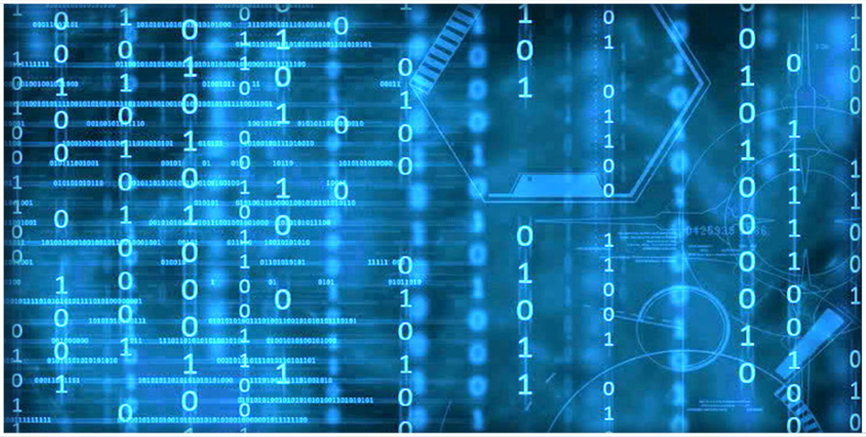
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 11 Issue 6

June 2020



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

SAI

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 11 Issue 6 June 2020
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

Modeling Real-World Load Patterns for Benchmarking in Clouds and Clusters

Kashifuddin Qazi
Department of Computer Science
Manhattan College
NY, USA

***Abstract*—Cloud computing has currently permeated all walks of life. It has proven extremely useful for organizations and individual users to save costs by leasing compute resources that they need. This has led to an exponential growth in cloud computing based research and development. A substantial number of frameworks, approaches and techniques are being proposed to enhance various aspects of clouds, and add new features. One of the constant concerns in this scenario is creating a testbed that successfully reflects a real-world cloud datacenter. It is vital to simulate realistic, repeatable, standardized CPU and memory workloads to compare and evaluate the impact of the different approaches in a cloud environment. This paper introduces Cloudy, which is an open-source workload generator that can be used within cloud instances, Virtual Machines (VM), containers, or local hosts. Cloudy utilizes resource usage traces of machines from Google and Alibaba clusters to simulate up to 16000 different, real-world CPU and memory load patterns. The tool also provides a variety of machine metrics for each run, that can be used to evaluate and compare the performance of the VM, container or host. Additionally, it includes a web-based visualization component that offers a number of real-time statistics, as well as overall statistics of the workload such as seasonal trends, and autocorrelation. These statistics can be used to further analyze the real-world traces, and enhance the understanding of workloads in the cloud.**

Keywords*—*Cloud computing; workload generator; cluster computing

I. Introduction

Cloud computing has become an important part of most organizations these days. By leasing compute resources from cloud providers, organizations can save on hardware and setup costs. Because of its popularity cloud computing has garnered substantial attention within the research community. Research is being consistently performed on a large scale on the cloud to make it faster, more efficient, and add more features.

Researchers have explored different aspects of cloud computing such as live migration [1] [2], vertical elasticity [3], horizontal elasticity, remote memory [4], workload prediction, container placement, virtual machine consolidation, and load balancing [5]. In order to evaluate proposed approaches similar to these, it is important to have a standard tool that can be used to benchmark them. This entails an environment that simulates the resource usage patterns seen in the real world. For example, it is important to test live migration approaches with virtual machines using up realistic amounts of memory over time. The environment should also offer a number of features to be useful as an evaluation benchmark in cloud computing based research. First, the workload generation should be non-intrusive, i.e. it should run separate from the approach being tested. Second, the tool should preferably also log a variety of performance and system statistics. These statistics are extremely important to observe the positive or negative effects of the approach under test. Third, the testbed should allow setup within a Virtual Machine (VM), container, cloud instance, or physical host. Fourth, the testbed required could involve a cluster of machines, and the tool should be able to simulate workloads on multiple computers. Finally, the tool should ideally be open source.

This paper introduces an open-source tool called Cloudy that models and runs workloads within cloud instances, VMs, containers, or physical hosts. It is easy to use, and can be downloaded, installed and run without the need for additional configurations in the system and without affecting any other components on the system. The tool uses data traces from more than 16000 machines from Google and Alibaba clusters to provide real-world patterns of memory and CPU usage in real time over multiple days. This ensures a large number of unique workloads that can be run on different machines in a cluster. Additionally, Cloudy features an online visualization dashboard that can be used to observe the CPU and memory usage of a machine, as well as obtain other important performance statistics such as operations per second, number of page faults, etc. over time. Finally, the workload generated can be scaled in terms of both usage and time, giving a finer level of control to the user. It is envisioned that this tool could benefit experimental evaluations of cloud-based research, and provide an easy-to-use standard to compare different approaches. Earlier versions of this tool have been previously used by the authors in [4], [6].

Fig. 1 shows one possible use-case for Cloudy. In order to evaluate a cloud-based framework or approach, a baseline set of performance statistics are obtained by running Cloudy within Infrastructure-as-a-Service (IaaS) instances. Since Cloudy offers more than 16000 unique trace patterns, each instance in this set can run a different real-world workload pattern. Next, Cloudy is restarted with the same workloads as before, this time, along with the approach to be evaluated. The performance statistics are collected again. A comparison of these statistics against the baseline statistics can help researchers evaluate the efficacy of the approach being tested.

The rest of the paper is divided as follows. Section II discusses some existing cloud benchmarking tools and highlights the difference between those tools and Cloudy. Section III describes the implementation, and internals of Cloudy in detail. Section IV reports various experimental results to

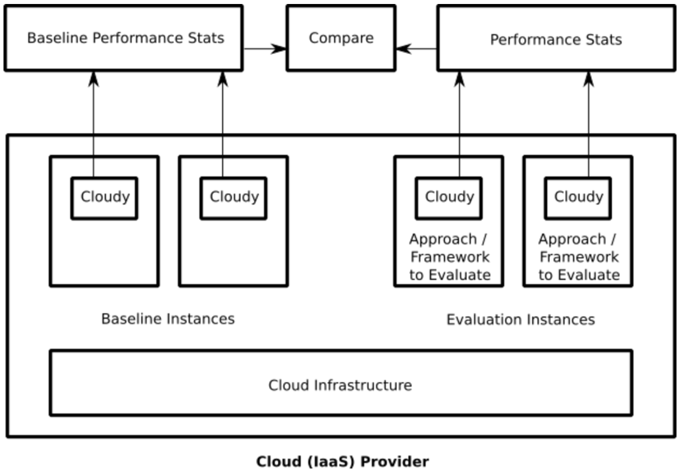
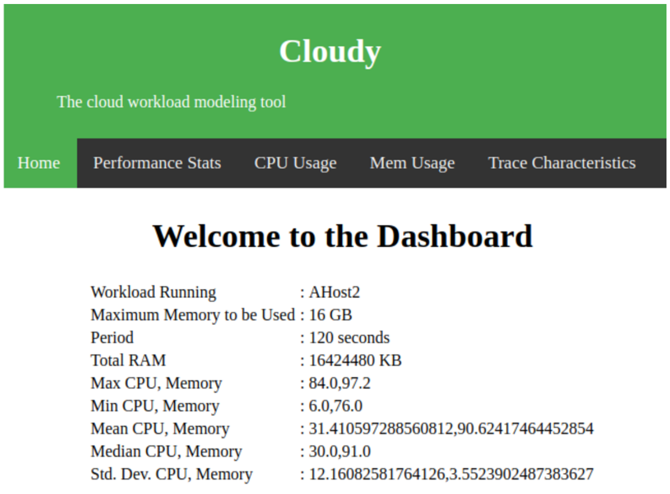


Fig. 1. Use Case of Cloudy.



Welcome to the Dashboard

Workload Running	: AHost2
Maximum Memory to be Used	: 16 GB
Period	: 120 seconds
Total RAM	: 16424480 KB
Max CPU, Memory	: 84.0,97.2
Min CPU, Memory	: 6.0,76.0
Mean CPU, Memory	: 31.410597288560812,90.62417464452854
Median CPU, Memory	: 30.0,91.0
Std. Dev. CPU, Memory	: 12.16082581764126,3.5523902487383627

Fig. 2. Screenshots of Cloudy Web-based Visualization Dashboard.

demonstrate the accuracy and efficacy of Cloudy in recreating the real-world patterns. It also illustrates its utility in further analyzing the traces. Section V notes additional discussions and considerations when using Cloudy. Section VI concludes the paper.

II. Related Work

A number of research endeavors and software exist in literature for evaluating and benchmarking in cloud-based environments. These tools generally fall into one of three categories - benchmarking the cloud itself, testing performance of a required application in different clouds, and general benchmarking tools that solve resource-intensive problems to use resources.

Cloud Bench [7] automates cloud-scale evaluation and benchmarking through the running of controlled experiments, where complex applications are automatically deployed according to user-defined experiment plans. It helps assess the stability, scalability and reliability of different cloud configurations. Similarly, Expertus [8], [9] is a code generation-based approach with the main goal of automating distributed application configuration and testing in IaaS clouds. Cloud Crawler [10], [11] approaches the same problem by providing users with an environment where they can describe a variety of performance evaluation scenarios for a given application. The tool then automatically configures, executes and collects the results of the scenarios described. Cloud WorkBench [12] is another cloud benchmarking service that supports the automatic execution of systematic performance tests in the cloud by leveraging the notion of Infrastructure-as-Code (IaC).

These approaches have a different goal compared to Cloudy. They do not non-obtrusively run a real-world workload in the background. Instead, the workload that will run is the application that a developer intends to move to the cloud. The approaches test and evaluate the given application under different cloud scenarios and offer advice on suitable placement strategies. They are useful for selecting an appropriate configuration of cloud resources for a given application.

RUBiS [13] is a free, open-source auction site prototype modeled after eBay.com. It can be used to evaluate application design patterns and application servers' performance scalability. The website can simulate a real-world load by performing actions such as selling, browsing and bidding. While RUBiS does simulate a real-world application, it is restricted to a scenario consisting of a webserver, specifically for an auction-like site.

Another actively maintained open-source tool that comes with a collection of pre-configured benchmarks is Google's PerfKit Benchmarker [14]. It also offers an optional dashboard for performance analysis. The main goal is to define a canonical set of benchmarks to measure and compare cloud offerings. However, PerfKit does not offer any features that allow generating loads according to real-world patterns.

As opposed to all these approaches, Cloudy focuses on generating CPU and memory load patterns that mirror real-world loads.

III. Method

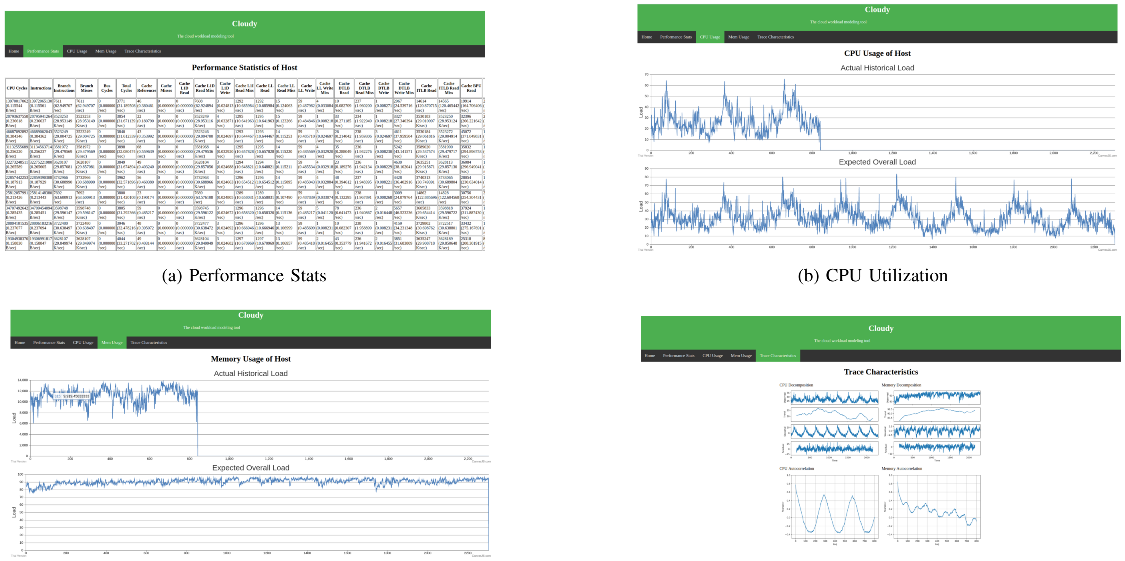
The methodology of Cloudy is discussed in the following subsections from two perspectives: the end-user's perspective (installation, execution, interaction) and implementation (internal components).

A. User's Perspective

From the end-user's perspective, setting up and interacting with Cloudy is a straightforward process. The entire framework with all the required dependencies can be cloned from the Gitlab repository [15] into the VM, local host, or cloud instance of choice. Cloudy can then be installed by running the provided install script (install.sh). Once all the dependencies and file placements are automatically handled, the workload can be started by running the workload.sh script, passing the name of the trace to use (TN), maximum memory to use in GB (MMG) and time scaling in seconds (TSS) as arguments.

./workload.sh *TN* [-m*MMG*] [-t*TSS*]

The argument trace name is the name of the underlying real-world trace that Cloudy will use to generate CPU and



Performance Statistics of Host

CPU Cycles	Instructions	Branch Instructions	Branch Misses	Bus Cycles	Total Cycles	Cache References	Cache Misses	Cache L1D Read	Cache L1D Read Miss	Cache L1D Write	Cache L1I Read Miss	Cache LL Read	Cache LL Read Miss	Cache LL Write	Cache LL Write Miss	Cache DTLB Read	Cache DTLB Read Miss	Cache DTLB Write	Cache DTLB Write Miss	Cache ITLB Read	Cache ITLB Read Miss	Cache BPU Read
13970017062 (0.115544 B/sec)	13972065130 (0.115561 B/sec)	7611 (62.949707 /sec)	7611 (62.949707 /sec)	0 (0.000000 /sec)	3771 (31.189508 /sec)	46 (0.380461 /sec)	0 (0.000000 /sec)	0 (0.000000 /sec)	7608 (62.924894 /sec)	3 (0.024813 /sec)	1292 (10.685984 /sec)	1292 (10.685984 /sec)	15 (0.124063 /sec)	59 (0.487982 /sec)	4 (0.033084 /sec)	10 (0.082709 /sec)	237 (1.960200 /sec)	1 (0.008271 /sec)	2967 (24.539716 /sec)	14614 (120.870715 /sec)	14565 (120.465442 /sec)	19914 (164.706406 /sec)
28793637558 (0.236618 B/sec)	28795941264 (0.236637 B/sec)	3523253 (28.953149 K/sec)	3523253 (28.953149 K/sec)	0 (0.000000 /sec)	3854 (31.671139 /sec)	22 (0.180790 /sec)	0 (0.000000 /sec)	0 (0.000000 /sec)	3523249 (28.953116 K/sec)	4 (0.032871 /sec)	1295 (10.641963 /sec)	1295 (10.641963 /sec)	15 (0.123266 /sec)	59 (0.484846 /sec)	1 (0.008218 /sec)	33 (0.271185 /sec)	234 (1.922949 /sec)	1 (0.008218 /sec)	3327 (27.340394 /sec)	3530183 (29.010097 K/sec)	3523250 (28.953124 K/sec)	32396 (266.221642 /sec)
46687092892 (0.384346 B/sec)	46689062043 (0.384362 B/sec)	3523249 (29.004725 K/sec)	3523249 (29.004725 K/sec)	0 (0.000000 /sec)	3840 (31.612339 /sec)	43 (0.353992 /sec)	0 (0.000000 /sec)	0 (0.000000 /sec)	3523246 (29.004700 K/sec)	3 (0.024697 /sec)	1293 (10.644467 /sec)	1293 (10.644467 /sec)	14 (0.115253 /sec)	59 (0.485710 /sec)	3 (0.024697 /sec)	26 (0.214042 /sec)	238 (1.959306 /sec)	3 (0.024697 /sec)	4611 (37.959504 /sec)	3530184 (29.061816 K/sec)	3523272 (29.004914 K/sec)	45072 (371.049831 /sec)
31132555689 (0.256220 B/sec)	31134563714 (0.256237 B/sec)	3581972 (29.479569 K/sec)	3581972 (29.479569 K/sec)	0 (0.000000 /sec)	3898 (32.080474 /sec)	68 (0.559639 /sec)	0 (0.000000 /sec)	0 (0.000000 /sec)	3581968 (29.479536 K/sec)	4 (0.032920 /sec)	1295 (10.657828 /sec)	1295 (10.657828 /sec)	14 (0.115220 /sec)	59 (0.485569 /sec)	4 (0.032920 /sec)	35 (0.288049 /sec)	236 (1.942276 /sec)	1 (0.008230 /sec)	5242 (43.141571 /sec)	3589020 (29.537574 K/sec)	3581990 (29.479717 K/sec)	35832 (294.896753 /sec)
32273248511 (0.265589 B/sec)	32275221980 (0.265605 B/sec)	3628107 (29.857081 K/sec)	3628107 (29.857081 K/sec)	0 (0.000000 /sec)	3849 (31.674894 /sec)	49 (0.403240 /sec)	0 (0.000000 /sec)	0 (0.000000 /sec)	3628104 (29.857056 K/sec)	3 (0.024688 /sec)	1294 (10.648821 /sec)	1294 (10.648821 /sec)	14 (0.115211 /sec)	59 (0.485534 /sec)	4 (0.032918 /sec)	23 (0.189276 /sec)	236 (1.942134 /sec)	1 (0.008229 /sec)	4630 (38.102041 /sec)	3635251 (29.915871 K/sec)	3628113 (29.857130 K/sec)	36084 (296.949041 /sec)
22857442253 (0.187913 B/sec)	22859390308 (0.187929 B/sec)	3732966 (30.688990 K/sec)	3732966 (30.688990 K/sec)	0 (0.000000 /sec)	3962 (32.571896 /sec)	56 (0.460380 /sec)	0 (0.000000 /sec)	0 (0.000000 /sec)	3732963 (30.688966 K/sec)	3 (0.024663 /sec)	1296 (10.654512 /sec)	1296 (10.654512 /sec)	14 (0.115095 /sec)	59 (0.485043 /sec)	4 (0.032884 /sec)	48 (0.394612 /sec)	237 (1.948395 /sec)	1 (0.008221 /sec)	4428 (36.402916 /sec)	3740313 (30.749391 K/sec)	3733065 (30.689804 K/sec)	28054 (230.634014 /sec)
25812057991 (0.213426 B/sec)	25814148380 (0.213443 B/sec)	7692 (63.600913 /sec)	7692 (63.600913 /sec)	0 (0.000000 /sec)	3800 (31.420108 /sec)	23 (0.190174 /sec)	0 (0.000000 /sec)	0 (0.000000 /sec)	7689 (63.576108 /sec)	3 (0.024805 /sec)	1289 (10.658031 /sec)	1289 (10.658031 /sec)	13 (0.107490 /sec)	59 (0.487839 /sec)	4 (0.033074 /sec)	16 (0.132295 /sec)	238 (1.967891 /sec)	1 (0.008268 /sec)	3009 (24.879764 /sec)	14862 (122.885696 /sec)	14828 (122.604568 /sec)	30756 (254.304431 /sec)
34707492642 (0.285435 B/sec)	34709454094 (0.285451 B/sec)	3598748 (29.596147 K/sec)	3598748 (29.596147 K/sec)	0 (0.000000 /sec)	3805 (31.292366 /sec)	59 (0.485217 /sec)	0 (0.000000 /sec)	0 (0.000000 /sec)	3598745 (29.596122 K/sec)	3 (0.024672 /sec)	1296 (10.658320 /sec)	1296 (10.658320 /sec)	14 (0.115136 /sec)	59 (0.485217 /sec)	5 (0.041120 /sec)	78 (0.641473 /sec)	236 (1.940867 /sec)	2 (0.016448 /sec)	5657 (46.523236 /sec)	3605833 (29.654414 K/sec)	3598818 (29.596722 K/sec)	37924 (311.887430 /sec)
28804101535 (0.237077 B/sec)	28806183216 (0.237094 B/sec)	3722480 (30.638497 K/sec)	3722480 (30.638497 K/sec)	0 (0.000000 /sec)	3946 (32.478216 /sec)	48 (0.395072 /sec)	0 (0.000000 /sec)	0 (0.000000 /sec)	3722477 (30.638472 K/sec)	3 (0.024692 /sec)	1296 (10.666946 /sec)	1296 (10.666946 /sec)	13 (0.106999 /sec)	59 (0.485609 /sec)	1 (0.008231 /sec)	10 (0.082307 /sec)	238 (1.958899 /sec)	1 (0.008231 /sec)	4159 (34.231348 /sec)	3729802 (30.698762 K/sec)	3722517 (30.638801 K/sec)	33432 (275.167691 /sec)
19304938370 (0.158830 B/sec)	19306991817 (0.158847 B/sec)	3628107 (29.849974 K/sec)	3628107 (29.849974 K/sec)	0 (0.000000 /sec)	4044 (33.271702 /sec)	49 (0.403144 /sec)	0 (0.000000 /sec)	0 (0.000000 /sec)	3628104 (29.849949 K/sec)	3 (0.024682 /sec)	1297 (10.670969 /sec)	1297 (10.670969 /sec)	13 (0.106957 /sec)	59 (0.485418 /sec)	2 (0.016455 /sec)	43 (0.353779 /sec)	236 (1.941672 /sec)	2 (0.016455 /sec)	3851 (31.683809 /sec)	3635247 (29.908718 K/sec)	3628189 (29.850648 K/sec)	25318 (208.301915 /sec)

(a) Performance Stats

(b) CPU Utilization

(c) Memory Utilization

(d) Trace Characteristics

Fig. 3. Screenshots of Cloudy Web-based Visualization Component.

memory usage. It could be the name any one of the 16000 traces available in the repository. This argument is mandatory. The maximum memory to use is an important parameter that can be tweaked based on testing requirements. By default, when generating a pattern, Cloudy will use up to all the available memory. However, specifying a maximum will restrict Cloudy from zero to the maximum memory specified. It is important to note, that in either case, the pattern generated will look exactly the same and follow the underlying trace chosen, it will simply be scaled to the maximum memory specified. Similarly, the time scaling pattern allows the user to specify the duration of the entire workload. By default, each data point in the underlying trace is considered to be at 5 minutes (which is the actual time frame). However, specifying a different time scale (for example 120 seconds) will make Cloudy consider each underlying data point at the new time scale (every 120 seconds in the example). Again, the overall pattern of the trace will not be affected, instead this will simply stretch or shrink the entire trace in time.

Once suitable arguments are chosen (or left to default), Cloudy starts utilizing CPU and memory over time, according to the trace chosen.

The installer also sets up a webserver and front-end dashboard on the same machine, which can be accessed by browsing to the ip address of the machine, as long as port 80 is accessible. Fig. 2 is a screenshot of Cloudy's dashboard. This is the landing page of the ip address of a machine running Cloudy. The dashboard reports summarized statistics on the current state of the machine, and the underlying trace being used. This information includes the amount of memory available, as well as the name of the trace being run, and the maximum, minimum, median, mean, and standard deviation of both, memory and CPU usage of the trace.

