 10-13), for each request Q_j we produce the fused ranking *CombinedListj*, afterwards, we calculate the corresponding average precision.
- (line 14), we estimate the mean averaged precision of the actual SRQ set.
- (line 16-19), We calculate the mean averaged precision of the cross-validation loop (denoted MAP'). This score is associated to the actual ns.
- (line 20), we return the optimal number of segments $(ns*)$.

In the following, we show a scenario that illustrates the processing performed by the probabilistic fusion (i.e. RPC and PF).

Algorithm 3: Algorithm CVBT

Input: $dataset: \ the \ set \ of \ services$,
$CRQ: \ the \ collection \ of \ requests$, $m:$
$Integer \ (the \ number \ of \ matching \ functions)$,
$np: \ Integer \ (the \ number \ of \ parts)$,
$the \ defaut \ value \ is \ 5$,
$Rel:$
$a \ binary \ matrix \ of \ dimension \|requests\|.|dataset|$,
$Q_j: \ the \ current \ request \ to \ be \ handled$, $m:$
$Integer \ (the \ number \ of \ matching \ functions)$

Output: $ns*:$
$Integer \ (the \ optimal \ number \ of \ segments)$

1 $Parts = divide(CRQ, np)$
2 $MAP* = 0; \ ns* = 2;$
3 **for** $ns = 2 \ to \ round(|dataset|/2)$ **do**
4 $SRQ = CRQ$
5 $MAP'[ns] = 0$
6 **for** $i = 1 \ to \ np$ **do**
7 $SRQ = SRQ - Parts[i]$
8 $MRelP = RPC(dataset, SRQ, m, ns, Rel)$
9 **for** $each \ Q_j \ in \ CRQ$ **do**
10 $CombinedList_j = FP(MRelP, |dataset|, Q_j)$
11 $AP[j] = AveragePrecision(CombinedList_j, Rel)$
12 $MAP[i] = MeanAveragedPrecision(AP)$
13 **for** $i = 1 \ to \ np$ **do**
14 $MAP'[ns] = MAP'[ns] + MAP[i]$
15 $MAP'[ns] = MAP'[ns]/np$
16 $return \ (ns*)$

V. Experimental Study

This section presents our experiments related to the probabilistic fusion as well as the individual matching functions. We also show a comparison with respect to the Borda [33] fusion scheme and other state of the art methods.

A. Evaluation Scheme

To assess the effectiveness and the efficiency of the proposed fusion scheme, we use the test collection *OWLTC V2.2*[1]. The latter contains real-world web service descriptions, extracted mainly from public IBM UDDI registries. As depicted in Table II, the benchmark contains:

1) 1007 service descriptions,
2) 29 sample requests,
3) a manually identified relevance set for each request. This information allows the computation of recall and precision.

Since we set np to 5 (np is the number of parts), then 80% of the request set is utilized for learning the relevance probabilities. In addition, all requests will be used for evaluating *MAP* and some other metrics *($recall@N, Prec@N, R-prec$) defined below:*

[1] http://projects.semwebcentral.org/projects/owls-tc/

TABLE II. OWLSTC V2 TEST COLLECTION

Class	Number of services	Number of requests
Travel	197	6
Education	286	6
Food	34	1
Medical care	73	1
Communication	59	2
Weapon	40	1
Economy	395	12

TABLE III. AVERAGE EXECUTION TIME OF THE PROBABILISTIC FUSION

Average fusion time (PF function)	Average learning time (RPC function)	Sum
13830 ms	13659 ms	27489 ms

TABLE IV. AVERAGE EXECUTION TIME FOR ALL METHODS

Approach	Probabilitic fusion	Borda	Cos	EJ	IL	JS	LOG
Average time	27489 ms	700 ms	28630 ms	26391 ms	21741 ms	27659 ms	14941

- *R-Precision(R-prec or R-P)*: measures the precision after all relevant items have been retrieved [40].
- Precision at N *(Prec@N)*: measures the precision after N items have been retrieved [40].
- Recall at N *(recall@N)*: measures the recall after N items have been retrieved [40].

We also measure the average execution time of the probabilistic fusion, the Borda fuse model and the individual matching functions. Our algorithms have been implemented in Java and the experiments were conducted on a Core I3 1.80 GHz machine with 4GB of RAM, running on Windows7.

In Table III, we show the average execution time of the learning phase (RPC function), the fusion phase (PF function), and the total time. Since the aforementioned algorithms have a polynomial complexity, then they remain scalable for large services datasets.

In Table IV, we compare the performance of the probabilistic fusion with respect to the other approaches. We observe that all running times fluctuate between 21.000 and 29.000 Milli.Sec, except for Borda. The latter exhibits a performance around 700 ms. This is due to the fact that Borda is a simple sum of the service positions. We also notice that, the logic-based approach is more efficient than the other individual methods, because we implemented the subsumption test with a logic "or". This implementation is enabled by a binary encoding scheme inspired from [41]. By coding the ontology with binary words we significantly decrease the subsumption test cost.

Tables V and VI show the behavior of PF for both recall and precision. In general, we observe that the performance rises as the number of segments ns increases (for all values of K). We also notice that the best performance is provided by $ns = 500$.

TABLE V. THE RECALL WITH RESPECT TO ns

ns	TOP 10	TOP 20	TOP 30	TOP 40	TOP 50	TOP 60
100	0.419	0.661	0.777	0.829	0.867	0.908
150	0.425	0.678	0.793	0.85	0.884	0.911
200	0.423	0.679	0.801	0.86	0.898	0.924
251	0.421	0.679	0.815	0.871	0.904	0.931
300	0.426	0.685	0.806	0.868	0.904	0.932
350	0.434	0.691	0.811	0.873	0.911	0.936
400	0.433	0.691	0.811	0.872	0.911	0.936
500	0.432	0.705	0.828	0.892	0.93	0.949

TABLE VI. THE PRECISION WITH RESPECT TO ns

ns	TOP 10	TOP 20	TOP 30	TOP 40	TOP 50	TOP 60
100	0.906	0.746	0.616	0.506	0.432	0.381
150	0.896	0.76	0.624	0.516	0.437	0.382
200	0.893	0.762	0.633	0.524	0.446	0.389
251	0.893	0.768	0.642	0.529	0.449	0.393
300	0.896	0.77	0.636	0.528	0.448	0.391
350	0.917	0.775	0.637	0.531	0.453	0.393
400	0.913	0.775	0.637	0.531	0.453	0.393
500	0.91	0.789	0.652	0.542	0.461	0.398

TABLE VII. MEAN RECALL FOR ALL METHODS

Algorithm	TOP 10	TOP 20	TOP 30	TOP 40	TOP 50	TOP 60
EJ	0.33	0.59	0.73	0.79	0.83	0.85
IL	0.33	0.59	0.7	0.79	0.84	0.86
JS	0.33	0.59	0.72	0.79	0.83	0.86
LOG	0.3	0.46	0.6	0.66	0.72	0.69
COS	0.33	0.59	0.72	0.78	0.82	0.86
PF	**0.43**	**0.70**	**0.82**	**0.89**	**0.93**	**0.94**
BORDA	0.39	0.58	0.73	0.81	0.84	0.9

TABLE VIII. MEAN PRECISION FOR ALL METHODS

Algorithm	TOP 10	TOP 20	TOP 30	TOP 40	TOP 50	TOP 60
EJ	0.81	0.64	0.58	0.51	0.42	0.36
IL	0.81	0.64	0.58	0.5	0.42	0.36
JS	0.8	0.65	0.57	0.49	0.41	0.36
LOG	0.73	0.53	0.48	0.42	0.37	0.31
COS	0.81	0.64	0.57	0.48	0.4	0.36
PF	**0.90**	**0.76**	**0.63**	**0.52**	**0.44**	**0.39**
BORDA	0.83	0.67	0.58	0.5	0.42	0.38

According to Tables VII and VIII, we observe that PF is more effective than the remaining approaches. The PF results are achieved by setting ns to 500. As demonstrated in the experiments, PF largely outperforms the Borda fuse model.