The web-based interface also provides other detailed analysis of the system and trace being used. Fig. 3 shows screenshots of the remaining four sections of the visualization. Fig. 3a shows a report of the performance metrics collected through the entire run of the workload. These include the CPU cycles, page fault count, context switches, cache-related statistics, etc. and are also recorded in a logfile. Fig. 3b, 3c show the real time graphs for CPU and memory usage respectively. These sections also report the graphs for the entire trace for both CPU and memory usage. Finally, Fig. 3d calculates and generates statistics to evaluate the overall trace. These statistics include decomposition of the trace into trend, seasonal, and residual components, as well as autocorrelation plots. The statistics are generated for both the CPU and memory traces. These features are discussed in more detail in the section on Results.

Currently, the install script provided supports Ubuntu-based AWS ec2 instances. However, Cloudy can still be run without any modification on most Linux-based machines within different commercial cloud providers (such as Google's Compute Engine).

B. Implementation

Fig. 4 shows an overview of Cloudy. There are three main components of Cloudy, which include the workload trace, the load generator, and the web-based visualization component. As depicted in the figure, the workload traces are individual files, with the percent memory and CPU usage of 16000 machines, stored on a remote file hosting server. One of these traces is selected for each run of the model. When Cloudy is run, the initialization step downloads the trace file specified by the trace name (TN) argument onto the VM or container being tested. This trace file is picked up by the load generator, which follows the trace to generate matching memory and CPU loads over time. Finally, the visualization component, which also exists in the VM or container, can be accessed by any browser over the internet through the ip address of the VM or container to view details and statistics about the workload. The next subsections discuss the details of each of these components.

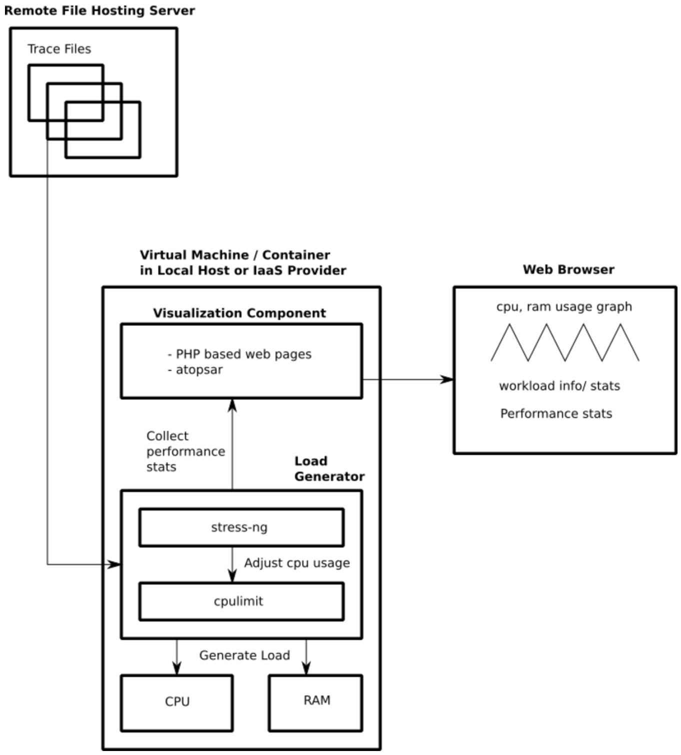


Fig. 4. Overview of Cloudy.

1) Data Traces: The data traces, which are stored on a remote server, hold the CPU and memory usage over time of one machine each. The files are structured so that each line has comma separated CPU and memory usage (in percentage of total) within a 5 minute period. There are a total of 16000 traces, 12000 of which belong to the Google cluster [16], and 4000 belong to the AliBaba cluster [17]. Cloudy uses these trace files as a guide to generating workloads. Next, the two cluster traces, and the mechanism used for extracting the relevant traces from the two datasets are described.

Google Data Traces: The Google cluster data trace consists of 29 days' worth of logs for about 12000 machines, from a Google cluster in a datacenter in the US, starting at 19:00 EDT on Sunday, May 1, 2011. In this context, a Google cluster is a set of machines, packed into racks, and connected by a high-bandwidth cluster network. A set of these machines (cell) is allocated work by a cluster-management system. Work arrives at a cell in the form of jobs which are comprised of one or more tasks, and these tasks run on machines. Each task is a Linux program made up of multiple processes and runs on a single machine. The usage data for the tasks were collected from the management system and the individual machines. The data is represented as percent CPU and memory usage of each task at 5 minute intervals.

The trace contains a number of tables describing different information. These tables include:

- Machine events - describes addition, removal, updates of machines
- Machine attributes - describes machine properties such as kernel version, clock speed, etc.
- Job events - describes when each job was submitted, scheduled, run, etc.
- Task events - describes which machines tasks are located in, resources requested, etc.
- Task constraints - describes constraints on placement of tasks, if any
- Task resource usage - describes mean CPU usage, memory usage, disk I/O time, etc. for each task at each time instance

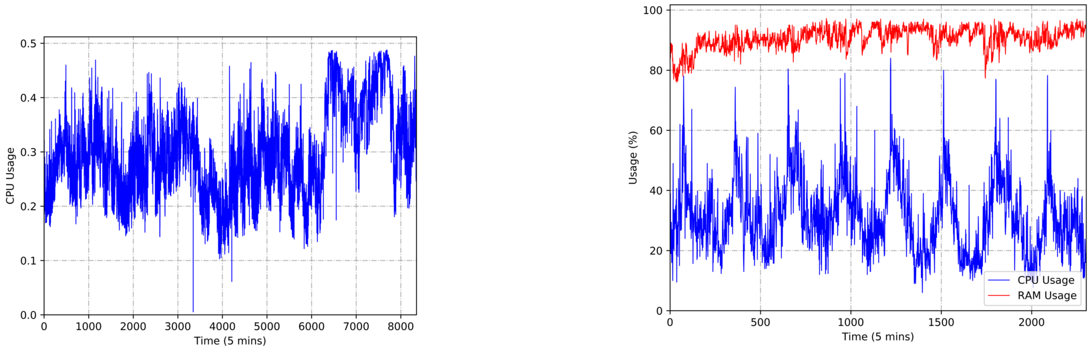
Of these tables, the task resource usage is of particular interest. Since the Google cluster data does not directly provide the CPU and memory usage on a particular machine, it has to be calculated. For a given instance in time, this is done by adding up the usages of all the tasks residing on the machine at that time. A python script was written to collate all the tasks on the same machine, and then calculate the sum of their usages at each time interval (5 minutes). The final traces are stored in separate files for each machine. The files are named GHost0 to GHost11999. The end result is a set of 12000 files with 8352 data points each (29 days at 5 minute intervals) specifying the percent of CPU and memory usage. Fig. 5a shows the CPU usage of a sample host from the Google cluster dataset for the 29 days of the trace.

Alibaba Cluster Data Traces: The Alibaba cluster data trace includes about 4000 machines for the Alibaba website, during a period of 8 days, and consists of six tables (each is a file). These tables include:

- machine_meta.csv: the meta info and event information of machines
- machine_usage.csv: the resource usage of each machine
- container_meta.csv: the meta info and event information of containers
- container_usage.csv: the resource usage of each container
- batch_instance.csv: information about instances in the batch workloads
- batch_task.csv: information about instances in the batch workloads

As opposed to the Google cluster data, the Alibaba data traces directly specify the percent CPU and memory usage of each machine at a given time. This can be obtained from the machine_usage.csv file. Using a python script, the usages of the machines were separated, and arranged in 5 minute intervals. The final traces are stored as separate files for each machine. The files are named AHost0 to AHost3999. The end result is a set of 4000 files with 2304 data points each (8 days at 5 minute intervals) specifying the percent of CPU and memory usage. Fig. 5b shows both the CPU and memory usage of a sample host from the Alibaba traces for the 8 days of the trace.

2) Load Generation: The load generator runs in a loop, reading a pair of CPU and memory values from the given trace file periodically. The period is dictated by the timescaling



(a) Google Host

(b) Alibaba Host

Fig. 5. Sample Hosts CPU and Memory Usage.

(TSS) argument given when running Cloudy. As mentioned, by default, the load generator reads a new pair of values every 5 minutes. Each period, the generator aims to generate a physical workload that matches both the CPU and memory usage specified by the pair of values. In order to generate this workload, a utility must be chosen that solves a generic problem, thus utilizing CPU and memory. For example, allocating and modifying large arrays can be used to simulate memory usage, while linear algebra solvers can simulate CPU usage. While it is fairly trivial to run a utility that simulates a certain amount of memory usage or CPU usage, it is extremely difficult to choose a single tool that utilizes an exact, arbitrary amount of both memory and CPU usages as required.

Cloudy approaches load generation in two steps. At any point in time, first the memory load required for the current period is generated by running a suitable utility. However, any such utility, will end up working at full available CPU capacity. Therefore, in the next step, a limit on the amount of CPU that can be used is applied to the running utility to match the CPU usage required for the current period.

To achieve the first step of memory load generation, the benchmarking utility stress-ng is used. This utility allows stress-testing a system in a number of selectable ways. Stress-ng has a variety of stressors including floating point, integer, or bit manipulation for CPU, i/o devices, network, schedulers, etc. Cloudy utilizes stress-ng's memory stressor to generate controlled, memory intensive loads. The memory stressor can be given a size of memory to use, and the stressor continuously calls mmap for the specified size and writes to the allocated memory. Since the trace files provide memory usage as a percentage, the load generator calculates the size of the memory to use based on the maximum memory (MMG) argument (if given) or the total memory available (default).

Once the required amount of memory is being used, the second step begins. The program cpulimit can be given a CPU usage percentage, and the PID of a process to limit the real CPU usage of the process to the desired percentage. Using this, the load generator limits the CPU usage of the running stress-ng process to the usage required for the current period.

At this point, both the CPU and memory usage of the machine match the values specified by the trace for the current period. These usages continue until the next period, when the current stress-ng process terminates, and the previous two steps are repeated for the next pair of values from the trace file.

3) Visualization: The Visualization component of Cloudy consists of some backend scripts for data collection and calculation and a frontend. The statistics that are recorded for visualization, are all returned from the stress-ng utility, and are collected at the end of each period. These include the operations per second, page fault count, etc. and are recorded in a logfile while stress-ng runs.

Additionally, to view the actual CPU and memory usage of the VM or container in real time, the program atopsar is used. Atopsar can report statistics on a system level and return periodic information about the usage.

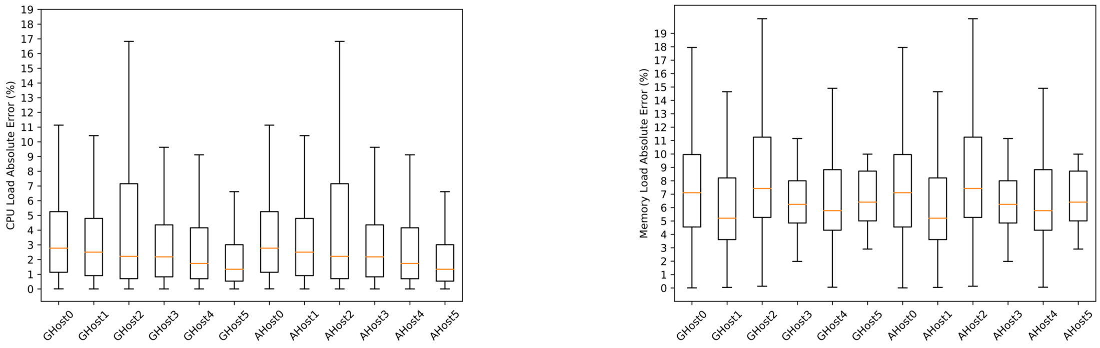
In order to retrieve the information in a suitable fashion, the logging features of both stress-ng and atopsar have been modified. The modifications only include changes to the output formats so that the outputs can be redirected to the logfiles, without the need for additional scripts to clean the data.

Finally, a backend Python script is used to calculate and plot the decompositions and autocorrelation values for both CPU and memory from the current trace file.

The front end of the visualization component is built using PHP. When installing Cloudy from the git repository, the entire Visualization component is included, and the front-end as well as the modified versions of stress-ng and atopsar are automatically installed.

IV. Experimental Evaluations and Results

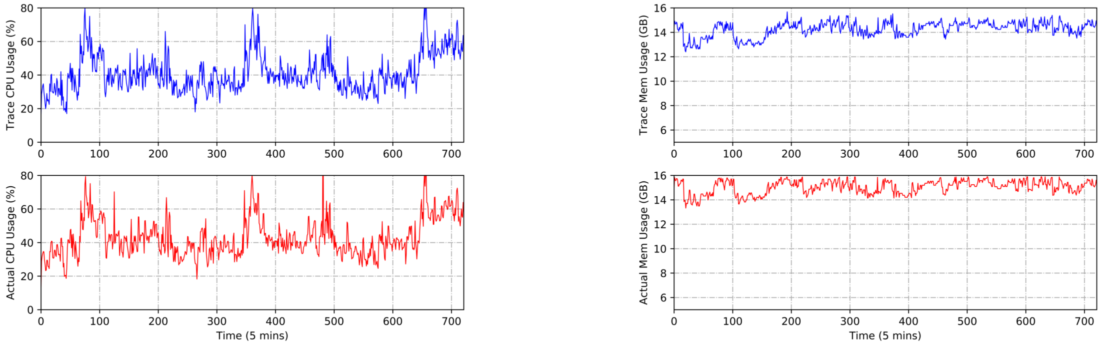
In order to evaluate Cloudy, multiple runs with different traces were performed on Amazon Web Services' ec2 instances (t2.xlarge: Ubuntu 18.04, 4 cores, 16 GB RAM, 40 GB EBS). For the experiments in this paper, the maximum memory to use was set to 16 GB, and the scaling was at the default of 5 minutes. Currently, 2000 traces are available in the gitlab repository. These include 1000 traces each from Google and Alibaba workloads (GHost0 to GHost999 and AHost0 to AHost999). The experiments that follow use samples from these 2000 traces. All the 16000 traces are currently being placed on a suitable ftp server, and are available on request.



(a) CPU

(b) Memory

Fig. 6. Absolute Actual vs. Trace Load Error.



(a) CPU

(b) Memory

Fig. 7. Actual vs. Trace Usage.

The following subsections evaluate two aspects: the accuracy of Cloudy when recreating patterns from underlying traces, and characteristics of the traces that can be gleaned using Cloudy.

A. Cloudy Evaluation

One of the important aspects of evaluating the efficacy of Cloudy is to analyze how closely the generated CPU and memory usages follow the usages in the underlying data traces. For these experiments, 12 traces (AHost0-5 and GHost0-5) were separately run for their entire duration, and evaluated on the ec2 instances. This implies that Cloudy was run for 29 days for each of the GHost traces, and 8 days for each of the AHost traces. The logged actual CPU and memory usage over these 12 runs was then compared to the usages according to the underlying traces. The absolute error at each period for each host was calculated as $abs(usage_{actual} - usage_{trace})$ Fig. 6 plots boxplots of the absolute errors for each of the 12 hosts.

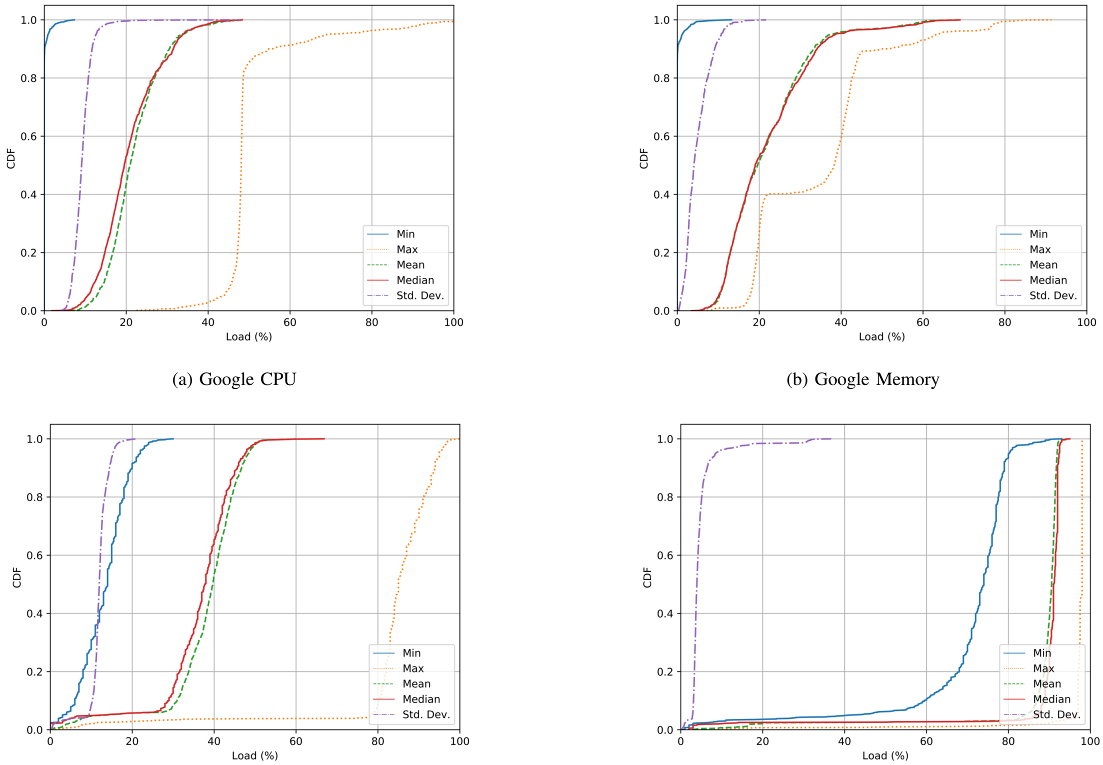
The plots show that for the 12 hosts, the median CPU error is mostly at about 2-3%. At worst, the generated CPU usage deviates by about 17% for AHost2. The few extremely high error moments can be attributed to external factors, such as the underlying OS performing system tasks, etc. Even then, for AHost2, 75% of the errors are at or below 7% and 50% of the errors are at or below about 2%. Similarly, the median memory error stays in the range of 5-7% for all 12 hosts. This demonstrates that generally, with an error of less than 7%, Cloudy accurately recreates the CPU and memory usage of the underlying trace.

The average CPU and memory errors for 12 hosts are given in Table I.

TABLE I. Actual vs. Trace Load Absolute Errors

Trace Name	CPU (%)	Memory (%)
AHost0	7.96	6.07
AHost1	6.83	17.53
AHost2	8.48	5.51
AHost3	9.23	4.79
AHost4	4.89	4.97
AHost5	7.72	4.91
GHost0	3.95	8.07
GHost1	3.88	5.97
GHost2	5.71	9.05
GHost3	3.34	6.31
GHost4	3.44	6.52
GHost5	2.19	6.43

The figure and table indicate that the percent memory usage generally has a median error of about 6%. To put this in absolute memory terms, since 16 GB instances were used, 6% equates to about 0.96 GB. This additional memory usage corresponds to the memory requirements of the underlying



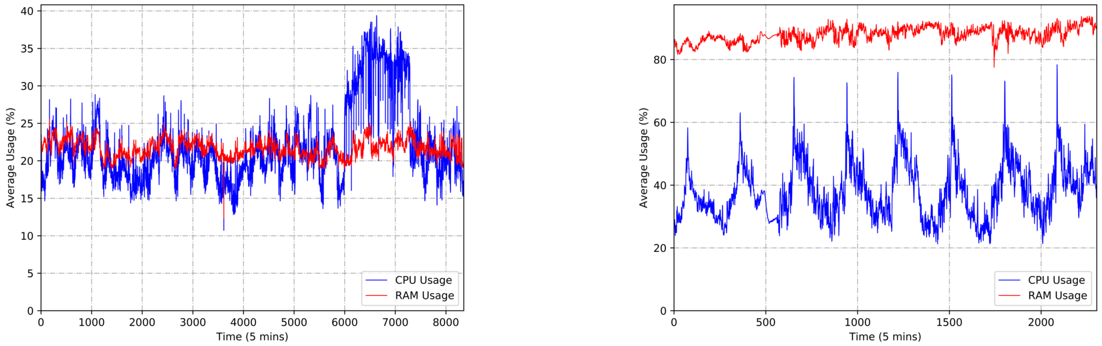
(a) Google CPU

(b) Google Memory

(c) AliBaba CPU

(d) AliBaba Memory

Fig. 8. CDFs of Min, Max, Mean, Median, and Std. Dev. for CPU and Memory of all Workloads.



(a) Google

(b) AliBaba

Fig. 9. Average CPU and Memory usage of 1000 machines over time.

operating system (OS) and its processes. If greater accuracy in memory usage is required, the maximum memory to use argument can be tweaked while starting Cloudy, to accommodate for the memory requirements for the underlying OS. For reference, Fig. 7 shows the actual load generated vs trace load for a sample workload (AHost4). It can be observed that the generated load closely matches the pattern of the load indicated by the underlying trace.

B. *Workload Characteristics*

This subsection evaluates and discusses the behavior of the underlying traces that Cloudy follows to generate the workloads. There are two main purposes of these evaluations. First, to provide the reader with an idea of the nature and type of the underlying traces. Second, to demonstrate the various types of analysis that can be performed on the workloads when using Cloudy.

In order to meet these goals, the following subsections discuss some aggregated statistics such as minimum, maximum, mean loads, and standard deviations, as well as seasonal decomposition of the loads, autocorrelation of the loads, and cross-correlation of CPU loads with memory loads. All of these characteristics for a running workload can be viewed through the visualization component of Cloudy. For this set of experiments, all 2000 traces were used. The CPU and memory usages were separated, resulting in 4000 total traces.

1) Aggregated Statistics: Fig. 8 shows four CDFs that summarize the aggregated values of the Google and Alibaba traces. The reported parameters are the maximum, minimum, mean, median, and standard deviation values over the entire duration of the traces. For the Google CPU traces, the average maximum and minimum values are 55% and 0.16%, respectively, while for Google memory traces, the average maximum and minimum values are 34.45% and 0.21%, respectively. For the Alibaba CPU traces, the average maximum and minimum values are 83.7% and 13.13%, respectively, while for the Alibaba memory traces, the average maximum and minimum values are 96.54% and 69.81%, respectively. From the figures, the standard deviations indicate, that in both Google and Alibaba traces, memory usage is generally less variable around its mean, as opposed to CPU usage that varies substantially within a single trace. Further, the Alibaba traces in general, show higher memory and CPU usages as opposed to the Google traces. Finally, the Alibaba traces show substantially high memory usage for most traces.

To offer an overall view of the traces, Fig. 9 shows the average CPU and memory usage for both Alibaba and Google traces at each instance of time. It can be seen that over all the observed workloads, the Alibaba CPU traces have a more obvious pattern than the Google traces. The memory traces in both cases, does not show an apparent pattern. However, as suggested before, it can be deduced that the Alibaba memory traces utilize more memory than the Google memory traces.

2) Seasonality and Trends: In order to analyze the periodic nature of the traces, as well as any inherent patterns, all the Google and Alibaba traces were decomposed into their trend, seasonal, and residual components. Fig. 10 shows one sample each of the Google CPU, Google memory, Alibaba CPU, and Alibaba memory traces. Decompositions for all the traces can be viewed through Cloudy. It is important to note that the x-axis scales for Google and Alibaba are different since their durations are different (24 days and 8 days respectively). Based on auditing the decompositions, similar trends and patterns exist across all Google and Alibaba traces. The Alibaba CPU traces demonstrate a clear seasonal pattern corresponding to one day. While the other three types of traces also demonstrate a seasonal pattern, the residual components for them do not seem to be simply noise (especially for the memory traces). This suggests the need for some further investigation into the inherent patterns within the memory traces.