This is due to the fact that Borda is very sensitive to the services with bad individual ranks. Consequently its global performance is unsatisfying. On the other hand, we notice that the four similarity measures $\{Cos, EJ, IL, JS\}$ have almost the same performance. The worst case is achieved by the logic-based approach.

The execution of *CVBT* is shown in Fig. 1. It depicts the

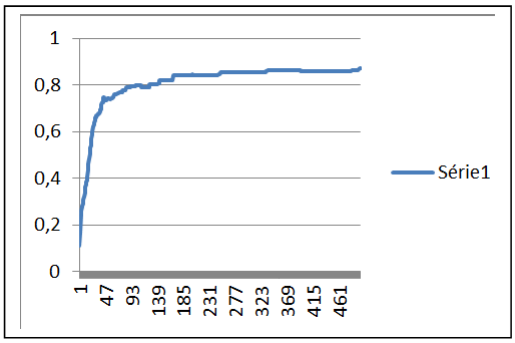


Fig. 1. Mean Average Precision (MAP) vs. ns

TABLE IX. AVERAGE R-PREC FOR ALL METHODS

Algorithm	R-Prec
PF	**0.86**
Borda	0.669
IL	0.709
LOG	0.594
JS	0.705
EJ	0.707
COS	0.692

TABLE X. PF VS S3 C CONTEST APPROACHES

Algorithm	MAP
PF (ns=500	**0.87**
Borda	0.73
OWLS-iMatcher2	0.84
JIAC-OWLSM	0.81
SPARQLent	0.71
OPOSSUM	0.57
ALIVE	0.5
OWLS MX3	0.86

relationship between MAP and ns parameter. In general, we distinguish two behaviours: firstly, when $ns \in \{2, .., 120\}$ we observe a rapid improvement of the estimated MAP. Second, when $ns \in 121, .., 500$, we observe a slow improvement of MAP. The optimal value is reached round 500.

From these results we conclude that the more the segment size is small, the better the performances.

As depicted in Table IX, the $R-Prec$ of PF is higher than the individual ranking algorithms as well as the Borda fuse model. In summary, PF produces a gain of 21% with respect to the highest individual $R - Prec$ (i.e. the information loss R-Prec) and 28% with respect to the Borda $R - Prec$.

Table X shows a comparison between the probabilitic fusion and the different systems that participate in the S3 contest 2009[2]. We notice that this competition is based on the same benchmark (i.e. OWLSTC.2). Table X clearly shows that our approach outperforms all existing matchmakers.

[2]International Contest S3 on Semantic Service Selection 2009, http://www-ags.dfki.uni-sb.de/ klusch/s3/

VI. CONCLUSION

In this paper, we have tackled the problem of retrieving and ranking web services. Our proposed framework takes into account multiple functional descriptors (input and output parameters) as well as several matching functions (logic reasoning and text similarities).

Simply speaking, our fusion algorithm leverages a set of relevance probabilities in order to infer an optimal fused ranking. These probabilities are largely dependent on the number of segments (ns). The setting of this regulating parameter is ensured by the cross-validation process.

The obtained results are very promising, and confirm the effectiveness of the proposed scheme.

In the nearest future, we aim to compare our approach with alternative fusion schemes, such as probabilistic dominance and majority-based voting. These approaches can be further enhanced by tuning their critical parameters with machine learning algorithms.

REFERENCES

[1] S. Sagayaraj and M. Santhoshkumar, "Heterogeneous ensemble learning method for personalized semantic web service recommendation."

[2] S. Kona, A. Bansal, G. Gupta, and T. D. Hite, "Semantics-based web service composition engine," in *Proc. of the 9th IEEE Int. Conf. on E-Commerce Technology (CEC 2007) / 4th IEEE Int. Conf. on Enterprise Computing, E-Commerce and E-Services (EEE 2007)*, 2007, pp. 521–524.