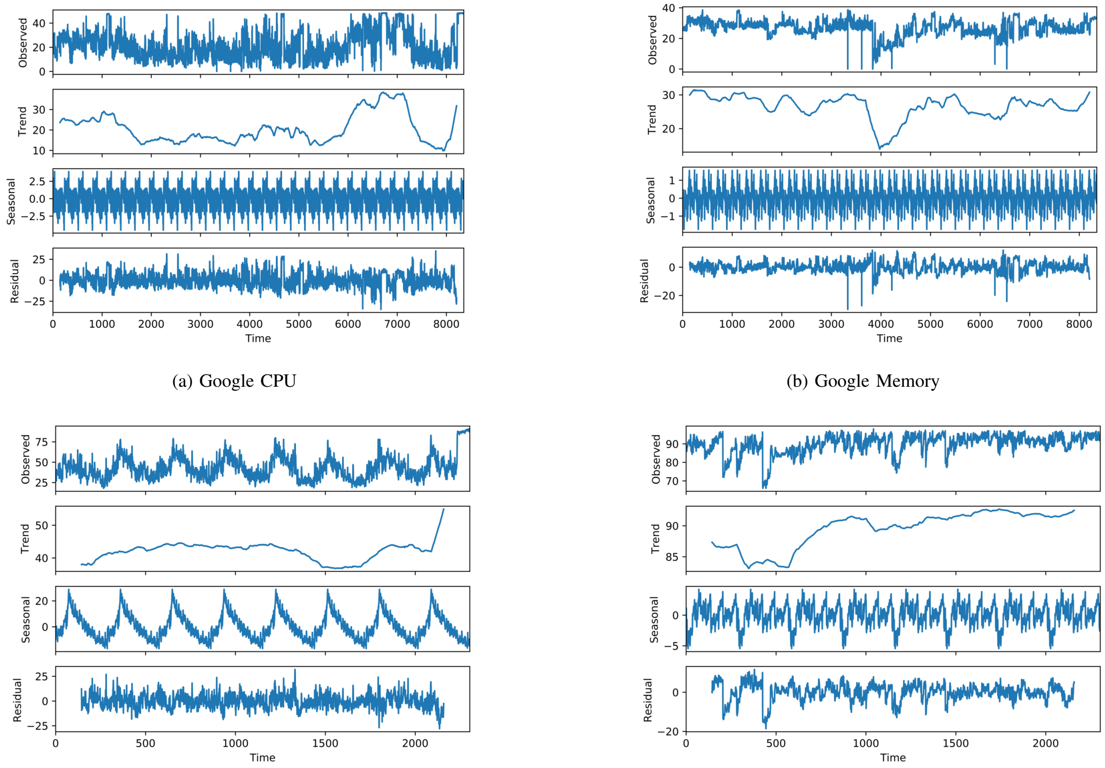
3) Autocorelation and Cross-corelation: In order to further understand whether any patterns exist in the traces, all 2000 traces were analyzed for autocorrelation. After calculating the autocorrelation function (ACF) values for each trace up to lag 800, the maximum value not at lag 0 were logged. Fig. 11a shows a boxplot of these maximum ACF values for all the traces, separated by type. The figure can provide a general idea of the amount of autocorrelation that exists on an average in these traces. It can be observed that the traces from the Alibaba CPU have higher median maximum ACF values as opposed to the other types of traces. This indicates higher autocorrelation in the Alibaba CPU traces. Similarly, on an average, lower autocorrelations can be seen in the Alibaba memory traces. The median maximum ACFs for Google CPU, Google memory, Alibaba CPU, and Alibaba memory traces are about 0.35, 0.35, 0.5, and 0.25 respectively. The observation supports the analysis from the previous section, that demonstrated high seasonality in the Alibaba CPU traces. This can be used as a starting point for further analysis into the patterns and predictabilty of the traces. Fig. 12 shows the autocorrelation plots for sample traces (one Google and one Alibaba). These plots are available from the Visualization component of Cloudy. The Alibaba CPU trace shows obvious, high peaks at non-zero lags, indicating a high degree of autocorrelation. While the Alibaba memory plot in this sample also shows a high degree of autocorrelation, that is not generally true for most other Alibaba memory traces. The Google traces do not show any prominent autocorrelation at any lag.

Another important aspect to consider for a workload on a machine is the relationship between the CPU and memory usage. Intuitively, since a running program is working with both CPU and memory, it stands to reason that for a given workload, there could be some positive or negative (in some cases) correlation between usages of the two. This analysis can prove extremely beneficial in a variety of load predicting algorithms, and can potentially provide better results than predicting on CPU or memory alone. With this in mind, cross-correlation between CPU and memory for the 2000 traces for up to lag 800 is reported. Similar to the analysis with autocorrelation, for each of the 2000 traces, the maximum value of the cross-correlation function (CCF) not at lag 0 were logged. Fig. 11b shows a boxplot of these maximum CCF values for all the traces. In this case, it can be seen that overall, there does not seem to be a strong cross-correlation between memory and CPU for either the Google or Alibaba traces. The Google traces have a slightly higher cross-correlation between memory and CPU usage, with a median maximum CCF of about 0.3, and 75% of the traces showing maximum CCF under 0.4. Compared to this, the Alibaba traces have a median maximum CCF of about 0.23, and 75% of the traces showing maximum CCF under 0.25.

V. Discussions and Future Work

The experimental results show an average memory error of approximately 1 GB. This is an important aspect to consider. The reason for this error is the memory that the underlying OS requires for its own purposes, even without Cloudy running. Typically, for the ec2 Ubuntu instances, this corresponds to a little under 1 GB. It is therefore recommended that when running Cloudy, the maximum memory to be used is specified keeping the underlying OS's requirement in mind. For example, in the scenarios described previously, Cloudy should be run at a maximum of 15 GB memory (instead of 16 GB). This will ensure that the resultant memory load matches the trace load even more closely, with negligible errors.

There are three aspects of Cloudy that are currently being worked on to make the tool more universal. The first aspect deals with the statistics logged and displayed. Currently, the performance statistics provided are recorded via stress-ng, and



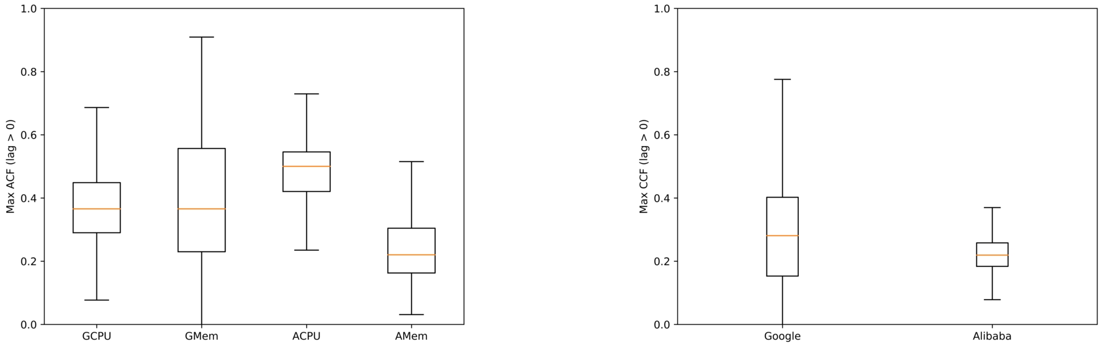
(a) Google CPU

(b) Google Memory

(c) AliBaba CPU

(d) AliBaba Memory

Fig. 10. Sample Decompositions of Workloads.



(a) Autocorrelation

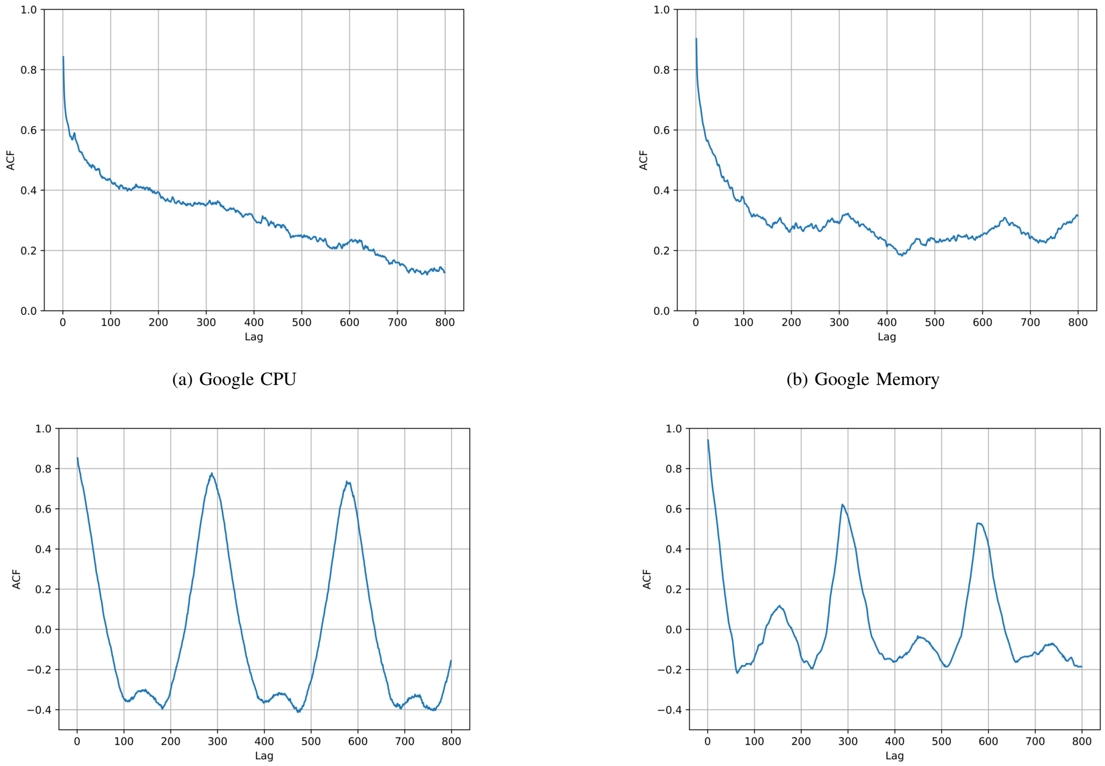
(b) Cross-correlation (CPU and memory usage)

Fig. 11. Maximum Correlation Function Values.

have to be used in that context. However, with only some slight additions and no changes to the behavior of the framework, other desired system-wide parameters can be recorded and displayed. Based on user input after release, the next update of Cloudy shall include other statistics as requested.

The second aspect is the utility used to create the load on memory, viz. stress-ng. Again, without any major changes to the behavior and code of the framework, any utility can be used to generate the memory load. For example, typical programs that are used to generate memory loads include array sorters, linear algebra solvers, matrix operators, etc. The next iteration of Cloudy aims to offer multiple stress-ng-like utilities that users can choose from, when running Cloudy. This will empower the user to select a work that is more representative of the types of load they envision in context of their testbed.

Finally, Cloudy has been tested and validated on Ubuntu based AWS ec2 instances. However, there is no part of the framework that prevents it from being run on any Linux-based distribution. Automatic install scripts for other distributions and cloud providers are currently been implemented, and shall be added to the git repository.



(a) Google CPU

(b) Google Memory

(c) AliBaba CPU

(d) AliBaba Memory

Fig. 12. Sample Autocorrelation of Workloads.

VI. Conclusion

This paper introduced a free, open-source, workload generator called Cloudy. The generator is aimed at researchers in cloud computing who need a testbed to evaluate their own research ideas. Cloudy is easy to install, non-intrusive, and can be used to quickly simulate real-world CPU and memory usage patterns in VMs, containers, cloud instances, or local machines. Through extensive experimental evaluations it was demonstrated that using Cloudy, the CPU and memory usage on a machine can closely follow one of 16000 real-world usage traces. Additional evaluations demonstrated the various analysis features of Cloudy that can allow users to further enhance their understanding of the underlying real-world loads, rather than running a black-box generator.

References

[1] M. R. Hines and K. Gopalan, "Post-copy based live virtual machine migration using adaptive pre-paging and dynamic self-ballooning," in *Proceedings of the 2009 ACM SIGPLAN/SIGOPS international conference on Virtual execution environments*. ACM, 2009, pp. 51–60.

[2] K. Z. Ibrahim, S. Hofmeyr, C. Iancu, and E. Roman, "Optimized pre-copy live migration for memory intensive applications," in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM, 2011, p. 40.

[3] Y. Al-Dhuraibi, F. Paraiso, N. Djarallah, and P. Merle, "Autonomic vertical elasticity of docker containers with elasticdocker," in *2017 IEEE 10th International Conference on Cloud Computing (CLOUD)*. IEEE, 2017, pp. 472–479.

[4] K. Qazi and S. Romero, "Remote memory swapping for virtual machines in commercial infrastructure-as-a-service," in *2019 4th International Conference on Computing, Communications and Security (ICCCS)*. IEEE, 2019, pp. 1–8.

[5] G. Moltó, M. Caballer, and C. De Alfonso, "Automatic memory-based vertical elasticity and oversubscription on cloud platforms," *Future Generation Computer Systems*, vol. 56, pp. 1–10, 2016.

[6] K. Qazi, "Vertelas - Automated user-controlled vertical elasticity in existing commercial clouds," in *2019 4th International Conference on Computing, Communications and Security (ICCCS)*. IEEE, 2019, pp. 1–8.

[7] M. Silva, M. R. Hines, D. Gallo, Q. Liu, K. D. Ryu, and D. Da Silva, "Cloudbench: Experiment automation for cloud environments," in *2013 IEEE International Conference on Cloud Engineering (IC2E)*. IEEE, 2013, pp. 302–311.

[8] D. Jayasinghe, J. Kimball, S. Choudhary, T. Zhu, and C. Pu, "An automated approach to create, store, and analyze large-scale experimental data in clouds," in *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)*. IEEE, 2013, pp. 357–364.

[9] D. Jayasinghe, G. Swint, S. Malkowski, J. Li, Q. Wang, J. Park, and C. Pu, "Expertus: A generator approach to automate performance testing in IaaS clouds," in *2012 IEEE Fifth International Conference on Cloud Computing*. IEEE, 2012, pp. 115–122.

[10] M. Cunha, N. Mendonca, and A. Sampaio, "A declarative environment for automatic performance evaluation in IaaS clouds," in *2013 IEEE Sixth International Conference on Cloud Computing*. IEEE, 2013, pp. 285–292.

[11] M. Cunha, N. Mendonça, and A. Sampaio, "Cloud Crawler: a declarative performance evaluation environment for infrastructure-as-a-service clouds," *Concurrency and Computation: Practice and Experience*, vol. 29, no. 1, p. e3825, 2017.

[12] J. Scheuner and P. Leitner, "Performance benchmarking of infrastructure-as-a-service (IaaS) clouds with cloud workbench," in *Companion of the 2019 ACM/SPEC International Conference on Performance Engineering*. ACM, 2019, pp. 53–56.

[13] "RUBiS," Aug 2019, posted at https://github.com/uillianluiz/RUBiS. (Accessed: May 2020).

[14] Google, "PerfKit benchmarker," github, 2020, posted at https://github.com/GoogleCloudPlatform/PerfKitBenchmarker (Accessed: May 2020).

[15] K. Qazi, "Cloudy," gitlab, 2020, posted at https://gitlab.com/kashifqazi/cloudy.

[16] J. Wilkes, "More Google cluster data," Google research blog, Nov. 2011, posted at http://googleresearch.blogspot.com/2011/11/more-google-cluster-data.html (Accessed: May 2020).

[17] Alibaba, "Alibaba production cluster data v2018," github, 2018, posted at https://github.com/alibaba/clusterdata/tree/v2018 (Accessed: May 2020).

Performance Evaluation of LoRa ES920LR 920 MHz on the Development Board

Puput Dani Prasetyo Adi[1,2], Akio Kitagawa[2]

Electrical Engineering Department, University of Merdeka Malang, Malang City, East Java, Indonesia[1]
Micro Electronics Research Laboratory, Kanazawa University, Kanazawa, Ishikawa, Japan[1, 2]

***Abstract*—This study contains the LoRa ES920LR test on obstruction or resistance conditions and its comparison with Free Space Path Loss (FSPL) using Drone means without obstacles. The ES920LR has 920 MHz frequency channel settings, 125 kHz Bandwidth, and SF 7-12, with 13 dBm Output Power. Comes with Sleep mode operation with Command Prompt based settings. The development board used is a leafony board, Leafony board is a board with a small size that is compatible with Arduino IDE, using a micro ATmega M328P microcontroller, this board is mounted tiled with complete facilities, e.g., a power supply board, four different sensor boards, MCU boards, communication boards e.g., WiFi and Bluetooth, specifically in this article using the LoRa ES920LR on the Leafony board, using LoRa because it requires a long range (km) and low power, and expansion boards that can be developed, Expansion to Leafony boards is expected to reduce the power consumption of the sensor node, lifetime, lif, small size, and lightweight. Furthermore, this algorithm is used to optimize LoRa Coverage and LoRa Lifetime. The results of receiving the FSPL LoRa 920LR have a Power Receiver (Pr) of 30 dB at a distance of 1 meter or A, at 500 meters 85 dB, 1500 meters 95 dB, 2.5km 100 dB. Attenuation is caused by distance, although not significant, other factors are obstacles or obstacles, bad weather (rain, snow).**

Keywords—Coverage; lifetime; low power; lightweight; long range; development; board; free Space; drone

I. Introduction

Lora ES920LR is a type of LoRa that works at 920 MHz. LoRa ES920LR is indeed used in Japan, principally it is important to study the Quality of Service (QoS) of LoRa as seen from various factors such as Radio Propagation, Ability of LoRa eg, Power save management, bandwidth, SF, Output Power, Sleep management mode, ability on Receive Signal Strength Indicator (RSSI)[14],[16], etc. In research [1] monitoring of soil and environmental information was collected in 17 km, in research [2] using Semtech products at a 915 MHz LoRa frequency. moreover, compatibility and configuration on the Application server e.g., TTN (The Things Network), Ambient.io, or Thingspeak are important that communication between a LoRa module can run properly. Previous research used the same sensor, Pulse sensor [3],[15], in research [4], LoRa was used in the analysis of WaterGrid-Sense, as a full-stack node based on LoRa deployed on a smart water management system.

Furthermore, it is important to pay attention to the Power Consumption factor as in research [5], [6], the results of data compression in wireless sensors nodes that use LoRa technology to transmit data. An energy consumption comparison is made with other data communication protocols used in WSN. In managing the LoRa and LoRaWAN networks, an appropriate protocol is needed in setting the Medium Access Control (MAC) random access, in research [7], using ALOHA protocol analysis, ie, normalizing the communication of LoRaWAN networks using a Reservation-ALOHA (R-ALOHA). With the continued development of LoRa and LoRaWAN-based End nodes, a protocol that is capable of handling millions of end nodes is needed, it also requires continuous renewal in terms of performance, security [8], robustness, in research [9] discussing the revolution of LoRaWAN network technology considering the IoT requirements. In addition to Software Development, Protocols with security and methodology settings, LoRa acceleration is also developed with various devices, such as Leafony Board and PSoC [10]. In this study, a blackboard-based board developed specifically for LoRa ES920LR communication and applied to drones and GPS technology to obtain SF, SNR [17], and RSSI values, the application in subsequent studies is to get patients (Longitude and Latitude) that can be accessed with Google maps, furthermore, it can be seen that each patient's BPM is in a different location. Because of the shape of the Leafony board is small, it can be placed and flown by a drone.

II. Theory

A. Receiver Sensitivity of ES920LR

As the purpose of this research is an analysis of RSSI, SNR, and other parameters such as power consumption, Gain, Power Receiver, obstacles [Fig. 1], receiver sensitivity, Time on Air, and how to manage the uplink and downlink data on the server or internet gateway and application the server. Moreover, Power Receiver depends on several factors, including receiver gain and lost signal packet factor or attenuation on the Free Space Path Loss. An illustration of the signal reduction can be seen in Fig. 1. The LoRa ES920LR has a transmitting Power (PTX) of -13 dBm, and the Power Receiver depends on the current state of signal propagation including the presence of Obstacles for all circumstances, e.g., bad wheater (snow, rain [18],[21],[22]), the distance that will affect the Time On Air (ToA), The antenna used, Connector Loss and Gain, we can refer to as Receiver Sensitivity (s).The sensitivity of LoRa (-dB) can be seen in equation 1. Where S or sensitivity depends on the value of Bandwidth, Noise Fig. 2, and Signal Noise Ratio (SNR). Table I shows the relationship between SF, S, and ToA. Moreover, the factor to consider is the Fresnel Zone calculation factor [12] if the transmitter and receiver distance is ≥ 5 km [13].

$$S= -174 + 10 \log_{10}(BW)+ NF+SNR_{limit} \quad (1)$$

B. Time on Air (ToA) of ES920LR

The time needed for data from the Transmitter (Tx) to arrive right at the receiver (Rx) while in the air is Time On Air (ToA). [11] ToA is calculated using equation 2 and equation 3.

$$ToA = T_{\text{ preamble}} + T_{\text{ Payload}} \quad (2)$$

$$T_{preamble} = Nb_{\text{ Preamble}}\ (8)+\text{symbols added by radio } (4.25) \text{ x } T_{\text{symbol}} \text{ ,}$$

$$T_{payload} = Nb_{PayloadSymbol} \text{ x } T_{symbol}$$

$$N_{payload} = 8 + \max(\text{ceil } [(8PL-4SF+28+16CRC-20IH)/ 4(SF-2DE)]\ (CR+4),0) \quad (3)$$

Furthermore, the relationship between SF, Chips, SNRlimit, ToA, and BitRate is shown in Table II.

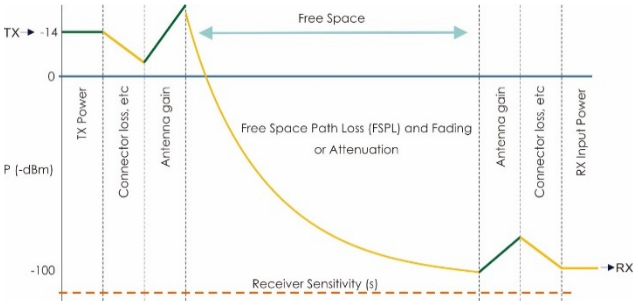


Fig. 1. Condition on the Signal Transmitting Process.

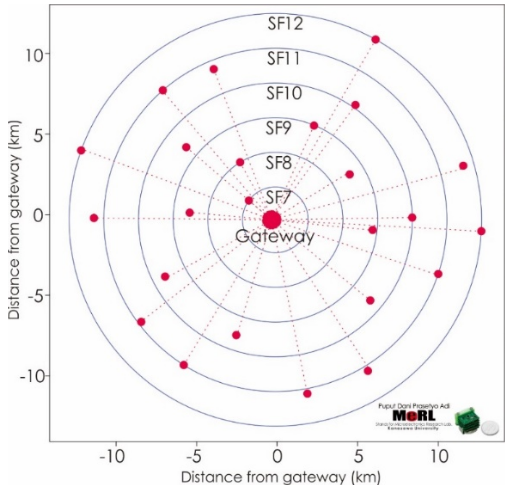


Fig. 2. Spreading Factor.