[3] M. Mohammed, C. M. Amine, and H. Fethallah, "Leveraging fuzzy dominance relationship and machine learning for hybrid web service discovery," *International Journal of Web Engineering and Technology*, vol. 11, no. 2, pp. 107–132, 2016.

[4] M. Klusch, B. Fries, and K. Sycara, "OWLS-MX: A hybrid Semantic Web service matchmaker for OWL-S services," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 2, pp. 121–133, 2009.

[5] M. Klusch and P. Kapahnke, "The iSeM matchmaker: A flexible approach for adaptive hybrid semantic service selection," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 15, pp. 1–14, 2012.

[6] K. Benouaret, D. Benslimane, A. Hadjali, M. Barhamgi, Z. Maamar, and Q. Z. Sheng, "Web service compositions with fuzzy preferences: A graded dominance relationship-based approach," *ACM Transactions on Internet Technology (TOIT)*, vol. 13, no. 4, p. 12, 2014.

[7] S. Berg, "Condorcet's jury theorem, dependency among jurors," *Social Choice and Welfare*, vol. 10, no. 1, pp. 87–95, 1993.

[8] M. Santhoshkumar and S. Sagayaraj, "Ranking semantic web services by matching triples and query based on similarity measure," *International Journal of Information Technology*, pp. 1–9, 2019.

[9] N. Srinivasan, M. Paolucci, and K. Sycara, "Semantic web service discovery in the OWL-S IDE," in *Proc. of the 39th Annual Hawaii Int. Conf. on System Sciences (HICSS'06)*, vol. 6. IEEE, 2006, pp. 109b–109b.

[10] U. Keller, R. Lara, H. Lausen, A. Polleres, and D. Fensel, "Automatic location of services," *The Semantic Web: Research and Applications*, pp. 1–16, 2005.

[11] U. Küster and B. König-Ries, "Evaluating semantic web service matchmaking effectiveness based on graded relevance," in *Proc. of the 2nd Int. Conf. on Service Matchmaking and Resource Retrieval in the Semantic Web-Volume 416*. CEUR-WS. org, 2008, pp. 32–46.

[12] F. Baader and U. Sattler, "An overview of tableau algorithms for description logics," *Studia Logica*, vol. 69, no. 1, pp. 5–40, 2001.

[13] A. Segev and E. Toch, "Context-based matching and ranking of web services for composition," *IEEE Transactions on Services Computing*, vol. 2, no. 3, pp. 210–222, 2009.

[14] G. Cassar, P. Barnaghi, and K. Moessner, "Probabilistic matchmaking methods for automated service discovery," *IEEE Transactions on Services Computing*, vol. 7, no. 4, pp. 654–666, 2014.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003. [Online]. Available: http://portal.acm.org/citation.cfm?id=944937

[16] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. of the 15th Conf. on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.

[17] M. Schumacher, H. Helin, and H. Schuldt, *CASCOM: intelligent service coordination in the semantic web*. Springer Science & Business Media, 2008.

[18] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[19] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8, 1966, pp. 707–710.

[20] E. Garcia, "Cosine similarity and term weight tutorial," *Information retrieval intelligence*, 2006.

[21] B. Fuglede and F. Topsoe, "Jensen-shannon divergence and hilbert space embedding," in *Proc. of the Int. Symposium on Information Theory (SIT 2004)*. IEEE, 2004, p. 31.

[22] Z. Chouiref, A. Belkhir, K. Benouaret, and A. Hadjali, "A fuzzy framework for efficient user-centric web service selection," *Applied Soft Computing*, vol. 41, pp. 51–65, 2016.

[23] K. Benouaret, D. Benslimane, and A. Hadjali, "Selecting skyline web services for multiple users preferences," in *Proc. of the 19th IEEE Int. Conf. on Web Services (ICWS'12)*. IEEE, 2012, pp. 635–636.