TABLE I. SENSITIVITY, TOA, AND SPREADING FACTOR

Spreading Factor (SF)	Sensitivity (S)	Time on Air (ToA)
SF7	-123 dBm	41 ms
SF8	-126 dBm	42 ms
SF9	-129 dBm	144 ms
SF10	-132 dBm	288 ms
SF11	-134.5 dBm	577 ms
SF12	-137 dbm	991 ms

TABLE II. LORA SPREADING FACTOR COMPARATION WITH SNR_{LIMIT}

Spreading Factor (SF)	Chips /Symbol	SNR_{Limit}	ToA (10 byte Packet)	Bitrate
7	128	-7.5	56 ms	5470 bps
8	256	-10	103 ms	3125 bps
9	512	-12.5	205 ms	1758 bps
10	1024	-15	371 ms	977 bps
11	2048	-17.5	741 ms	537 bps
12	4096	-20	1483 ms	293 bps

C. Budget Link (-dB) of ES920LR

Calculation of LoRa Budget Link ES920LR LoRa specification can be seen on equation 4, Before is determined the LoRa ES920LR sensitivity value, then as a deduction for the Transmit Power (dB) ES920LR LoRa. Moreover, Fig. 3 is Time On Air from the ES920LR LoRa, and Fig. 4 is the Link Budget (-dB) ES920LR with a -13 dB Power transmit.

$$\text{Budget Link (dB)} = \text{Tx Power (dB)} - \text{Sensitivity (dB)} \quad (4)$$

In detail, the ToA of ES920LR with different bandwidths (125, 250, and 500 kHz) is shown in Table III.

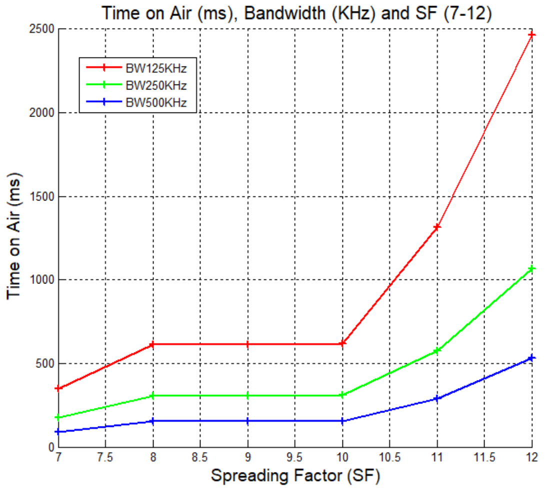


Fig. 3. Time on Air (ToA) ms.

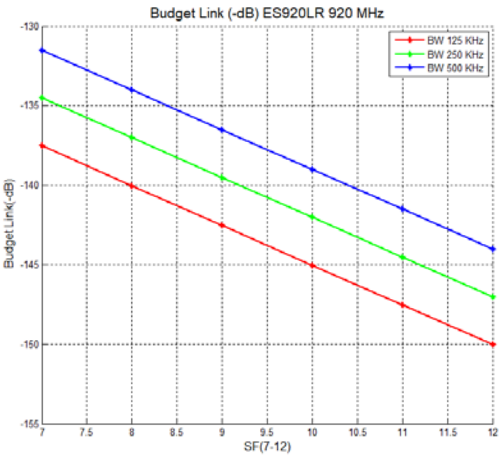


Fig. 4. Link Budget (-dB).

TABLE III. TIME ON AIR COMPARISON WITH BW, AND SF

SF	BW125KHz	BW250KHz	BW500KHz
7	348.42 ms	174.21 ms	87.1 ms
8	614.91 ms	307.46 ms	153.73 ms
9	615.42 ms	307.71 ms	153.86 ms
10	616.45 ms	308.22 ms	154.11 ms
11	1314.82 ms	575.49 ms	287.74 ms
12	2465.79 ms	1069.06 ms	534.53 ms

III. METHODS

A. Flowchart, Devices and Design

The flowchart in Fig. 6 is the steps used in this research, very clearly illustrated in the flowchart, that the initial step is to build a sensor node. Before using the Arduino Pro Mini, the Arduino board [23] is used to make it easier to get a tx and Rx pin [Fig. 7]. After success, furthermore, a communication test Between ES920LR LoRa [Fig. 5] was conducted.

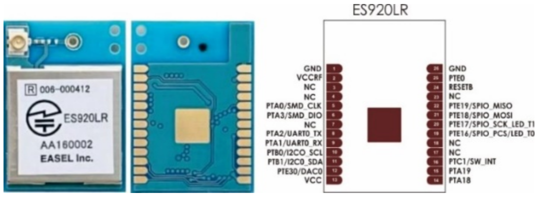


Fig. 5. ES920LR Pins.

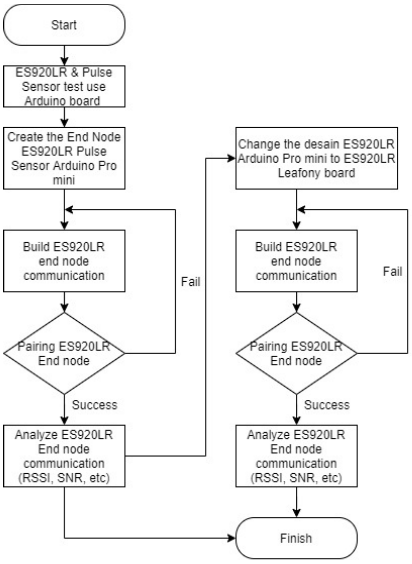

Start
ES920LR & Pulse Sensor test use Arduino board
Create the End Node ES920LR Pulse Sensor Arduino Pro mini
Build ES920LR end node communication
Fail
Pairing ES920LR End node
Success
Analyze ES920LR End node communication (RSSI, SNR, etc)
Change the desain ES920LR Arduino Pro mini to ES920LR Leafony board
Build ES920LR end node communication
Fail
Pairing ES920LR End node
Success
Analyze ES920LR End node communication (RSSI, SNR, etc)
Finish

Fig. 6. Flowchart on this Research.

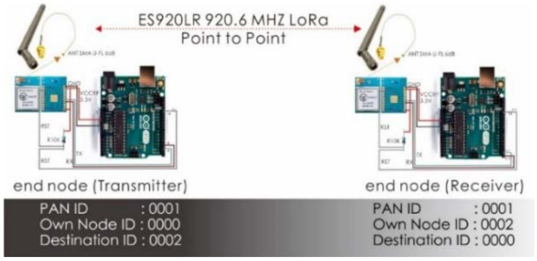
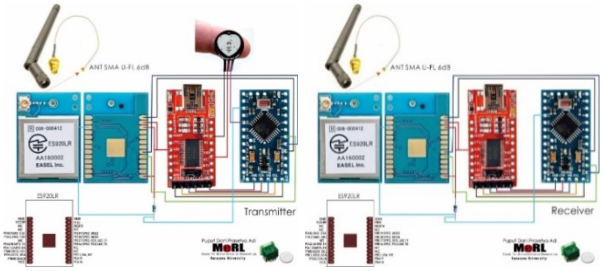


Fig. 7. Pairing ES920LR LoRa use Arduino board (Design_1).

Table IV is a detailed specification of the LoRa ES920LR, Fig. 7 and Fig. 8 are the steps in making a Sensor Node. Fig. 7 still uses Arduino Uno, while Fig. 8 has been changed to Arduino Pro mini. The transformation of Fig. 7 and 8 was completed in this research objective, namely using the Leafony Board in Fig. 11. Moreover, Fig. 9 is a mesh communication that can be carried out by the ES920LR in collaboration with the ES920GW and Application server. LeafBus in Leafony is shown in Fig. 10, there are 5 pins used by LoRa ES920LR.

TABLE IV. ES920LR SPESIFICATION

Spesification	Description
Model	ES920LR
JAPAN Government Certification /Standar ISM Band	ARIB STD-T108
Frequency	920.6 – 928.0 MHz
Modulation type	LoRa Modulation CSS (Chirps Spread Spectrum)
Number of Channels	37 ch (at 125 kHz bandwidth or less)
	18 ch (at 250 kHz bandwidth)
	12 ch (at 500 kHz bandwidth)
Bandwidth	62.5 kHz – 500 kHz
Spreading Factor	7-12
Transmission Speed	146 bps – 22 kbps
Transmission Output	13 dBm (20 mW)
Receiver Sensitivity	-118 dBm ~ -142 dBm
MCU	ARM Cortex M0+
Memory	Flash ROM : 128 KB, RAM : 16 KB
Power Consumption	Tx : 43 mA (13 dBm setting) Rx : 20 mA During Sleep : 1.7 uA (when the timer starts)
interface	UART, SPI, I2C, ADC, GPIO
Antenna	Wire Antenna, External Antenna (U.FL)
Power Supply Voltage	2.4 Volt to 3.6 Volt
Operating Temperature range	-40 ~ +85^o celcius
Connection Terminal	26QFN
Board Mounted PCB	SMT mounting type
Dimensions	24.00 x 17.0 x.2.3 mm
Construction Design Certification acquired	Certification Number : 006-000412



(a)Transmitter (b) Receiver

Fig. 8. Pairing ES920LR LoRa use Arduino Pro mini (Design_2).

Table V is the connection pin between ES920LR and Arduino, reset is on pin D12, because Pin 13 is used by the LED on the arduino Pulse sensor interrupt program.

On same time, setting the Receiver Pins, e.g., Table VI.

The following pseudocode can facilitate the understanding of how to pair or communicate between the ES920LR End nodes.

```
Pseudocode of LoRa E920LR_arduino

1.  Describe the E920LR Library
    #include <SoftwareSerial.h>
    #define LORA_RECV_RecvData 100
    #define ES920LR_RST_PIN 13
    #define LORA_RX 2
    #define LORA_TX 3
2.  Describe the PAN ID and Destination ID
    String dstId = "00010002";
3.  Determine the Maximum Sending the Data
    const int maxSendTimes =50;
4.  Determine How Long the Delay
    const int setCmdDelay = 100;
5.  Determine the type of Communication
    SoftwareSerial LoRa_Serial(LORA_RX, LORA_TX);
6.  Determine the Output Pin and delay
    pinMode(ES920LR_RST_PIN, OUTPUT);
    digitalWrite(ES920LR_RST_PIN, LOW);
    delay(100);
    digitalWrite(ES920LR_RST_PIN, HIGH);
    delay(1500);
7.  Describe the BoudRate or Speed the data sending
    Serial.begin(9600);
    LoRa_Serial.begin(9600);
8.  Describe the Case or Scope on the Program
    loraInit();
9.  Describe the type of data sending,delay and Looping
    String sendData = "";
    for (int i = 1; i <= maxSendTimes; i++) {
    sendData = dstId + "Times" + i;
    delay(2000);
    Serial.println(sendData);
    LoRa_Serial.println(sendData);
    while(LoRa_Serial.available()>0) LoRa_Serial.read();
    delay(4000);

10. Initialize the LoRa Parameter
    void loraInit()
        •  LoRa Serial initialize Node type (ED or Coor)
        •  LoRa Serial initialize Bandwidth of Node
        •  LoRa Serial initialize Spreading Factor (SF)
        •  LoRa Serial initialize The Channel of node
        •  LoRa Serial initialize PAN ID
        •  LoRa Serial initialize OWN ID
        •  LoRa Serial initialize DEST ID
        •  LoRa Serial initialize Acknowledge
        •  LoRa Serial describe the Retry count
        •  LoRa Serial describe the type of Transfer mode
        •  LoRa Serial describe the RSSI ON or OFF
        •  LoRa Serial describe the Transfer mode (Frame
           or Payload)
        •  LoRa Serial initialize a Save Command
        •  LoRa Serial initialize run command

11. Describe the Reading Command
    while(LoRa_Serial.available()>0) LoRa_Serial.read();
```

TABLE V. CONNECTION PIN ON TRANSMITTER

ES920LR	Arduino
GND	GND
VCCRF	3.3 Volt
TX	D2
RX	D3
VCC (3.3 Volt)	3.3 Volt
RST	D12

TABLE VI. CONNECTION PIN ON RECEIVER

ES920LR	Arduino
GND	GND
VCCRF	3.3 Volt
TX	TX 1
RX	RX 0
VCC (3.3 Volt)	3.3 Volt
RST	D13

The development system can be seen in Fig. 3. Fig. 9 uses the ES920LR Leafony Board that mesh with each other on end-devices and sends sensor data to the edge router or border router to the ES920GW as an Internet Gateway and displays it on the internet server.

The difference between Leafony boards and other boards is the port in Fig. 10. LeafBus Leafony boards are pins that are used to connect with other boards. LeafBus LoRa uses these five LeafBus pins i.e., 3V3, Reset, Pins 8 (F11), and Pin9 (F13) as Tx and Rx, and GND. Fig. 11 is the LoRa ES920LR design on the Leafony board, created using kiCAD Software, as described in Fig. 4, there are five LoRa pins identified i.e., Tx, Rx, 3V3, GND, and Reset Pin. And there is the addition of a 10 k Ohm resistor on the Reset pin to Pin 13 and V3V.

Therefore, there were three times the changes in the form of LoRa ES920LR end node, i.e., end node LoRa ES920LR use Arduino Uno, LoRa ES920LR use Pro mini, and LoRa ES920LR Leafony board, as shown in Fig. 12.

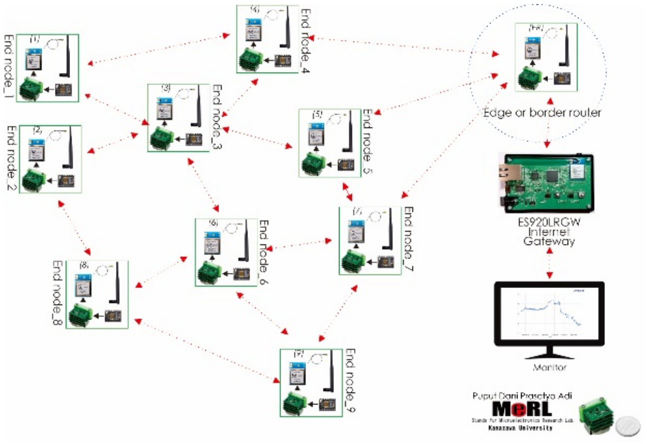


Fig. 9. Mesh Communication of Leafony board LoRa.

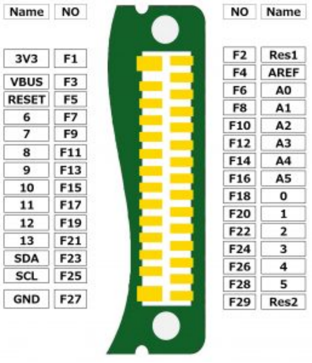


Fig. 10. LeafBus Leafony Board.

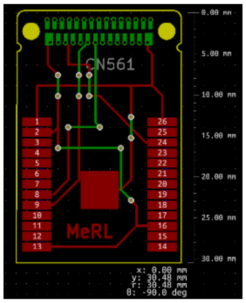


Fig. 11. Desain to LoRa ES920LR Leafony (Design_3).

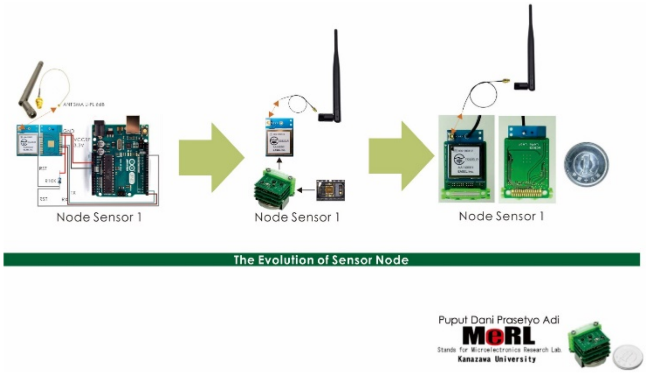


Fig. 12. Sensor node Evolution.

B. *ES920LR Radio Propagation*

Before the ES920LR was made in the form of a Leafony board, the Arduino Pro mini-board was tested first. Please note, the mini-board pro requires FTDI. FTDI functions as a regulator as well as a programmer board to program the Arduino Pro mini. The LoRa ES920LR transmitting data experiment uses drones using the FSPL approach. Moreover, Free Space Path Loss (FSPL) is a condition where the process of sending sensor data from the transmitter to the receiver does not pass through any obstacles, the application of Free Space Path Loss (FSPL) using a drone is one of the right efforts to get the value of Receive Signal Strength (RSS) without obstacles. Furthermore, Fig. 13 is one method of using drones in the process of transmitting LoRa data. There are two equations about FSPL (-dB) in RSSI (-dBm) or Power Receiver (Pr) parameters. Where the equation Pr (-dB) as equation x shows. Where the value of c = 3 x 10^8 or 299 792 458 m / s, with frequency (f) LoRa ES920LR is 920000000 Hz, so that the wavelength λ = 0.3258613673913043 meters or 32.5861367 cm or 325.861367 mm.

Equation 5 is the FSPL formula obtained from the specific wavelength and frequency parameters of the LoRa ES920LR, and equation 6 is a logarithmic equation with a value reduction of -147.55.

$$FSPL = (4\pi d/\lambda)^2 \text{ or } (4\pi df/c)^2 \text{ from } \lambda = c/f \quad (5)$$

$$FSPL\ (dB) = 20\log_{10}(d) + 20\log_{10}(f) - 147.55 \quad (6)$$

Developing an equation for LoRa Path Loss by considering the environment (n or exponent) or data transmission area as equation 7 [20].

$$PL(\mathrm{d}) = \mathrm{PL}\ (\mathrm{d0}) + 10\mathrm{nlog}\left(\frac{d}{d0}\right) + \mathrm{X}\sigma \quad (7)$$

Fig. 14 is an example of Lora's ES920LR test area with Obstacle buildings and trees. This will produce a different RSSI (-dBm), RSSI can't experience attenuation even with longer distances, this is wherefore the signal is a combination of reflected and direct signals. FSPL gives the direct signal greater amplitude of the waves which causes RSSI (-dBm) greater than those affected by obstacles that cause reflected, scattered, and diffraction.

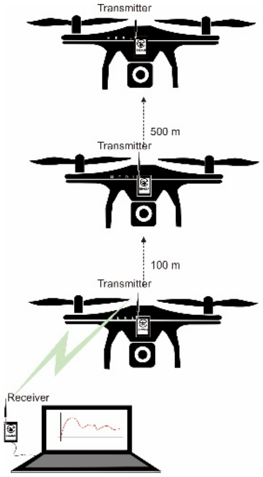


Fig. 13. LoRa Drone on FSPL Test.

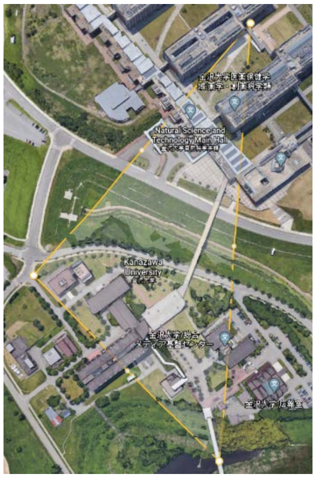
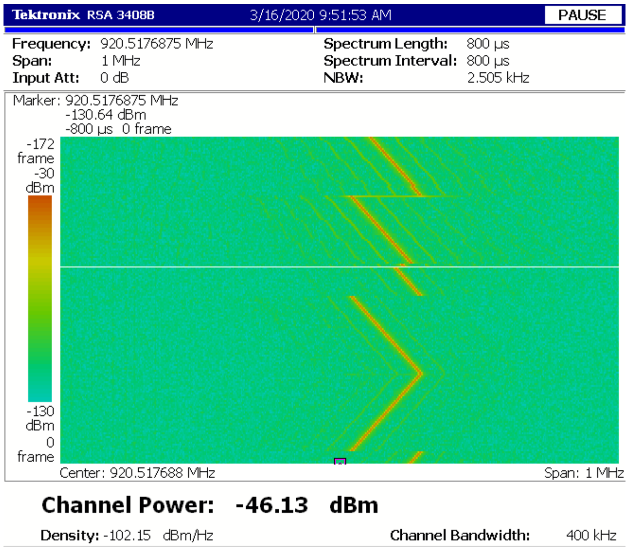


Fig. 14. ES920LR LoRa Test on Building and Trees Obstacles.

IV. Result and Discussion

An important part of this chapter is an analysis of the signals generated by ES920LR with an approach to Free Space Path Loss. For the signal analyzer used Textronix RSA 3408B, LoRa signal contains preamble and symbol, the amount of bytes of data is 255 bytes, on ES920LR 293 bytes on SF12, seen in Fig. 15 is a chirp which shows the existence of LoRa signal, the resulting chirp is up-chirp and down -chirp. Channel power shows how close the distance between transmitters - receiver, -46.13 dBm shows the value of Signal strength, in general, this value is greater based on the distance of the transmitter-receiver that is increasingly far away until a loss occurs.

Fig. 16 is another analysis of the LoRa signal i.e., Carrier Frequency, in this section, the signal power (-dBm) can be set, seen in the -30 dBm signal analyzer. The real signal power is -49.66 dBm (-83 dBm / Hz). Fig. 17 shows the value of FSPL (-dB) using the FSPL LoRa equation at 920 MHz frequency according to equations 4 and 5.



Tektronix RSA 3408B 3/16/2020 9:51:53 AM PAUSE

Frequency: 920.5176875 MHz Spectrum Length: 800 µs
Span: 1 MHz Spectrum Interval: 800 µs
Input Att: 0 dB NBW: 2.505 kHz

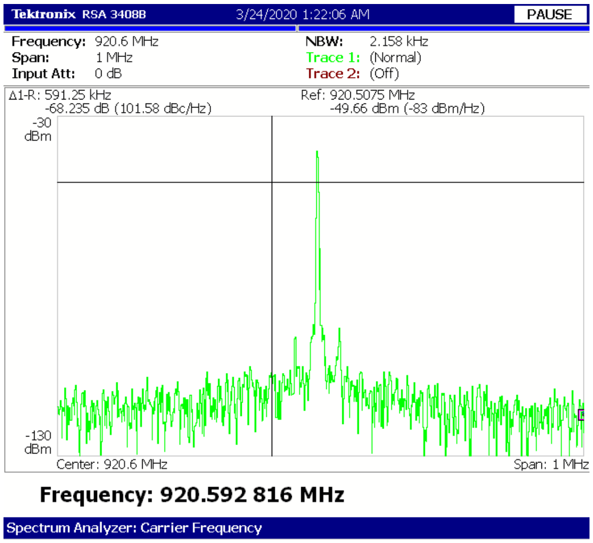
Marker: 920.5176875 MHz
-130.64 dBm
-800 µs 0 frame

Center: 920.517688 MHz Span: 1 MHz

Channel Power: -46.13 dBm

Density: -102.15 dBm/Hz Channel Bandwidth: 400 kHz

Fig. 15. Chirps LoRa 920LR.



Tektronix RSA 3408B 3/24/2020 1:22:06 AM PAUSE

Frequency: 920.6 MHz NBW: 2.158 kHz
Span: 1 MHz Trace 1: (Normal)
Input Att: 0 dB Trace 2: (Off)

Δ1-R: 591.25 kHz Ref: 920.5075 MHz
-68.235 dB (101.58 dBc/Hz) -49.66 dBm (-83 dBm/Hz)

Center: 920.6 MHz Span: 1 MHz

Frequency: 920.592 816 MHz

Spectrum Analyzer: Carrier Frequency

Fig. 16. Frequency Carrier of LoRa 920.592 MHz.

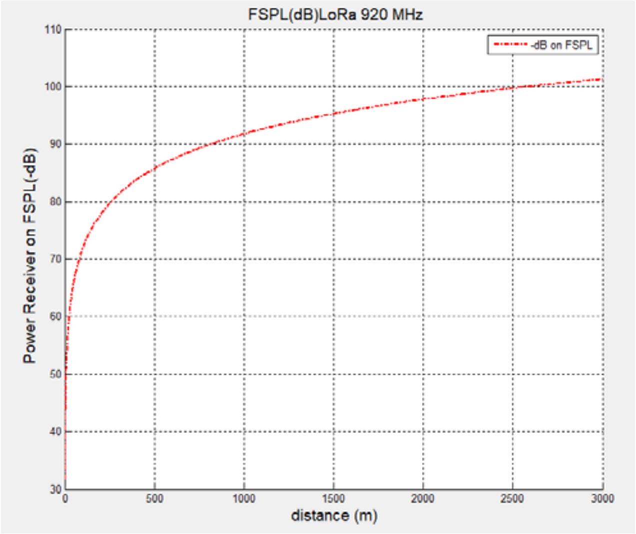


Fig. 17. FSPL use Drone LoRa ES920LR.

Fig. 18 is a real experiment from this research, where the transmitter and receiver have different distances, with the transmitter fixed position, and the receiver moves according to the specified point. RSSI data at a distance of 400 meters to 800 meters do not show regular attenuation, this is due to different levels of obstacles, at a distance of 400 meters there are many obstacles that block the direct signal to the receiver, but at a distance of 700-800 meters, the direct signal is greater than reflected signal. thus producing a combined signal that gives a smaller RSSI (-dBm) value or a stronger signal.

While Fig. 19 is an RSSI and SF approach, there is no equation or relationship between SF and RSSI in the equation, because SF talks about the time between Transmitter-receiver in sending data or signals. This is shown in equation 8. [16].

$$T_{sym} \; or \; T_S = \frac{2^{SF}}{BW} \qquad (8)$$

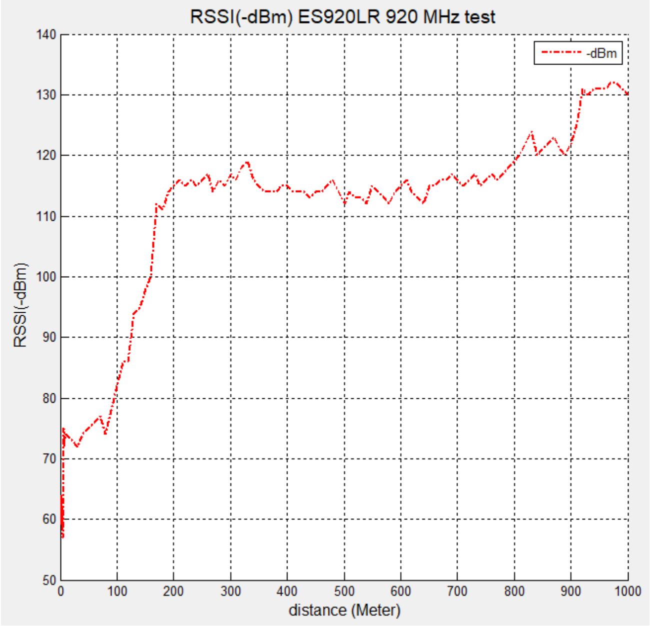


Fig. 18. RSSI (dBm) Real Experiment.

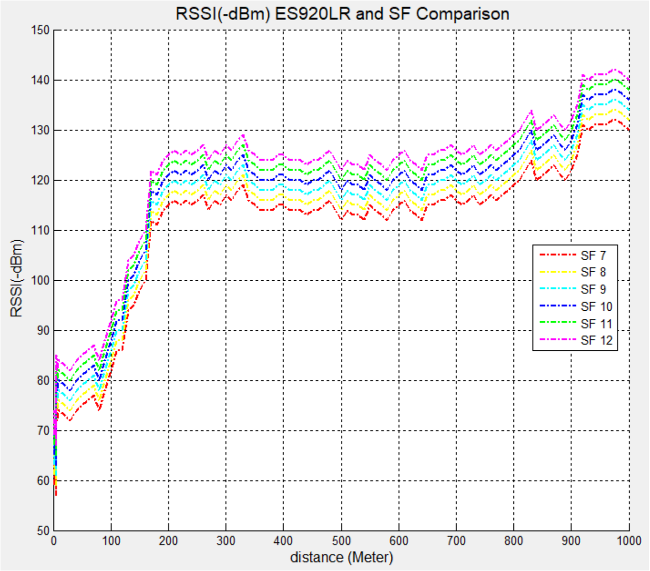


Fig. 19. SF and RSSI (dBm) Comparison.

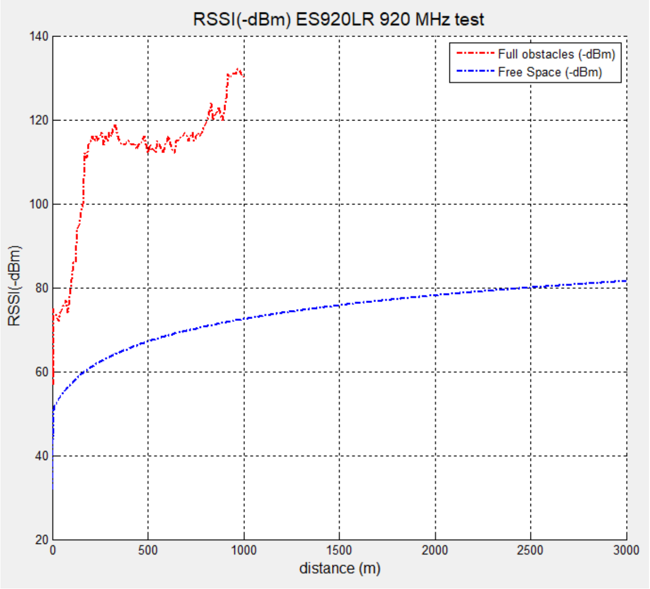


Fig. 20. FSPL and Obstacle (-dBm) RSSI Comparison.

Time is influenced by SF and BW, the greater SF (12), and the smaller the bandwidth, the greater travel time, and vice versa. This causes attenuation, the farther the distance between the transmitter and receiver (SF12). Then the weaker the signal produced, accordingly, the minimum RSSI is -120 dBm, so the greater the Spreading Factor (Fig. 2) [19], the weaker the signal produced.

Finally, if these signals are combined (Fig. 20), they will produce very different signals, between FSPL and full Obstacle, at a distance of 1000 meters, the full obstacle signal is lost and down. This is why the FSPL approach uses drones, the resulting value is indeed not significant on equations 4 and 5. However, it is close to the RSSI (-dBm) value.

In addition to using drones or FSPL [Fig. 17], testing is carried out using Tx and Rx ES920LR transmitting in areas of buildings and trees [Fig. 18]. So that the comparison is obtained as in Fig. 20. Fig. 15 and Fig. 16 are output Chrip and Signal LoRa ES920LR in real-time using Signal Analyzer. The change in the Spreading Factor causes a change in Time on Air and causes an attenuation signal to the receiver (Pr) or RSSI (-dBm) shown in Fig. 19.

V. Conclusion

Receiving a Signal Strength indicator (RSSI) on the ES920LR generally decreases based on distance or Time On Air (ToA), so that ToA can cause attenuation signal strength. Free Space Path Loss (FSPL) is a condition where there are no obstacles and signals that are better than this experimental research where experiments are conducted with many obstacles e.g., buildings and trees. Furthermore, drones are the solution to get FSPL values on Propagation radio. In the full obstacle situation the ES920LR loses the LoRa signal at 1 km distance, but can be re-tested by changing the measurement area so that the mileage is in accordance with the LoRa ES920LR specifications.

VI. Future Work

End node devices have to the ability to long life or survive longer with a Power Supply Battery, and a wide range of distances, small and lightweight e.g., Leafony LoRa board for the drone, in the network design that is made there are additional gateways to reduce the load heavily of the gateway in accommodating the end node sensor data to maintain the stability and value of Packet Receive Ratio (PRR (%)).

Acknowledgment

Thanks to the Micro Electronics Research Laboratory (MeRL) of Kanazawa University Japan, and Professor Akio Kitagawa for providing facilities, mentoring, for the progress of the author, so that this paper can be resolved properly.

References

[1] D. Taskın, S. Yazar, "A Long-range Context-aware Platform Design For Rural Monitoring With IoT In Precision Agriculture", agora university, romania, 2020, doi.10.15837/ijccc.2020.2.3821.

[2] Puput Dani Prasetyo Adi and Akio Kitagawa, “A Study of LoRa Performance in Monitoring of Patient’s SPO2 and Heart Rate based IoT” International Journal of Advanced Computer Science and Applications(IJACSA), 11(2), 2020. doi. 10.14569/IJACSA.2020. 0110232.

[3] Muhammad Niswar, Amil Ahmad Ilham, Elyas Palantei, Rhiza S. Sadjad, Andani Ahmad, Ansar Suyuti, Indrabayu, Zaenab Muslimin, Tadjuddin Waris, Puput Dani Prasetyo Adi, Performance evaluation of ZigBee-based wireless sensor network for monitoring patients' pulse status, 2013 International Conference on Information Technology and Electrical Engineering (ICITEE) DOI: doi/10.1109/ICITEED.2013. 6676255.

[4] Oratile Clement Khutsoane, Bassey Isong, Naison Gasela, Naison Gasela Adnan M. Abu-Mahfouz, "WaterGrid-Sense: A LoRa-based Sensor Node for Industrial IoT applications", IEEE Sensors Journal PP(99):1-1, DOI: 10.1109/JSEN.2019.2951345.

[5] D. I. Săcăleanu, R. Popescu, I. P. Manciu, and L. A. Perişoară, "Data Compression in Wireless Sensor Nodes with LoRa", ECAI 2018 - International Conference – 10th Edition Electronics, Computers and Artificial Intelligence 28 June -30 June, 2018, Iasi, ROMÂNIA.

[6] Philipp Mayer, Michele Magno, "LoRa vs. LoRa: In-Field Evaluation and Comparison For Long-Lifetime Sensor Nodes", DOI: 10.1109/IWASI.2019.8791362, 2019 IEEE 8th International Workshop on Advances in Sensors and Interfaces (IWASI), June 2019.

[7] Dina Hussein, Dina M. Ibrahim, "Improving LoRaWAN Performance Using Reservation ALOHA", Journal of Information Technology Management, ISSN : 2423-5059, DOI: 10.22059/jitm.2020.75792.

[8] Amira Naa, Naa Mohamed Tahar, Bakiri Mohammed, "Design and development of a secure sensor node based on the LoRa protocol and a chaotic encryption block",University of Sciences and Technology Houari Boumediene Faculty of Electronics and Computer Science Department of Electronics, August 2019.

[9] Dina M. Ibrahim, Dina Hussein, "Internet of Things Technology based on LoRaWAN Revolution", 2019 10th International Conference on Information and Communication Systems (ICICS), DOI: 10.1109/IACS.2019.8809176.

[10] Tuyen Phong Truong, Hai Toan Le, Tram Thi Nguyen, "A reconfigurable hardware platform for low-power wide-area wireless sensor networks", Journal of Physics Conference Series 1432:012068, DOI: 10.1088/1742-6596/1432/1/012068.

[11] Ferran Adelantado, Xavier Vilajosana, Pere Tuset-Peiro, Borja Martinez, Joan Melià-Seguí, Thomas Watteyne, "Understanding the Limits of LoRaWAN", IEEE Communications Magazine (Volume: 55 , Issue: 9 , Sept. 2017), DOI: 10.1109/MCOM.2017.1600613.

[12] Puput Dani Prasetyo Adi and Akio Kitagawa, “Performance Evaluation of E32 Long Range Radio Frequency 915 MHz based on Internet of Things and Micro Sensors Data” International Journal of Advanced Computer Science and Applications(IJACSA), 10(11), 2019. DOI: 10.14569/IJACSA.2019.010110.

[13] Adi, P.D.P, Kitagawa, "A performance of radio frequency and signal strength of LoRa with BME280 sensor", Telkomnika (Telecommunication Computing Electronics and Control),Issue 2, 1 April 2020, Pages 649-660, DOI: 10.12928/telkomnika.v18i2.14843.

[14] Puput Dani Prasetyo Adi and Akio Kitagawa, “ZigBee Radio Frequency (RF) Performance on Raspberry Pi 3 for Internet of Things (IoT) based Blood Pressure Sensors Monitoring” International Journal of Advanced Computer Science and Applications(IJACSA), 10(5), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100504.

[15] Puput Dani Prasetyo Adi and Akio Kitagawa, “Quality of Service and Power Consumption Optimization on the IEEE 802.15.4 Pulse Sensor Node based on Internet of Things” International Journal of Advanced Computer Science and Applications(IJACSA), 10(5), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100518.

[16] Jansen C.Liando, Amalia Gamage, et.al, "Known and Unknow Facts of LoRa : Experiences from a Large-scale Measurement Study", https://doi.org/10.1145/3293534.

[17] Amir Muaz Abdul Rahman, "Performance Analysis of LPWAN Using LoRa Technology for IoT Application", International Journal of Engineering & Technology, 7 (4.11) (2018) 212-216.

[18] MD Hossinuzzaman, Dahlila Putri Dahnil,"Enhancement of Packet Delivery Ratio during Rain Attenuation for Long Range Technology", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 10, 2019.

[19] Francisco Helder C. dos Santos Filho, "Performance of LoRaWAN for Handling Telemetry and Alarm Messages in Industrial Applications", MDPI, Sensors 2020, 20, 3061; doi:10.3390/s20113061.

[20] Puput Dani Prasetyo Adi and Akio Kitagawa, “Performance Evaluation WPAN of RN-42 Bluetooth based (802.15.1) for Sending the Multi-Sensor LM35 Data Temperature and RaspBerry Pi 3 Model B for the Database and Internet Gateway” International Journal of Advanced Computer Science and Applications(ijacsa), 9(12), 2018. http://dx.doi.org/10.14569/IJACSA.2018.091285.

[21] Fatin Hamimi Mustafa, "Effect of Rain Attenuations on Free Space Optic Transmission in Kuala Lumpur", Proceeding of the International Conference on Advanced Science, Engineering and Information Technology 2011, ISBN 978-983-42366-4-9.

[22] Sujan Shrestha, Dong-You Choi, "Rain attenuation statistics over millimeter wave bands in South Korea", Journal of Atmospheric and Solar, Terrestrial Physics 152 (2017) 1–10.

[23] Puput Dani Prasetyo Adi and Rahman Arifuddin, Design of Tsunami Detector Based Sort Message Service Using Arduino and SIM900A to GSM/GPRS Module, JEEMECS (Journal of Electrical Engineering, Mechatronic and Computer Science) Volume 1, No.1. 2018, DOI: doi/10.26905/jeemecs.v1i1.1982.

Adapting CRISP-DM for Idea Mining

A Data Mining Process for Generating Ideas Using a Textual Dataset

Workneh Y. Ayele
Stockholm University
Department of Computer and Systems Sciences, DSV
Stockholm University, Sweden

***Abstract*—Data mining project managers can benefit from using standard data mining process models. The benefits of using standard process models for data mining, such as the de facto and the most popular, Cross-Industry-Standard-Process model for Data Mining (CRISP-DM) are reduced cost and time. Also, standard models facilitate knowledge transfer, reuse of best practices, and minimize knowledge requirements. On the other hand, to unlock the potential of ever-growing textual data such as publications, patents, social media data, and documents of various forms, digital innovation is increasingly needed. Furthermore, the introduction of cutting-edge machine learning tools and techniques enable the elicitation of ideas. The processing of unstructured textual data to generate new and useful ideas is referred to as idea mining. Existing literature about idea mining merely overlooks the utilization of standard data mining process models. Therefore, the purpose of this paper is to propose a reusable model to generate ideas, CRISP-DM, for Idea Mining (CRISP-IM). The design and development of the CRISP-IM are done following the design science approach. The CRISP-IM facilitates idea generation, through the use of Dynamic Topic Modeling (DTM), unsupervised machine learning, and subsequent statistical analysis on a dataset of scholarly articles. The adapted CRISP-IM can be used to guide the process of identifying trends using scholarly literature datasets or temporally organized patent or any other textual dataset of any domain to elicit ideas. The ex-post evaluation of the CRISP-IM is left for future study.**

Keywords*—CRISP-IM; idea generation; idea evaluation; idea mining evaluation; dynamic topic modeling; CRISP-DM

I. Introduction

Under the umbrella of data science, the study of extracting value from data has advanced in complexity and size. Data science is more frequently used and is often favored over data mining these days [1]. According to [1], from a metaphorical discourse point of view, if data mining is analogous to gold mining, then data science is comparable with prospecting or searching for profitable mining sites. In this paper, for consistency reasons, data mining is used. Besides, the phrase “data mining” is used in the Cross-Industry-Standard-Process model for Data Mining (CRISP-DM) [2], which has been widely accepted for the past two decades [1]. Data mining is goal-orientated and more focused on processes, while data science is data-oriented and more focused on the exploration of values, including goal-driven values [1]. According to [1], the CRISP-DM process model is developed from goal-oriented perspectives, yet it is still applicable to data science projects.

The CRISP-DM is a generic data mining process model that provides an overview of life cycles of data mining projects [2]. CRISP-DM is popular both in the industry and academia [3], and according to user polls and many surveys, it is considered as the de facto standard for knowledge discovery and data mining projects [1]. The benefits of using the CRISP-DM are reduced cost and time, and minimized knowledge requirements for data mining projects. Moreover, expediting training, knowledge transfer, documentation, and capturing best practices are also the benefits of using CRISP-DM [4] (Chapman et al. 1999). More important, particularly for researchers and practitioners, data mining can be used in innovation endeavors [5]. Hence, CRISP-DM is useful for innovative activities.

Digital innovation is the process in which new or significantly changed artifacts are developed to be embodied in or enabled by IT [6]. In the twenty-first century, data is a goldmine with a potential for stimulating digital innovation. For example, scholarly literature databases are increasingly growing [7]. The accumulation of patent data could also be used to discover innovative solutions, new ideas to existing real-world problems [8]. Solving real-world problems could be supported by generating ideas and stimulating innovation through which economic development could be realized. Economic development arises from people generating ideas [7]. However, it is challenging to analyze large collections of data manually and hence demands innovative ways to deal with it. Another challenge, according to Stevens and Burley, is that initial ideas are seldom commercialized, and it takes thousands of new ideas for a single commercial success [9]. Also, [7] argues that it is becoming harder and harder to elicit innovative ideas. Therefore, it is valuable to have any means to generate as many ideas as possible.

Thorleuchter et al. defined a novel and possibly valuable idea as a text phrase comprising of domain-specific terms from the context of technical language usage rather than everyday language usage [10]. Moreover, a problem-solution pair could also be considered to be an idea [11]. In this paper, “idea” is referred to as: *“text phrase(s), sentence(s) describing new and useful information through expressing possible solution(s) to current problems”*. Ideas mining is used to generate ideas. Thorleuchter et al. define idea mining as “*the use of text mining to automatically process unstructured text data for extracting new and useful ideas”* [10, p.7183]. According to a number of authors, idea mining is introduced by Thorleuchter et al. [12 - 16]. Thorleuchter et al. applied the Euclidean distance measure, which is a distance-based similarity measuring algorithm, to generate ideas [10].

However, distance-based algorithms, which is used to measure the similarity between problem query and solutions from textual datasets, are not the only available solutions to idea mining. For example, deep learning [8], Information Retrieval (IR) [17], topic modeling [18], bibliometric [19], social network analysis [20], association rule mining [21], and collaborative filtering algorithms [11] could also be used for idea mining purposes. Therefore, in this paper, idea mining is described as "*the process of using text mining (e.g., topic modeling) in general, and more specific techniques such as machine learning techniques (e.g., deep learning, association rule mining, collaborative filtering, etc.), social and network theory (e.g., social network analysis), bibliometric, statistical methods, and IR to generate useful and new ideas from unstructured or semi-structured textual data.*"

In addition to distance-based algorithms and text mining techniques presented in the previous paragraphs, trend-based idea elicitation could be applied. Besides, trends relevant to user-demanded products are applicable to generate ideas [22]. Yet, finding ideas from untraceable weak trends that could yield a competitive advantage is like finding a needle in a haystack [23]. Fortunately, machine learning techniques enable the analysis of larger collections of datasets [24]. For example, topic modeling, a type of unsupervised machine learning, can be applied to identify hidden topics [25], and ideas [11]. Also, a topic modeling technique, such as Latent Semantic Analysis (LSA) is used to predict the relevance of ideas and provide valuable insights [26]. According to [27], the elicitation and analysis of trends can be helpful for decision-makers and stakeholders in academia and the industry [27]. Similarly, the elicitation of topics about emerging trends in science and technology is crucial for making decisions [28]. Trend analysis can be used for forecasting trends in technology [29]. Also, forecasting trends in technological advancement could be used to evaluate existing ideas collected or generated by incubators.

The interpretation of results from idea mining techniques needs human intervention. An exemplary scenario from [23] illustrates the relevance of human involvement in idea generation and evaluation. A business analyst of a "digital photography" company found an association between "digital photography" and "mobile phone" 20 years ago and ignored it. The business analyst could have convinced his company's CEOs to introduce a brand new technology, a mobile with an integrated camera. Mobile with an integrated camera is an idea that the business analyst should have proposed, and he most definitely regrets ignoring this idea today. Also, [30] argued that the creative process of generating and evaluating ideas should involve the use of machine learning and human interpretation. Besides, ideas should be evaluated for relevance using quality criteria perspectives such as technical, customer, market, financial, and social. Moreover, each quality criteria consist of concrete evaluation criteria, such as customer perspectives, including novelty, necessity, usability, and usefulness [31].

The use of standard process models for data mining projects contributes to the success of projects. Especially when experienced data scientists do the first projects, then projects will be efficiently, and reliability repeated [2]. Data mining projects rely on standard models for the success of involved people, especially with little time, limited options to experiment with different approaches, and lower technical skills [2]. Existing process models of idea mining mainly use a handful of techniques, namely similarity measure, association rules, and co-occurrence of terms analysis. Moreover, existing process models for idea mining are depicted as simple workflows and pictorial illustrations of processes. On the other hand, the CRISP-DM process model is independent of data mining technology and industry sectors [2]. As a result, it can be adapted for a given scenario. For example, Asamoah and Sharda applied design science research to adapt CRISP-DM for processing social media data to elicit analytical insights in healthcare [32]. Similarly, Spruit and Lytras aligned CRISP-DM's process and design science cycle by linking phases from the two methodologies [33]. CRISP-DM is widely accepted, and it is the de facto standard in data analytics applications [1, 33, 34].

The purpose of this paper is to introduce a reusable process model for idea generation based on CRISP-DM. This study aims to identify relevant phases and corresponding activities by designing and documenting a reusable process model for idea generation using Dynamic Topic Modeling (DTM) and subsequent processes. Therefore, in this study, an extension of a more general task and goal-oriented adaptation of the data mining process model based on CRISP-DM [2] for Idea Mining (CRISP-IM) is proposed. The idea mining processes and the detailed description of the activities could be adapted by documenting each phase followed under CRISP-DM. Yet, in this paper, the design science research approach is applied for the rigor of the study, and to address the "why" and the "how" questions for justifying the components of the proposed artifact. Design science is a relevant artifact development research approach used in information systems, computer science, and engineering research projects [35]. The adapted CRISP-IM model could be used for guiding idea generation in innovation projects. The design of the CRISP-IM is demonstrated using a running example of a data mining process, DTM, and subsequent statistical methods for eliciting ideas.