[24] A. Pahlevan, J.-L. Duprat, A. Thomo, and H. Müller, "Dynamis: effective context-aware web service selection using dynamic attributes," *Future Internet*, vol. 7, no. 2, pp. 110–139, 2015.

[25] P. Rodriguez-Mier, C. Pedrinaci, M. Lama, and M. Mucientes, "An integrated semantic web service discovery and composition framework," *IEEE transactions on services computing*, vol. 9, no. 4, pp. 537–550, 2016.

[26] M. Fariss, N. El Allali, H. Asaidi, and M. Bellouki, "Review of ontology based approaches for web service discovery," in *International Conference on Advanced Information Technology, Services and Systems*. Springer, 2018, pp. 78–87.

[27] N. Zhang, J. Wang, Y. Ma, K. He, Z. Li, and X. F. Liu, "Web service discovery based on goal-oriented query expansion," *Journal of Systems and Software*, vol. 142, pp. 73–91, 2018.

[28] M. Klusch, B. Fries, and K. Sycara, "Automated semantic web service discovery with OWLS-MX," in *Proc. of the 5th Int. Joint Conf. on autonomous agents and multiagent systems*. ACM, 2006, pp. 915–922.

[29] M. Klusch and P. Kapahnke, "Semantic web service selection with SAWSDL-MX," in *Proc. of the 7th Int. Semantic Web Conf.*, 2008, p. 3.

[30] A. Halfaoui, F. Hadjila, and F. Didi, "Qos-aware web service selection based on self-organising migrating algorithm and fuzzy dominance," *International Journal of Computational Science and Engineering*, vol. 17, no. 4, pp. 377–389, 2018.

[31] D. Skoutas, D. Sacharidis, A. Simitsis, and T. Sellis, "Ranking and clustering web services using multicriteria dominance relationships," *IEEE Transactions on Services Computing*, no. 3, pp. 163–177, 2010.

[32] E. A. Fox and J. A. Shaw, "Combination of multiple searches," *NIST special publication SP*, vol. 243, 1994.

[33] J. A. Aslam and M. Montague, "Models for metasearch," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '01. New York, NY, USA: Association for Computing Machinery, 2001, p. 276–284. [Online]. Available: https://doi.org/10.1145/383952.384007

[34] M. Farah and D. Vanderpooten, "An outranking approach for rank aggregation in information retrieval," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 591–598. [Online]. Available: https://doi.org/10.1145/1277741.1277843

[35] H. Fethallah, B. Amine, and H. Amel, "Hybrid Web Service Discovery Based on Fuzzy Condorcet Aggregation," in *East Europ. Conf. on Advances in Databases and Information Systems*. Springer, 2015, pp. 415–427.

[36] M. Montague and J. A. Aslam, "Condorcet fusion for improved retrieval," in *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, ser. CIKM '02. New York, NY, USA: Association for Computing Machinery, 2002, p. 538–548. [Online]. Available: https://doi.org/10.1145/584792.584881

[37] S. Samir, A. Sarhan, and A. Algergawy, "Context-based web service discovery framework with qos considerations," in *Proc. of the 11th Int. Conf. on Research Challenges in Information Science (RCIS 2017)*. IEEE, 2017, pp. 146–155.

[38] A. D. Eddine and B. F. M'hamed, "Improved multicriteria ranking based web service discovery approach," in *2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)*. IEEE, 2018, pp. 1–6.

[39] D. Lillis, F. Toolan, R. Collier, and J. Dunnion, "Probfuse: a probabilistic approach to data fusion," in *Proc. of the 29th annual int. ACM SIGIR conf. on research and development in information retrieval*. ACM, 2006, pp. 139–146.

[40] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. USA: Cambridge University Press, 2008.

[41] Y. Caseau, M. Habib, L. Nourine, and O. Raynaud, "Encoding of multiple inheritance hierarchies and partial orders," *Computational Intelligence*, vol. 15, no. 1, pp. 50–62, 1999.