The dataset used, in this study, is a textual corpus of scholarly articles about self-driving cars. Self-driving technology was chosen since it is a hot research and development topic, attracting many stakeholders. For example, the use of a self-driving car has positive impacts on safety and quality of life, fuel reductions, optimum driving, and crash reductions. Also, with regards to urban planning, efficient parking is a characteristic of self-driving vehicles [36]. However, there are growing concerns utilizing self-driving technology such as ethical issues [37], transportation system, affordability, safety, control, and liabilities [36], and the impact of self-driving technologies on urban planning [37].

This paper is structured into six sections. The second section presents related research, and the subsequent sections include approaches used to adapt CRISP-DM, followed by a description and demonstration of the resulting CRISP-IM, a discussion of the result, and conclusions and future research.

II. Related Research

According to [1], CRISP-DM has been considered as the de facto standard data mining process model for the past two decades. Also, CRISP-DM is currently used in data science applications. For example, the Data Science Trajectory (DST) model proposed by [1], illustrated in Fig. 1, is backward compatible with CRISP-DM and includes new activities that are common in data science projects, namely the exploration of data sources, goals, products, data values, results, and narratives. Moreover, data acquisition, simulation, architecting, and release are data management related activities. Among the six exploration tasks, result exploration is the process of relating results to business goals. The activities illustrated in the DST model are proposed by [1] to serve as templates for planning data science projects where project managers could include or exclude activities in their workflows based on their demands. Hence, Fig. 1 and the discussions in this chapter illustrate the possibility of adapting CRISP-DM.

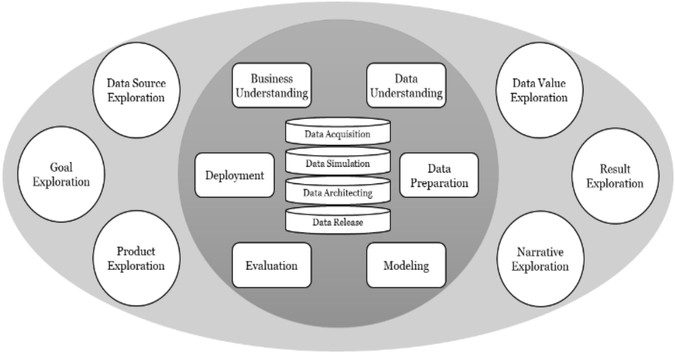


Fig. 1. The Data Science Trajectory (DTS) Map Containing Data Science Exploratory Activities in the Outer Circle, CRISP-DM (Goal-Directed Activities) in the Inner Circle, and Core Data Management Activities at the Center [1, p.5].

In this section, the adaptation of CRSIP-DM in different disciplines, and related research about process models for idea mining, are presented.

A. Previous Research - Adapting CRISP-DM

The CRISP-DM is adapted in different contexts and disciplines. The following list presents a list of adapted CRISP-DMs.

Bioinformatics

1) CRISP-DM is adapted in Bioinformatics for capturing the processes of computational and conventional biological processes of microarray DNA data analysis [38].

Software Engineering (Software Design)

2) Atzmueller and Roth-Berghofer extended CRISP-DM by adding explanation dimensions. The explanation dimensions identified are explanation goals (ontological knowledge), kind of explanations (instance knowledge), level of detail (pattern knowledge), and presentation styles (context knowledge) [39].

Cyber Forensic

3) Venter et al. adapted CRISP-DM in the cyber forensic domain for Event Mining (CRISP-EM) to define research gaps in evidence mining [40].

Data Mining Process Model

4) Martínez-Plumed et al. adapted CRISP-DM by designing a more flexible and Context-aware Standard Process for Data Mining – (CASP-DM), which inherits flexibility and versatility from the CRISP-DM life cycle and put more emphasis in that the sequence of phases is not rigid: context changes may affect different tasks so it should be possible to move to the appropriate phase [41].

Healthcare

5) Based on design science research Asamoah and Sharda adapted CRISP-DM for processing big and social network data to generate analytical insights in healthcare [32].

6) Catley et al. adapted CRISP-DM for Temporal Data Mining (CRISP-TDM) by incorporating a multi-dimensional time-series data of medical data streams [42].

7) Adapting CRISP-DM for addressing data mining problems in medicine (CRISP-MED-DM) [43].

8) Spruit and Lytras aligned the design science cycle and CRISP-DM's knowledge discovery process by linking phases from the two methodologies [33].

B. Previous Research about Idea Mining Models

An idea could be elicited from scholarly literature, patents, reports, the Internet, documents [44], networks of experts [45], social media [46], and crowdsourcing [47]. Furthermore, it is possible to elicit creative ideas of product and service from customers’ comments using forums [21].

Previous authors claim that idea mining is introduced by Thorleuchter et al. [10] [12-16]. The idea mining introduced by Thorleuchter et al. uses a text mining technique, a similarity measure using Euclidian distance, to elicit ideas [10]. Similarly, Alksher et al. used similarity measures for the same purpose [44]. However, other idea mining works use different analytics methods. For example, social network analysis [20], deep learning [8], bibliometric [19], topic modeling [18], and IR [17] are used for idea mining.

In spite of the existence of idea mining techniques, the use of standard processes models for idea mining is hardly touched. Existing process models are merely simple flow charts, Business Process Modeling Notations (BPMN) models, or simple nonstandard process diagrams.

1) Simple workflows or simple diagrams as process models: There are five research works presented below, which uses workflows or simple process diagrams to capture idea mining processes.

1) Thorleuchter et al. demonstrated that it is possible to generate ideas using unsupervised machine learning

technique, clustering, to generate ideas. Ideas are generated by measuring the similarity between textual data and problem query using distance metrics, Euclidean distance measures, and finally, applying idea mining measures using Jaccard's coefficient [10]. The process model is illustrated in Fig. 2.

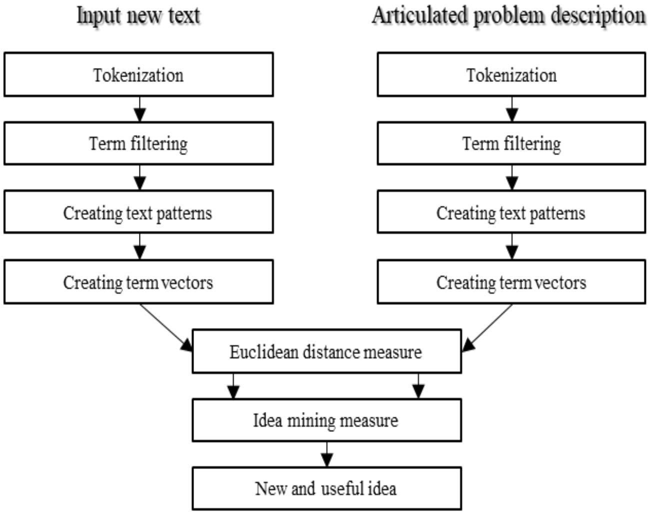


Fig. 2. Thorleuchter Et Al. Proposed a Process Model for Idea Generation using Similarity Measuring Techniques [10, p.7183].

2) It is possible to use the Apriori algorithm for association rule mining by exploring co-occurrence of terms using: key terms repository, associated terms repository, and suggestive terms repository about customer complaint data and online forum data to generate ideas as problem-solution or product-service as an output, the process model is illustrated in Fig. 3. [21].

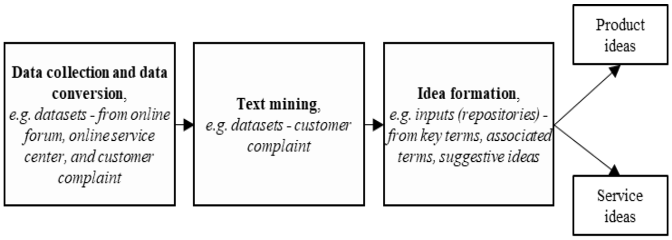

Data collection and data conversion,
e.g. datasets - from online forum, online service center, and customer complaint
Text mining,
e.g. datasets - customer complaint
Idea formation,
e.g. inputs (repositories) - from key terms, associated terms, suggestive ideas
Product ideas
Service ideas

Fig. 3. Kao Et Al. Process Model for Generating Product and Service Ideas [21, p.3].

3) Alksher et al. proposed an idea mining process and framework, as illustrated in Fig. 4. The idea mining uses similarity measures, Euclidean distance measure, to measure the similarity between a term vector of textual data and a new text provided by users [44].

4) Association rule mining and clustering using distance similarity measures are used to generate new ideas, and the process model is illustrated in Fig. 5 below [48].

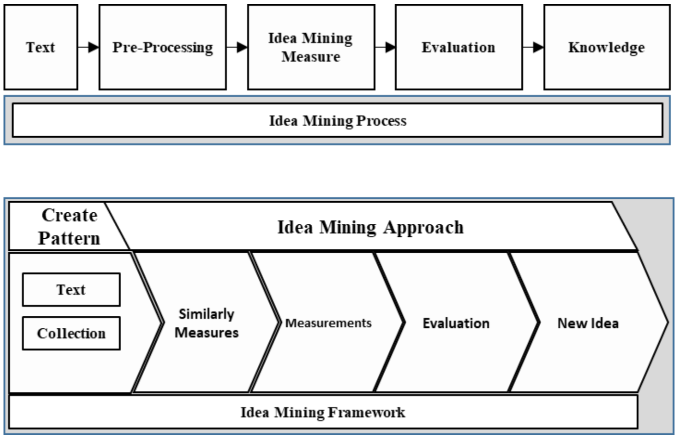


Fig. 4. Alksher et al. Idea Mining Process Model and Framework [44, p.89-90].

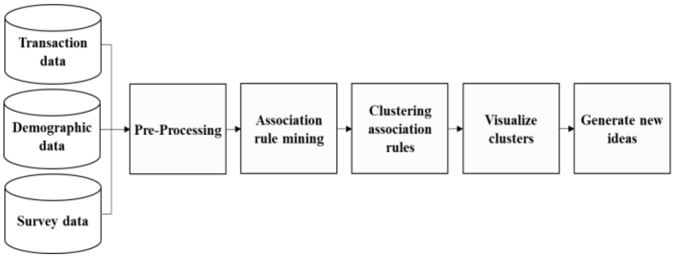


Fig. 5. Idea Mining Process Model using Association Mining [48, p.426].

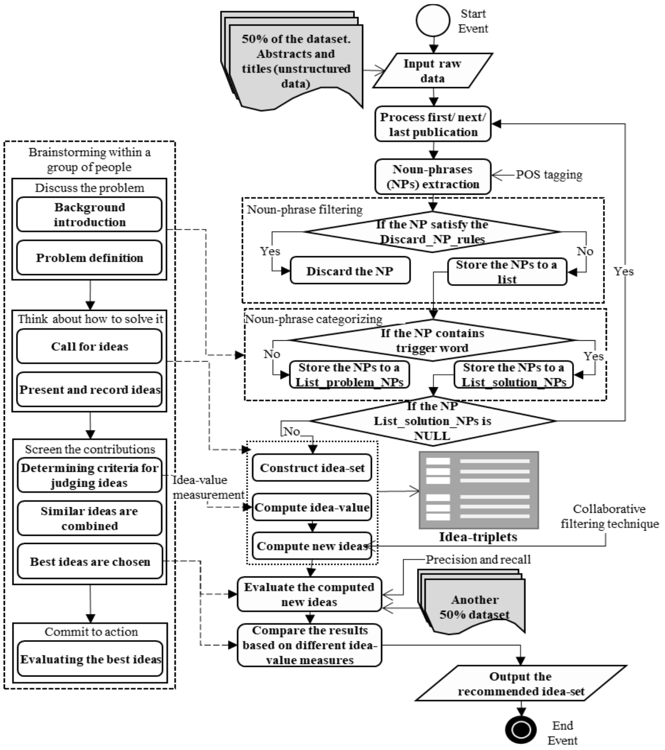

Start Event
50% of the dataset. Abstracts and titles (unstructured data)
Input raw data
Process first/ next/ last publication
Noun-phrases (NPs) extraction
POS tagging
Noun-phrase filtering
If the NP satisfy the Discard_NP_rules
Yes
No
Discard the NP
Store the NPs to a list
Yes
Noun-phrase categorizing
If the NP contains trigger word
No
Yes
Store the NPs to a List_problem_NPs
Store the NPs to a List_solution_NPs
If the NP List_solution_NPs is NULL
No
Brainstorming within a group of people
Discuss the problem
Background introduction
Problem definition
Think about how to solve it
Call for ideas
Present and record ideas
Screen the contributions
Determining criteria for judging ideas
Similar ideas are combined
Best ideas are chosen
Commit to action
Evaluating the best ideas
Idea-value measurement
Construct idea-set
Compute idea-value
Compute new ideas
Idea-triplets
Collaborative filtering technique
Precision and recall
Evaluate the computed new ideas
Another 50% dataset
Compare the results based on different idea-value measures
Output the recommended idea-set
End Event

Fig. 6. Liu et al. Idea Mining Process Model using POS and Collaborative Filtering [11, p.6].

5) A set of co-occurring terms, arranged as problem-solution pair is defined as ideas [11]. Liu et al. used Part-Of-Speech (POS) tagging to detect ideas as concepts, noun-phrases. Liu et al. used noun-phrases to classify as a set of problem-solution pairs and added relevance value to making it to triplets. The relevance value is calculated using the idea frequency-inverse document frequency score (phrase co-occurrence). Finally, collaborative filtering is applied on triplets to generate a ranked list of ideas. The process model is illustrated in Fig. 6.

2) Business Process Modeling Notations: There is one research work proposing a process model for idea mining presented below.

1) Idea mining process model for generating ideas is proposed using BPMN, see Fig. 7, where textual web data is processed using text mining using Part-Of-Speech tagging, and Natural Language Processing (NLP) such as stemming and lemmatization [46].

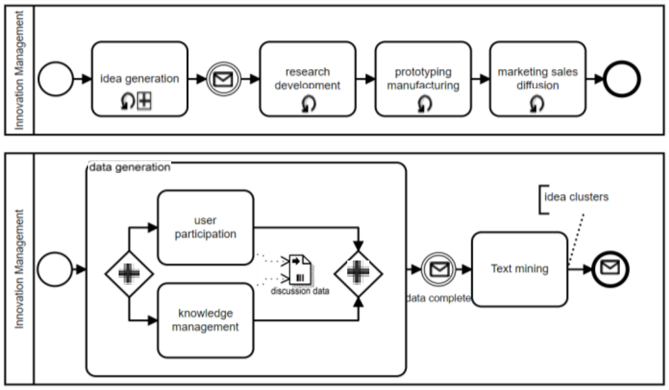


Fig. 7. Idea Mining Process Model Designed Following Business Process Modeling Notation (BPMN) [46, p.3-4].

III. Approaches Followed for Adapting CRISP-DM

It is possible to document the use cases, and activities followed while using CRISP-DM to introduce the CRISP-IM for facilitating reusability. Yet, the use of design science, which is an approach used for designing artifacts in computer science, information systems, and related disciplines [35], motivates and justifies as to why the components of the proposed CRISP-IM are relevant. The use of design science also adds value to the rigor of the research. The CRISP-DM for Idea Mining (CRISP-IM) method is adapted after running a DTM that is based on the (Latent Dirichlet Allocation) LDA algorithm. The result of the DTM was also further used to identify predictors and forecast trends using visualization of time-series patterns, and prediction. The development of the artifact, CRISP-IM, was done by following the design science approach, as illustrated in [49].

A. Dynamic Topic Modeling and Subsequent Analysis

Dynamic Topic Modeling (DTM): Large collections of unstructured textual datasets demand machine learning techniques for analyzing and gaining interpretable insights [24]. The DTM model by [50] is based on LDA. LDA is a topic modeling technique designed to elicit hidden topics without temporal information [25]. However, the DTM proposed by Blei and Lafferty could be used to identify topics and their evolutions. Also, to represent topics using multinomial distributions, it uses a state-space-model. Furthermore, it uses Kalman filters vibrational approximation and non-parametric wavelet regression to infer latent topic approximations [50]. In this research, DTM was used to elicit the evolution of topics about self-driving cars. The implementation of the DTM used was written in Python, and the subsequent analysis was done using Excel and RStudio.

Idea generation and evaluation: The interpretation of the result needs human intervention. For example, the elicitation of ideas is done by examining patterns, trends, and foresight generated from the DTM and the subsequent statistical processing, such as correlation and time series analysis using regression. The use of terms association analysis can be used to make strategic decisions to generate novel product idea development [23]. Topic modeling techniques, such as LDA group co-occurring terms together in respective topics [25]. Elicited topics are labeled with descriptive names, as suggested and illustrated by [25, 51]. Besides, the creative process of generating and evaluating ideas should involve experts and machine learning [30].

B. Design Science Approach: Method for Designing CRISP-IM

The design science approach, followed in this paper, consists of six activities: the identification of problems, objectives of the solution, design & development, demonstration, evaluation, and communication [49].

1) Identification of the problem: Method for problem identification is carried out using a literature review. Existing research on idea generation and evaluation, in particular, the use of DTM and succeeding statistical analysis is hard to find. Furthermore, previous work regarding idea mining, in particular, idea mining through the use of DTMs and succeeding statistical analysis, overlooks the utilization of standard models, such as CRISP-DM.

2) Objectives of the solution: The objectives of the artifacts are:

- To be able to follow the guidelines of CRISP-IM to preprocess textual data and identify the best models for topic modeling.
- To be able to follow the guidelines of CRISP-IM to identify topics and their trends.
- To be able to follow the guidelines of CRISP-IM to evaluate the quality of topics identified in terms of interpretability for idea generation and to generate ideas.

3) Design and development: Each phase of the CRISP-IM is customized by mapping it with the DTM process carried out to identify emerging trends in [52]. Additionally, the key elements of the phases of the CRISP-IM are inspired by the CRISP-DM [2] model. The detailed specification of components is identified using the DTM, and succeeding statistical analysis processes followed.

The design and development of the CRISP-IM mainly focus on documenting and facilitating the reusability of the innovative process, idea generation. The introduced CRISP-IM guides data analysts to identify trends and generate insightful outputs to generate and evaluate ideas.

4) Demonstration: The running example is based on the DTM and the succeeding statistical analysis to demonstrate and motivate parts of the CRISP-IM. Also, the components of each phase of CRISP-IM are motivated from previous research work that involves data collection, preprocessing, selecting the best machine learning model, generating topics, identifying trends, and interpreting the work as presented in [52].

5) Evaluation and communication: An empirical evaluation of the CRISP-IM following the design science research approach is left for future study. This paper is written to communicate the result to academia and the industry.

IV. Result: Adapting CRISP-DM to CRISP-IM

In this section, the adapted CRISP-DM, CRISP-IM, is presented. CRISP-DM to CRISP-IM mapping is illustrated in Table I and Fig. 8. The phases of CRISP-IM are renamed with more relevant and descriptive names, see Table I. Similarly, Venter et al. labeled each phase to suit their process in cyber forensic [40].

A. Phase 1: Technology Need Assessment

In this phase, a business needs assessment is done through elicitation of business opportunities and challenges being addressed. The inputs in this phase are needs from within a company, and reports from previous idea mining. Also, the identification of goals and success criteria, resources, cost-benefit analysis, risks, and contingencies are carried out.

1) Motivation for including Phase 1
Why is this phase needed?

Business need assessment is a critical task in data mining projects. Data analytics is used to unlock the hidden potential of large datasets, referred here as idea mining, to produce insights. Despite ever-growing research activities and the resulting findings, it has become difficult to find innovative ideas [7]. The inputs to business need assessment are needs within the organization, goals, success criteria, resources, cost-benefit analysis, and risk and contingencies [2].

How can the activities in this phase be done?

Business need assessment can be done through the use of corporate foresight [23] and requirement elicitation. Topic modeling also enables the identification of valuable insights [26]. Forecasting improves creative performance as part of idea generation [53]. Moreover, it is possible to use scientometric, machine learning, and visual analytics to elicit trends and temporal patterns [54].

TABLE I. Adapting CRISP-DM for Idea Mining, CRISP-IM

CRISP-DM	CRISP-IM
Business Understanding	Technology Need Assessment
Data Understanding	Data Collection and Understanding
Data Preparation	Data Preparation
Modeling	Modeling for Idea Extraction
Evaluation	Evaluation and Idea Extraction
Deployment	Reporting Innovative Ideas

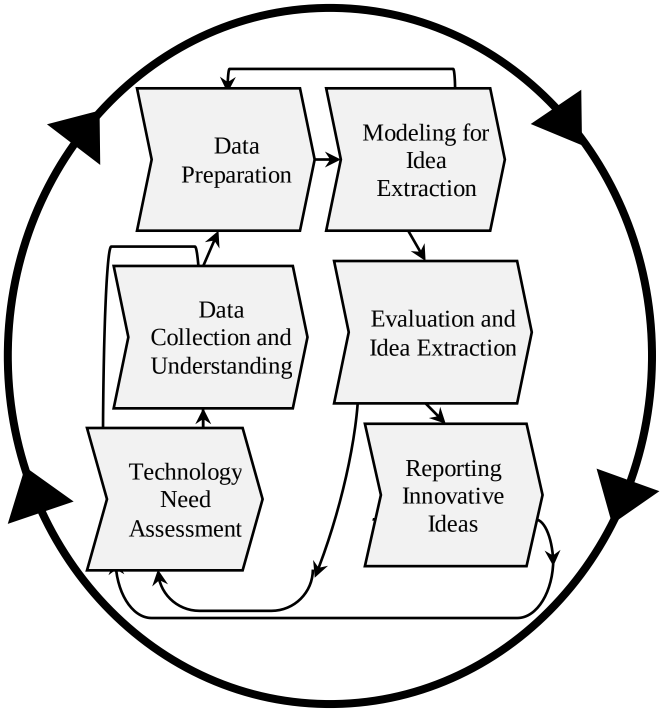

Data Preparation
Modeling for Idea Extraction
Data Collection and Understanding
Evaluation and Idea Extraction
Technology Need Assessment
Reporting Innovative Ideas

Fig. 8. CRISP-DM for Idea Mining (CRISP-IM).

2) Demonstration: The goal of the running example [52] is to identify emerging trends and patterns to elicit innovative ideas. Hence, [52] identified what could be gained by understanding the project.

The goals of the data mining by [52] are:

- To unlock the potential of ever-growing research findings in academia
- To identify the emerging trends about self-driving cars in academia
- To generate and evaluate ideas for innovation in self-driving cars using scholarly literature as a data source
- To support idea evaluation and generation activities by generating insightful trends

The data source chosen was Scopus, and tools and computing resources used were:

- Python, Excel, a PC with 8GB RAM, a 64 bit Windows 10 Operating System, and an Intel i7 CPU with 2.7 GHz

The analysis of cost-benefit, risk, and contingencies was not done as the purpose of the work by [52] is academic research.

B. Phase 2: Data Collection and Understanding

In this phase, data is collected after articulating a search query to extract relevant datasets. Data cleaning activity in this phase includes reformatting data into a structured dataset by removing anomalies. Also, exploring and describing data, identification of missing values, removing redundant information, and checking data consistency are also carried out in this phase. Phase 2 could also lead back to Phase 1 – Technology Need Assessment when the data quality is not good enough to extract valid information. For example, when there is insufficient data, we can go back to Phase 1 and include other data sources [2].

1) Motivation for including Phase 2
Why is this phase needed?

The most valid activities in this phase are collecting and cleaning the initial data, exploring and describing the data, and finally verifying data quality and adequacy. If the data is not of good quality and is not adequate, go back to Phase 1 to identify other data sources or repositories [2].

How can the activities in this phase be done?

Data can be extracted from the chosen data sources, such as Scopus. Scopus has better coverage of journals [55] and contains the latest and larger datasets of scholarly literature than Web of Science [56]. Exploring and understanding data can be carried out using a spreadsheet application [37, 57].

2) Demonstration: Scopus was used as a data source for collecting the dataset. Therefore, a query was formulated to extract the data. A total of 5425 documents were downloaded in CSV format. The documents were retrieved in batches since Scopus limits the maximum downloadable documents to 2000. Removal of duplicates could also be done using reference management software such as Mendeley [58], see Fig. 9. Data understanding was done using Excel and Notepad. A preliminary scanning of the dataset using Excel and Notepad was done to verify data quality. A initial dataset was prepared using Mendeley and Zotero.

C. Phase 3: Data Preparation - Preprocessing

In this phase, the inclusion of relevant data, cleaning of data, generating derived attributes, merging, and formatting data to make it suitable for modeling are performed.

1) Motivation for including Phase 3
Why is this phase needed?

The goal of this phase is to prepare data for modeling. Therefore, the activities performed in this phase address quality issues, identify and apply format for modeling, and clean the data as suggested by [2].

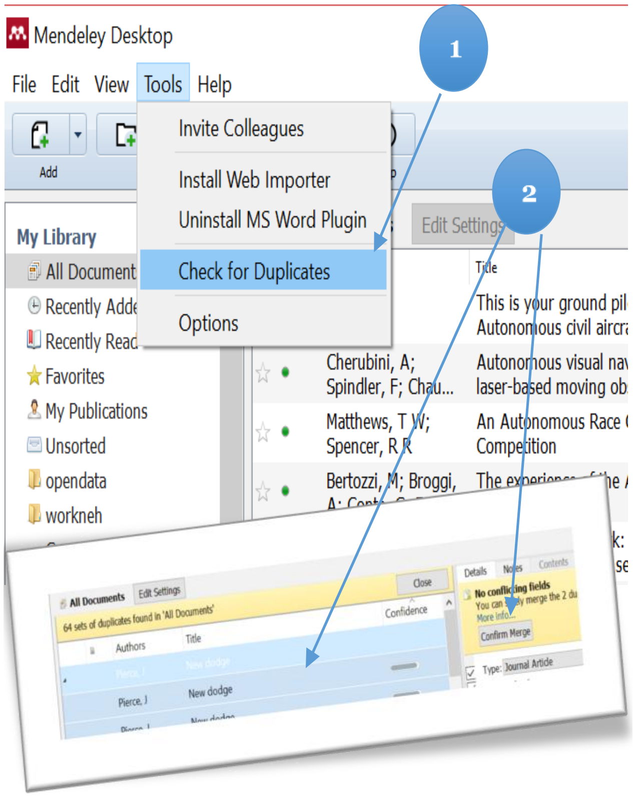


Fig. 9. Screenshot Illustrating how Duplicates are Managed using Mendeley.

How can the activities in this phase be done?

Effective text mining processes are predicated on suitable and relevant preprocessing techniques, and preprocessing techniques are applied to unstructured data to generate structured data suitable for text mining models [59].

2) Demonstration: The dataset was preprocessed using Python. The data cleaning activities listed above were done. Preprocessing was aided by word clouds, visualization of term frequency, and list of terms with corresponding frequencies to update the stopword list, as illustrated in Fig. 10. Also, noisy and irrelevant terms such as “IEEE”, “Copyright”, etcetra were removed. After the tokenization of each document, the lemmatization of terms was done to convert terms to their root forms. Instead of lemmatization, stemming could also be done, as illustrated in [37]. After tokenization, bigrams were computed and added to the tokens list. Finally, the main inputs for the DTM, corpus (bag of words representation), and dictionary were generated. The dictionary representation of the documents was readjusted to contain tokens that are available in more than 100 documents or less than 95% of the documents [52] to minimize the feature dimensionality problem and to include most productive and interpretive data. The combination possibilities and the scale of feature values in text mining are usually greater than standard data mining systems [59].

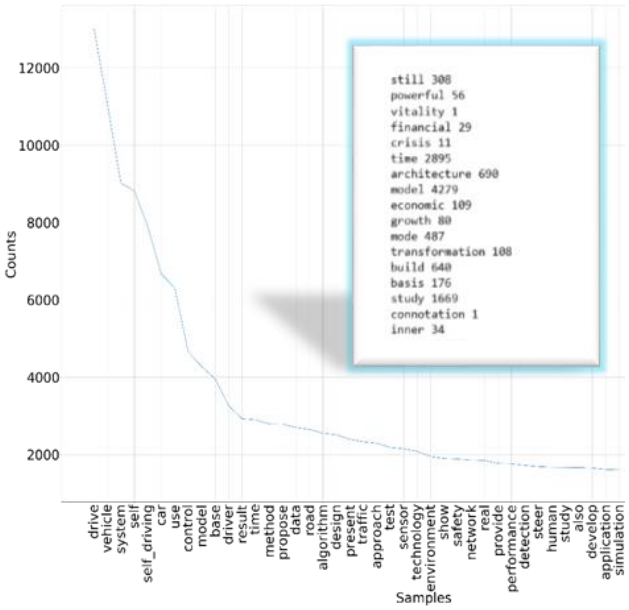


Fig. 10. Word Frequency Visualization and List of Terms with Corresponding Frequencies to Identify Irrelevantly and Misspelled Terms.

D. *Phase 4: Modeling for Idea Extraction*

In this phase, it is possible to do data analysis using tools and techniques such as text mining, network analysis, statistical analysis of linguistic features, etcetera. In this study, the chosen technique is DTM and the succeeding statistical analysis. If the model generated for the DTM is not of good quality, then it is possible to go back to data preparation, Phase 3, if data need to be reformatted to fit specific requirements by the chosen technique.

1) Motivation for including Phase 4
Why is this phase needed?

In this phase, the most important task is selecting the best model after identifying the data mining modeling technique, building, and assessing it [2]. The quality of the result depends on the quality of the model used.

How can the activities in this phase be done?

The inputs to the model generation process are the number of topics, preprocessed text corpus, and dictionary. Perplexity measures are used to determine the number of topics [37] or coherence scores [60]. The best fit model could be generated by comparing the semantic interpretability of models [62].

2) Demonstration: The DTM model, which is based on LDA [50], was used to generate topics and their evolution using a dataset about self-driving cars [52]. The DTM proposed by [50], analyzes the evolution of topics in a specific set of chosen datasets. The inputs, corpus, dictionary, and the number of topics are compulsory inputs for running the Python implementation, ldaseqmodel library[1], of the DTM algorithm, which was used in this research. In this phase, models suitable for idea extraction from topic evolution are selected.

1) Selecting the best model for idea extraction. In this paper, DTM based on LDA is chosen. However, DTM topic modeling techniques based on Latent Semantic Analysis (LSA) or Non-negative Matrix Factorization (NMF) could also be chosen based on the generated quality of output, as illustrated in phase 5 by comparing coherence score of these models. LDA-based topic modeling is favored over other topic models such as NMF, and NMF is overlooked [62]. O'callaghan et al. used a coherence score, a technique to measure semantic interpretability of topics, to compare models [62].

- Determine the optimum number of topics - it is possible to determine an optimum number of topics through perplexity [37] or coherence scores [60].

2) Generating the DTM model

3) Generating the output – topics and their evolution identified through visualization and interpretation of results, including succeeding statistical analysis.

- After a preliminary review of terms under each topic, it is possible to find inconsistencies such as the presence of abbreviations or acronyms. Also, the quality of the preprocessed data determines the quality of the model and the result. The optimum number of topics identified could also be affected by the quality of the input corpus. As a result, it is also possible to find overlapping topics despite your efforts in preprocessing and determining the optimum topic number. So you might consider going back to Phase 3 to preprocess accordingly. The quality of the output could also be affected by the NLP, preprocessing, the strategy you are following. For example, if you are choosing either stemming or lemmatization and get poor quality, you can go back to Phase 3, then do a separate preprocessing to compare both strategies.

E. *Phase 5: Evaluation and Idea Extraction*

1) The result of Phase 4 is evaluated against the goals of DM listed in Phase 1, and if the result is not in line with the goals of the project, then we need to go back to Phase 1 and redo the whole process again to meet specified goals. The most relevant activities in this phase are listed below.

1) Assessment of model performance

2) Labeling of topics

3) Identifying trends and illustrations using visualizations

4) Prediction of trends through time series analysis when enough time series data is available

5) Run correlation test on candidate terms in the time series to elicit correlations

6) Elicit ideas and align assessment of data mining results with goals and success criteria specified in Phase 1.

- Analyze the result for idea elicitation and evaluation using business analysts and technical experts. In addition, statistical indicators of significance are also

[1] https://radimrehurek.com/gensim/models/ldaseqmodel.html

used for making Go or No Go decision based on the quality of the result. Furthermore, candidate ideas generated as an output should be evaluated using quality criteria.

- Based on the analysis of the result, it should be possible to make decisions to determine if the result obtained is of acceptable quality and proceed to the next phase by analysts involved. If analysts are not convinced then we need to determine what our next steps are, for example, make an assessment if you need to re-clean the data, by asking questions such as "*Do we need to go back to Phase 1*?", where we repeat the whole processes again from Phase 1- Technology Need Assessment. In Phase 1, you will need to rearticulate goals and success criteria and then continue to Data Collection and Understanding.

2) Motivation for including Phase 5
Why is this phase needed?

Before generating reports or deploying generated models, it is important to evaluate models [2]. Besides, idea extraction is the main objective of the CRISP-IM model.

How can the activities in this phase be done?

Assessing the model – Calculation of the coherence score, which measures the quality of topics in terms of coherence of terms within topics, is used for assessing the quality of the model [63].

Labeling of topics – The naming of generated topics with descriptive names is not done by machines but by individuals involved in the activity [25, 51].

Identifying trends and illustrations using visualizations – Rohrbeck suggested the use of a mixture of people-centric and bibliometric mechanisms to identify weak signals that have potential [23], so it is possible not to miss every possible indicator of trends. Furthermore, text mining techniques are also used for the elicitation of foresight to predict and anticipate the market future, from weak signals [61].

Predicting trends through time series analysis when enough time series data is available – Insights and foresight generated using text analytics, topic modeling, can be used to elicit ideas [26]. Similarly, it is possible to elicit trends and identify temporal patterns using machine learning, scientometric, and visual analytics [54]. Furthermore, when there is enough observation to run a time series prediction, then it is possible to run forecasting of trends for a chosen topic. For example, according to [64], if you have observations of at least four, it is arguably advisable to use regression. Besides, [65] suggest that at least 50 observations are required to use advanced models such as Autoregressive Integrated Moving Average (ARIMA) model for time-series predictions.

Running correlation tests on candidate terms – Correlation is used to identify relationships or associations between variables [66]. Rohrbeck suggested the identification of association is valuable to identify potential product ideas [23]. Statistical significance tests should be used to determine the quality of post-processing using correlations, regressions, and time-series analysis results. Moreover, weak signals of change in the evolution of trends of topics should not be ignored.

Eliciting ideas and align assessment of data mining results with goals and success criteria specified in Phase 1 – Elicitation of ideas can be done by using topic modeling [26] and using association of terms found in publications and patents [23]. Evolving trends identified from topics can be as inputs for decision making in research and real-world technology [28]. Analysis of trends results in forecasting trends in technology [29], forecasting while idea generation improves idea evaluation [53].

Idea generation and evaluation activities should include the use of criteria such as customer perspectives – with attributes such as necessity, novelty, usefulness, usability, and other perspectives such as technical, market, financial, and social [31]. In addition to expert judgment to evaluate the quality and acceptability of the result, it is possible to run statistical significance tests on predictions [52].

3) Demonstration: Before labeling topics with descriptive names, an overall assessment of the coherence of terms under each topic was done. The calculation of the coherence score was done based on the algorithm by [63], and the implementation was based on the genism[2] python library. Several numbers of topics were assigned to models to choose a model with the highest coherence score, and the coherence score was generated for each number of topics, as illustrated in Fig. 11.

Labeling of topics was done by assigning descriptive names to topics identified, as suggested in [25] and [51]. Besides, thematic analysis of terms under each topic was done to label topics following [67]. However, when acronyms and abbreviations are present, labeling of topics is difficult [68]. Therefore, acronyms and abbreviations were interpreted after identifying their interpretations accordingly. The elicitation of trends was carried out using graphical illustrations of the result, see Fig. 12, for illustration. The visualization of term probabilities throughout the timeline enables the explanation and interpretation of trends [69], see Fig. 12.

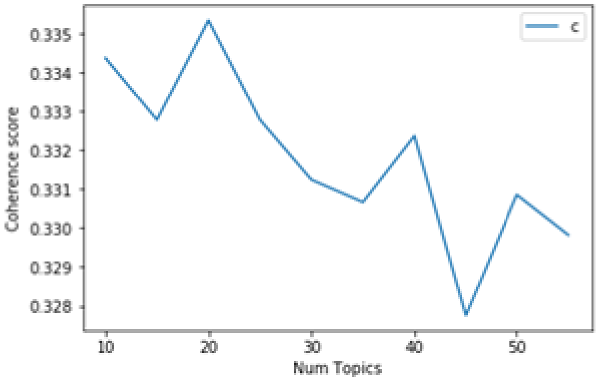


Fig. 11. Topic Coherence Score.

[2] https://radimrehurek.com/gensim/models/coherencemodel.html

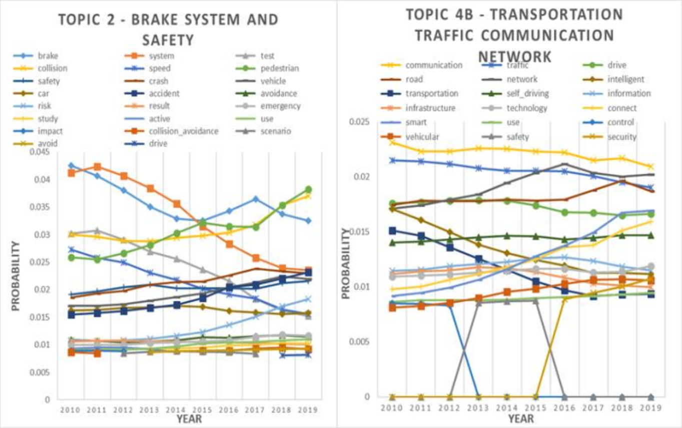
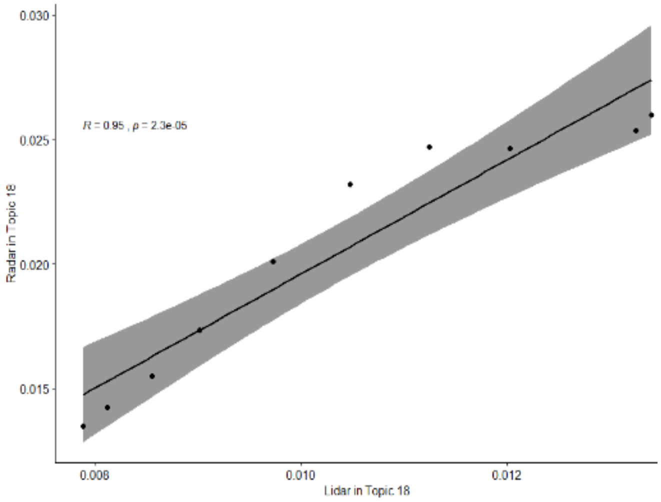


Fig. 12. Illustration of Topic Evolution.

The term "Smart," for example, has an increasing trend as illustrated in Topic 4B, Fig. 12, indicating that there is an increasing interest in academia in the concept smart in relation to terms such as road, traffic, communication, and so on. This trend implies issues regarding smart cities are also gaining attention in academia. On the other hand, Topic 2 shows that pedestrian, collision, accident, and break have increasing trends; also, safety is trending, implying that there is an increased demand for control and safety of self-driving cars. Innovative ideas related to these trends are likely to have a higher relevance. Finally, if there is enough observation of time series data, then it is possible to run a time series analysis for forecasting and to extract valuable insights.

Thematic analysis and visualization were used to elicit and interpret trends. Identified trends can be used to extrapolate and generate ideas and reports. Generated ideas can be used for evaluating the relevance and timeliness of ideas being commercialized by incubators, innovators, and R&Ds. The thematic analysis of evolving topics was done to analyze, elicit trends, and report patterns and themes [67]. It is possible to identify the correlation of terms using scatterplots of a chosen topic, as presented in [52], and it is illustrated in Fig. 13. Also, it is possible to generate innovative product ideas from associated terms [23].



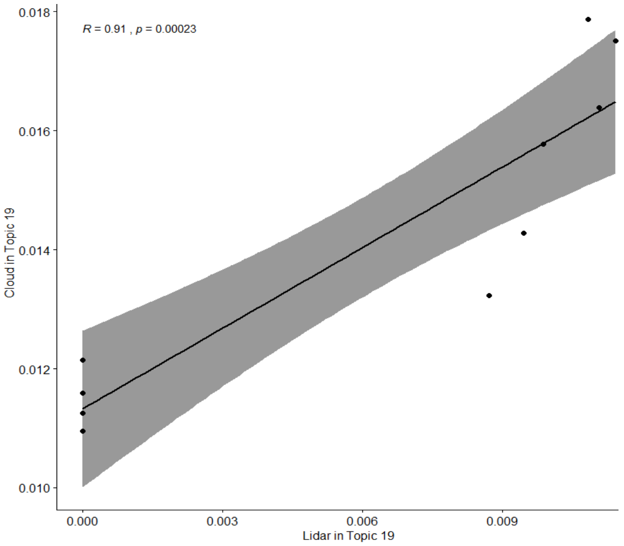


Fig. 13. Illustration of Correlation between Lidar and Radar on the Top and Lidar and Cloud on the Bottom.

In the demonstration [52], there are ten observations, which is a ten years' period. Since the number of observations is ten years' period, it is not advisable to run advanced time series analysis like ARIMA model [65], hence simple regression analysis for time series prediction, as suggested by [64], was done. For the illustration of prediction using regression, see Fig. 14.

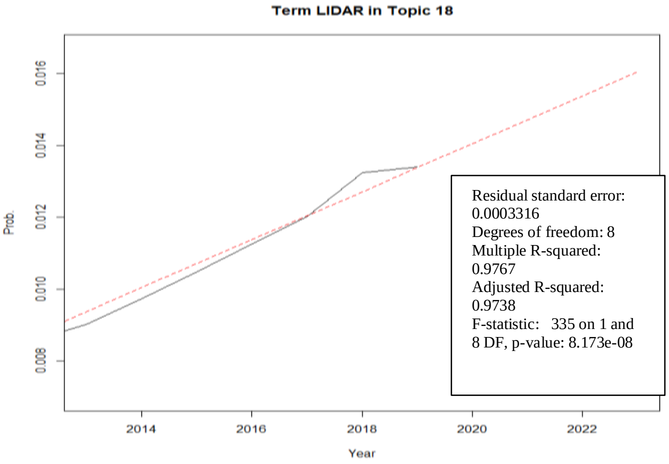


Fig. 14. Illustration of Forecasting Trends, the Choice of the Time-Series Model could be Evaluated using Statstical Significanc Measures such as Residual Standard Error, and P-Values.

Finally, identified innovative lists of ideas and research agendas were documented [52]

F. Phase 6: Reporting Innovative Ideas

In this last phase, analysis and interpretation of generated ideas and information from the result are reported. The main tasks suggested in this phase are reporting results, documenting best practices, and deployment planning. If the model generated in Phase 5 is of high quality, and if the task, idea generation, is done frequently, then it could be integrated with existing applications, and finally, maintenance and monitoring could be done if the model is integrated with existing applications.

1) Motivatioan for including Phase 6
Why is this phase needed?

The model alone is not the end product of data mining projects, and hence knowledge elicited need to be articulated and presented in a reusable way so that it can be used by customers for the future. Planning deployment, maintenance, and monitoring, producing, and reviewing final reports are activities in the last phase suggested in CRISP-DM [2].

How can the activities in this phase be done?

In this phase, generating reports is done and documenting actions that need to be done to make use of the created models so that it can be reused [2]. The documentation of innovative ideas and best practices can be done using project management tools and word processing applications.

2) Demonstration: The reporting of the results is done through textual interpretation and explanation through the visualization of topic evolution. Additionally, lessons learned, implementation codes, use cases, insights leading to quest further, and best practices are documented for future analysis. Finally, ideas for research and commercialization are documented and reported [52].

V. Discussions

This paper aims to build a set of guidelines for processing scholarly articles to generate and evaluate ideas using machine learning. In addition to machine learning, idea generation can be done using a network of experts [45], social media and online forums [21, 46], crowdsourcing [47], and innovation contests [70-71]. Ideas generated through machine learning can be used as an input in the front end of innovation activities and could ultimately be commercialized. For example, according to [72], an idea "gluten-free-beer" generated from a machine learning extracted from online forums was used by Lakefront Brewery Inc. to introduce the first gluten-free beer. Thus, idea generation activities are beneficial to companies. Idea mining unlocks potential marketing possibilities for organizations. On the other hand, the activities of idea generation, idea mining, can benefit from and following standard process models. Typically, CRISP-DM facilitates reusability, learning, knowledge transfers, cost and time reduction, and documentation [2].

In this paper, CRISP-DM is adapted to capture the processes of DTM for idea mining process, CRISP-IM, by using a dataset of self-driving cars. Similarly, CRISP-DM is adapted in different research domains such as cyber forensic for Event Mining (EM) as CRISP-EM [40], healthcare for processing big social network data [32], and other contexts as illustrated in Sections 2 and 3. Also, standard process models enable people with lower technical skills to easily follow and carry out sophisticated data mining tasks [2]. Likewise, CRISP-IM facilitates the reuse of best practices and expedites successful idea mining tasks. The result of this study could be used as a guideline for processing scientific literature, patents, and any textual information with a temporal variable, for evaluating and eliciting ideas through the analysis of trends.

CRISP-DM does not deal with tasks related to project organization, management, and quality issues [3]. However, the adapted CRISP-IM is designed to facilitate reusability with minimal knowledge requirements, documenting best practices, and research contribution purposes. Therefore, the proposed model does not focus on project planning and organization, similar to other adapted models, for example [40].

VI. Conclusions And Future Research

The growth of research findings has become exponential, and user-generated textual data is growing at an unprecedented pace while it is hard to find innovative ideas. It is possible to unlock the potential of digital data sources such as patents, social media, crowdsources, and etcetera by applying machine learning. For example, it is possible to identify research agendas and innovative ideas by analyzing research findings. The primary purpose of this research is to facilitate idea generation by simplifying idea mining and unlocking the untapped potential of growing unstructured or semi-structured textual data. To facilitate reusability, knowledge management, ease of learning, are the benefits of using standard process models. The proposed CRISP-IM adds values to the front-end of the innovation processes by streamlining the elicitation of innovative ideas.

Idea mining in this paper was used to demonstrate the process of conducting exploratory data analytics, and machine learning (unsupervised learning) on scholarly articles to elicit new ideas. It is also possible to generate ideas using a variety of other techniques. For example, to generate ideas, the following algorithms and methods could be used: Euclidean distance measure, deep learning, IR, topic modeling, bibliometric, social network analysis, association rule mining, statistical analysis, and collaborative filtering algorithms. Idea mining could be conducted to extract valuable and new information from mainly unstructured textual data. Text mining is the main technique used in idea mining activities, which involves IR and most data mining tasks such as association rule mining, similarity measuring techniques, and topic modeling. Text mining a field of study within computer science that uses techniques of data mining, IR, machine learning, NLP, and knowledge management [59]. Therefore, it is possible to undertake idea mining through the use of techniques of text mining, social network analysis, and bibliometric. Therefore, idea mining activities involve text mining, bibliometrics, statistics, and social network analysis.

Future studies should address the limitation of this study, ex-post evaluation, and extend this study by including other techniques and data sources for idea mining. Future studies could evaluate the applicability of CRISP-IM in different contexts. Finally, it is suggested to extend the CRISP-IM to include other types of DTMs such as Dynamic LSI, and Dynamic NMF.

Acknowledgment

I wish to thank Alexey Voronov (Ph.D.) and Mahedre DW Amanuel (M.Sc.) at RISE for their support in their inputs at the beginning of the project. I would also like to thank Gustaf

Juell-Skielse (Associate Professor), Paul Johannesson (Professor), Panagiotis Papapetrou (Professor) for offering constructive feedback. This study was conducted as part of the research project IQUAL (2018-04331) funded by Sweden's Innovation Agency (Vinnova).

REFERENCES

[1] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. H. Orallo, M. Kull, N. Lachiche, and P. A. Flach, CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories, IEEE Transactions on Knowledge and Data Engineering, 2019.

[2] R. Wirth, and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," Citeseer in Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, pp. 29-39, April 2000.

[3] G. Mariscal, O. Marban, and C. Fernandez, A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review, Vol. 25(2), 2010, pp. 137-166.

[4] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "The CRISP-DM user guide," in 4th CRISP-DM SIG Workshop in Brussels, March 1999.

[5] D. Kong, Y. Zhou, Y. Liu, and L. Xue, L. Using the data mining method to assess the innovation gap: A case of industrial robotics in a catching-up country, Technological Forecasting and Social Change, 119, 2017, pp. 80-97.

[6] R. Fichman, R. G., Dos Santos, B. L., & Zhiqiang (Eric) Zheng. Digital innovation as a fundamental and powerful concept in the information Systems curriculum. Mis Quarterly, Vol, 38(2), 2014, pp. 329-343.

[7] N. Bloom, C. I. Jones, J. Van Reenen, and M. Webb, Are ideas getting harder to find? (No. w23782), National Bureau of Economic Research, 2017.

[8] T. Hope, J. Chan, A. Kittur, and D. Shahaf, "Accelerating innovation through analogy mining," in 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD, Part F1296, 2017, pp. 235–243.

[9] G. A. Stevens, and J. Burley, 3,000 raw ideas= 1 commercial success!. Research-Technology Management, Vol. 40(3), 1997, pp. 16-27.

[10] D. Thorleuchter, D. Van den Poel, and A. Prinzie, Mining ideas from textual information. Expert Systems with Applications, 37(10), 2010, pp. 7182-7188.

[11] H. Liu, J. Goulding, and T. Brailsford, "Towards computation of novel ideas from corpora of scientific text," Springer, Cham in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2015, pp. 541-556.

[12] E. K. Özyirmidokuz, and M. H. Özyirmidokuz, "Analyzing customer complaints: A web text mining application," in Proceedings of INTCESS14-International Conference on Education and Social Sciences, Istanbul, 2014, pp. 734-743.

[13] E. A. Stoica, and E. K. Özyirmidokuz, Mining customer feedback documents. International Journal of Knowledge Engineering, Vol. 1(1), 2015, 68-71.

[14] T. C. Dinh, H. Bae, J. Park, J. Bae, (2015). A framework to discover potential ideas of new product development from crowdsourcing application. arXiv preprint arXiv:1502.07015.

[15] M. Alksher, A. Azman, R. Yaakob, E. M. Alshari, A. K., Rabiah, and A. Mohamed, "Effective idea mining technique based on modeling lexical semantic," Journal of Theoretical and Applied Information Technology, Vol. 96(16), 2018, pp. 5350-5362.

[16] A. Azman, M. Alksher, S. Doraisamy, R. Yaakob, and E. Alshari, "A Framework for Automatic Analysis of Essays Based on Idea Mining," Springer, Singapore, in Computational Science and Technology, 2020, pp. 639-648)..

[17] J. Chan, J. C. Chang, T. Hope, D. Shahaf, and A. Kittur, A. "SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers," Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 31:1–31:21, 2018.

[18] W. S. Lee, and S. Y. Sohn, Discovering emerging business ideas based on crowdfunded software projects. Decision Support Systems, 116, 2019, pp. 102–113.

[19] T. Ogawa, and Y. Kajikawa, "Generating novel research ideas using computational intelligence: A case study involving fuel cells and ammonia synthesis," in: Technological Forecasting and Social Change, 120, 2017, pp.41–47.

[20] P. Chen, S. Li, and M. Hung, "Co-occurrence analysis in innovation management: Data processing of an online brainstorming platform," Proceedings of PICMET '13: Technology Management in the IT-Driven Services (PICMET), 2013, pp. 688–694.

[21] S. C. Kao, C. H. Wu, and S. W. Syu, A creative idea exploration model: Based on customer complaints, 5th MISNC 2018, 2018.

[22] A. Salovaara, and P. Mannonen, "Use of future-oriented information in user-centered product concept ideation," Springer, Berlin, Heidelberg in IFIP Conference on Human-Computer Interaction,. pp. 727-740, September 2005.

[23] R. Rohrbeck, (2014). Trend scanning, scouting and foresight techniques. Springer, Cham in Management of the Fuzzy Front End of Innovation, 2014, pp. 59-73..

[24] D. M. Blei, and J. D. M. Lafferty, Topic models. In Text Mining, Chapman and Hall/CRC, 2009, pp. 101-124.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, 3, pp. 993-1022, January 2003.

[26] B. Steingrimsson, S. Yi, R. Jones, M. Kisialiou, K. Yi, and Z. Rose, (). Big Data Analytics for Improving Fidelity of Engineering Design Decisions (No. 2018-01-1200), SAE Technical Paper, 2018.

[27] A. A. Salatino, F. Osborne, and E. Motta, "AUGUR: forecasting the emergence of new research topics," ACM In Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, pp. 303-312, May 2018..

[28] H. Small, K. W. Boyack, and R. Klavans, "Identifying emerging topics in science and technology," Research Policy, Vol. 43(8), 2014, pp. 1450-1467.

[29] H. You, M. Li, K. W. Hipel, J. Jiang, B. Ge, and H. Duan, Development trend forecasting for coherent light generator technology based on patent citation network analysis. Scientometrics, Vol. 111(1), 2017, pp.297-315.

[30] D. Dellermann, N. Lipusch, and M. Li, "Combining Humans and Machine Learning: A Novel Approach for Evaluating Crowdsourcing Contributions in Idea Contests," in Multikonferenz Wirtschaftsinformatik, 2018.

[31] M. Stevanovic, D. Marjanovic, and M. Storga, "A model of idea evaluation and selection for product innovation," in DS 80-8 Proceedings of the 20th International Conference on Engineering Design: Innovation and Creativity Vol. 15(8), 2015, pp.193-202.

[32] D. A. Asamoah, and R. Sharda, Adapting CRISP-DM Process for Social Network Analytics: Application to Healthcare, AMCIS, 2015.

[33] M. Spruit, and M. Lytras, "Applied Data Science in Patient-centric Healthcare: Adaptive Analytic Systems for Empowering Physicians and Patients," Telematics and Informatics, Vol. 35(4), 2018, pp. 643-653.

[34] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model," Procedia CIRP, 79, 2019, pp. 403-408.

[35] K. Peffers, M. Rothenberger, T. Tuunanen, and R. Vaezi, "Design science research evaluation," In International Conference on Design Science Research in Information Systems, Springer Berlin Heidelberg, pp. 398-410, May 2012.

[36] D. Howard, and D. Dai, "Public perceptions of self-driving cars: The case of Berkeley, California", In Transportation Research Board 93rd Annual Meeting Vol. 14, No. 4502, 2014.

[37] W. Y. Ayele, and G. Juell-Skielse, "Unveiling topics from scientific literature on the subject of self-driving cars using latent Dirichlet allocation," IEEE in 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 1113-1119,. November 2018.

[38] S. González, V. Robles, J. M. Peña, and E. Menasalvas, Instantiation and adaptation of CRISP-DM to Bioinformatics computational processes. DOI, 10(1.217), 2011.

[39] M. Atzmueller, and T. Roth-Berghofer, "The mining and analysis continuum of explaining uncovered," in International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer, London, 9828, pp. 273-278, December 2010.

[40] J. Venter, A. de Waal, and C. Willers, Specializing CRISP-DM for evidence mining. In IFIP International Conference on Digital Forensics (. Springer, New York, NY., pp. 303-315, January 2007.

[41] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, P. Flach, J. Hernández-Orallo, M. Kull, and M. J. Ramírez-Quintana, CASP-DM: Context Aware Standard Process for Data Mining. arXiv preprint arXiv:1709.09003, 2017

[42] C. Catley, K. Smith, C. McGregor, and M. Tracy, "Extending CRISP-DM to incorporate temporal data mining of multidimensional medical data Streams: A neonatal intensive care unit case study," in 2009 22nd IEEE International Symposium on Computer-Based Medical Systems, pp. 1-5, August 2009.

[43] O. Niaksu, CRISP data mining methodology extension for medical domain. Baltic Journal of Modern Computing, Vol. 3(2), 2015, pp. 92.

[44] M. A. Alksher, A. Azman, R. Yaakob, R. A. Kadir, A. Mohamed, and E. M. Alshari, 2016. "A review of methods for mining idea from text," IEEE in Third International Conference on Information Retrieval and Knowledge Management (CAMP), 2016, pp. 88-93.

[45] J. Björk, and M. Magnusson, "Where do good innovation ideas come from? Exploring the influence of network connectivity on innovation idea quality," Journal of Product Innovation Management, Vol. 26(6), 2009, pp. 662-670.

[46] P. Kruse, A. Schieber, A. Hilbert, and E. Schoop, Idea mining–text mining supported knowledge management for innovation purposes. AMCIS, 2013.

[47] M. Rhyn, I. Blohm, and J. M.. Leimeister, "Understanding the Emergence and Recombination of Distant Knowledge on Crowdsourcing Platforms," in 38th International Conference on Information Systems: Transforming Society with Digital Innovation, ICIS, 2018.

[48] A. M. Karimi-Majd, and M. Mahootchi, A new data mining methodology for generating new service ideas. Information Systems and E-Business Management, Vol. 13(3), 2015, pp. 421–443.

[49] K Peffers, T. Tuunanen, M. A. Rothenberger, S. Chatterjee, "A design science research methodology for information systems research," Journal of management information sys-tems Vol. 24 (3), 2007, pp.45-77

[50] D. M Blei, and J. D. Lafferty, "Dynamic topic models," ACM in Proceedings of the 23rd international conference on Machine learning, pp. 113-120,. June 2006.

[51] A. Hindle, N. A. Ernst, M. W. Godfrey, and J. Mylopoulos, "Automated topic naming to support cross-project analysis of software maintenance activities," in Proceedings of the 8th Working Conference on Mining Software Repositories, pp. 163-172, May 2011,

[52] W.Y. Ayele, and G. Juell-Skielse, "Eliciting Evolving Topics, Trends and Foresight about Self-driving Cars Using Dynamic Topic Modeling," in: Arai K., Kapoor S., Bhatia R. (eds) Advances in Information and Communication. FICC 2020. Advances in Intelligent Systems and Computing, vol 1129, Springer, Cham, 2020..

[53] T. McIntosh, T. J. Mulhearn, and M. D. Mumford, Taking the good with the bad: The impact of forecasting timing and valence on idea evaluation and creativity. Psychology of Aesthetics, Creativity, and the Arts, 2019.

[54] W. Y. Ayele, and I. Akram, "Identifying Emerging Trends and Temporal Patterns about Self-driving Cars in Scientific Literature," in: Arai K., Kapoor S. (eds) Advances in Computer Vision. CVC 2019. Advances in Intelligent Systems and Computing, vol 944. Springer, Cham, 2020.

[55] P. Mongeon, and A. Paul-Hus, "The journal coverage of Web of Science and Scopus: a comparative analysis", Scientometrics, Vol. 106(1), 2016, pp. 213-228.

[56] C. A. Aghaei, H. Salehi, M. Yunus, H. Farhadi, M. Fooladi, M. Farhadi, and E.N Ale, "A comparison between two main academic literature collections: Web of Science and Scopus databases", 2013.

[57] H. Gulati, "Predictive analytics using data mining technique," in 2nd International Conference on Computing for Sustainable Global Development (INDIACom)," IEEE pp. 713-716, March 2015.

[58] J. Rathbone, M. Carter, T. Hoffmann, and P. Glasziou, Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module. Systematic reviews, Vol. 4(1), 2015, pp. 6.

[59] R. Feldman, and J. Sanger, The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press, 2007.

[60] H. Nguyen, and D. Hovy, "Hey Siri. Ok Google. Alexa: A topic modeling of user reviews for smart speakers," in Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), pp. 76-83, November 2019.

[61] M. El Akrouchi, H. Benbrahim, and I. Kassou, "Early warning signs detection in competitive intelligence," in Proceedings of the 25th International Business Information Management Association Conference—Innovation Vision 2020, pp. 1014-1024.

[62] D. O'callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," Expert Systems with Applications, Vol. 42(13), 2015, pp. 5645-5657.

[63] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," ACM in Proceedings of the eighth ACM international conference on Web search and data mining, pp. 399-408, February 2015..

[64] D. K. Simonton, (1977). Cross-sectional time-series experiments: Some suggested statistical analyses. Psychological Bulletin, Vol. 84(3), 1977, pp. 489.

[65] G. E. Box, and G. C. Tiao, (1975). Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical association, Vol. 70(349), 1975, pp. 70-79.

[66] A. G. Asuero, A. Sayago, and A. G. Gonzalez, The correlation coefficient: An overview. Critical reviews in analytical chemistry, Vol. 36(1), 2006, pp. 41-59.

[67] V. Braun, V. and Clarke. "Using thematic analysis in psychology," in Qualitative research in psychology, Vol. 3(2), 2006, pp.77-101

[68] G. Maskeri, S. Sarkar, and K. Heafield, "Mining business topics in source code using latent dirichlet allocation," in Proceedings of the 1st India software engineering conference, pp. 113-120, February 2008,

[69] T. Ha, B. Beijnon, S. Kim, S. Lee, and J. H. Kim, "Examining user perceptions of smartwatch through dynamic topic modeling," Telemat. Informatics, vol. 34, no. 7, 2017, pp. 1262–1273.

[70] N. Steils, and S. Hanine, Creative contests: knowledge generation and underlying learning dynamics for idea generation. Journal of Marketing Management, vol. 32, no. (17-18), 2016, pp 1647-1669.

[71] W.Y. Ayele, G. Juell-Skielse, A. Hjalmarsson, P. Johannesson, and D. Rudmark, Evaluating Open Data Innovation: A Measurement Model for Digital Innovation Contests. In *PACIS* (p. 204), 2015.

[72] K. Christensen, J. Scholderer, S.A. Hersleth, T. Næs, K. Kvaal, T. Mollestad, N. Veflen, and E. Risvik, How good are ideas identified by an automatic idea detection system? Creativity and Innovation Management, vol. 27, no. 1, 2018, pp. 23–31.

A Pedestrian Detection and Tracking Method for Robot Equipped with Laser Radar

Zhu Bin[1], Zhang Jian Rong[2], Wang Yan Fang[3], Wu Jin Ping[4]
AI research center, School of Mechanical and Engineering
Jiangxi college of applied technology
Ganzhou, China

Abstract*—In order to detect and track pedestrians in complex indoor backgrounds, a pedestrian detection and tracking method for indoor robots equipped with Laser radar is proposed. Firstly, The SLAM (Simultaneous Location and Mapping) technology is applied to obtain 2D grid map for a strange environment; then, Monte Carlo localization is employed to obtain the posterior pose of the robot in the map; then, an improved likelihood field background subtraction algorithm is proposed to extract the interesting foreground in changeable environment; then, the hierarchical clustering algorithm combining with an improved leg model is proposed to detect the objective pedestrian; at last, an improved tracking intensity formula is designed to track and follow the objective pedestrian. Experimental results in some complex environments show that our method can effectively reduce the impact of confusing scenarios which are challenges for other algorithms, such as the motion of the chair, the suddenly passing by person and when the objective pedestrian close to the wall and so on, and can detect, track and follow pedestrians in real time with high accuracy.

Keywords—Laser radar; likelihood field model; pedestrian detection; pedestrian tracking; simultaneous location and mapping

I. Introduction

Under the fast development of artificial intelligence, the research and application of intelligent service robot attracts more and more scholars and researcher’s attention, and the application fields of robots cover all aspects of human life, such as restaurant service, Home-based services, shopping guide, accompanying dance and so on [1]. The main work of these robots is to interact with people, so the detection, tracking and avoidance of pedestrians are particularly important.

At present, pedestrian detection and tracking methods mainly depend on ordinary camera [2-3], infrared camera [4-7], laser radar and so on. Because the ordinary camera based method are easy to be affected by light, field size, shooting angle and so on, so these kinds of method are only suitable for the application scene with stable lighting and relatively fixed shooting position. The infrared camera based method can be employed in large workplace and is not affected by visible light and these methods can work normally even in dark environment. However, due to the technical limitation, many drawbacks exist， such as low signal-to-noise ratio, low contrast, low resolution, and the cost is high. Because the laser radar based methods have the advantage of high precision and low cost and is not affected by illumination changes, it is widely used. For the consideration of performance and price, 2D laser radar is always suitable. However, two-dimensional plane information does not include the depth of the image and what’s more, the longer the detection distance, the lower the resolution. Authors in [8-9] proposed 14 features of human leg, and used AdaBoost strong classifier to detect pedestrians. Authors in [10-11] proposed improved convolutional neural network to classify wheelchairs and human legs. However, in complex background, due to the lack of two-dimensional information, the classifier would output wrong results, such as the chair leg near the corner would be misjudged as a human leg. In order to reduce the interference of complex background, [12] proposed a background subtraction method, which reduces the interference of background in fixed laser radar and fixed scene, but can’t be applied to mobile robot. Author in [13] firstly generate the local grid map in the environment, then match and align the grid map of the front and back frames to get the undetermined foreground. At the same time, assuming that the human leg corresponds to the minimum value of the laser radar distance histogram, the final foreground can be obtained by the operation of the laser radar distance histogram of the undetermined foreground and the laser radar distance histogram of the human lag. This method can be well applied in the relatively open environment such as corridor, but it can’t deal with the situation that pedestrian is still and environment is complex.

In order to solve the problem of background interference in pedestrian detection and tracking, this paper propose a newly method, first, the environment map is constructed, and then the likelihood domain model is used to segment the foreground from the background; at last, the improved Kalman filter is used to track and follow the pedestrian in the complex background.

II. Flowchart of our Method

As shown in Fig. 1, after the SLAM technology is applied to construct the environment map, the Monte Carlo localization is applied to determine the location of the robot in the map. Then, the foreground is extracted by likely likelihood domain model, the data cluster technology is employed to generate the steady background, which would enhance the accuracy of the foreground extraction. Then the foreground is judged whether the objective pedestrian exist, if the objective pedestrian is localized, then the robot will follow the pedestrian automatically. Otherwise, the robot will standstill or cruise randomly and waiting for the result of the next frame.

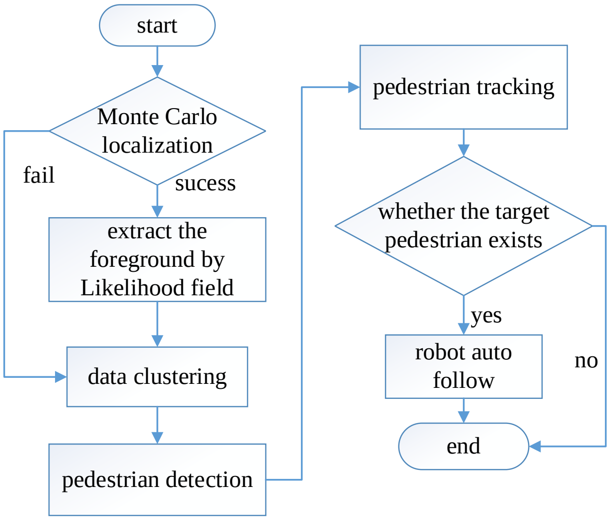


Fig. 1. The Flowchart of the Method.

III. Likelihood Field Background Difference

A. The Map Construction

The SLAM technology of mobile robot is first proposed by R.Smith, M.Self and P.Cheeseman [14], which locate the robot's pose when the robot move in an unknown environment, and then build a map according to the robot's pose, so as to achieve the purpose of Self-orientation and mapping at the same time. Grisetti G et al. [15] improved the traditional SLAM by rao-blackwellized particle filter, and proposed adaptive sampling to reduce particle loss, forming the current GMapping algorithm. GMapping is used to build environment map, because the method has the advantages of low requirement for the performance of the laser radar, low calculation consumption and high accuracy of the mapping.

Fig. 2 is the grid map of the experimental scene using our own device and GMapping technology. In the figure, the white part indicates that the robot can move in these areas; the black part indicates that the robot cannot move in these areas; the gray part means area still not be detected; numbers represent the position of where to operate background difference experiment.

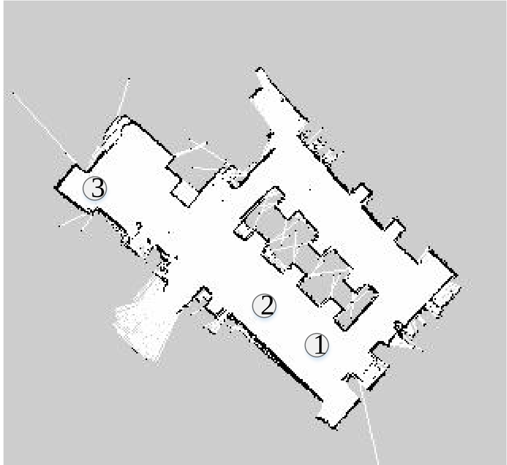


Fig. 2. The Grid Map of the Experimental Scene.

B. Monte Carlo Localization

Monte Carlo localization [16] is an algorithm which can be used to determine the position and direction for a robot in the grid map using odometer information and laser radar data. The algorithm first initializes a particle swarm in normal distribution using to standard mean and variance, then updates the pose of all particles in the particle swarm by the odometer data and the motion model, then obtains the importance weight of the particles by calculating the correspondence between the laser radar data and the map under the corresponding pose, and finally the maximum possibility rule is applied to resample the particle swarm, and the pose with the largest weight is treated as a posteriori pose. What's more, the random particles are usually added into the resampling step to recover the robot from global positioning failure and local optimal solution.

C. Likelihood Field Model

The likelihood field model is first applied to eliminate the uncertainty of the signal obtained by various sensors, the possibility of the value of signal intensity obtained by various sensors is employed to determine the final value of the signal, doing so, the output of the sensor would be more robust to the influence of the noise and the fluctuation of the environmental factors such as voltage, temperate, humidity and so on. In this paper, the likelihood field model is used to obtain a steady background, then under the circumstance that a constructed grid map and the position and direction of the robot is determined, the foreground can be extracted accurately.

The likelihood field model can be represented by a conditional probability distribution $p(z_t\,|\,x_t,m)$:

$$p(z_t\,|\,x_t,m)=\prod_{k=1}^{K}p(z_t^k\,|\,x_t,m) \tag{1}$$

Where, z_t is the measurement value at time t, $x_t=(x\quad y\quad \theta)^T$ is the pose of the robot, and m is the environmental map. Suppose that K measurement points are available for the data obtained by the laser radar sensor, and each measurement value obey noise independent distribution. Where, k is the number of the measurement point and z_t^k is the measurement value of the k-th measurement point at time t.

But traditional likelihood field model do not consider the influence of the noise and the uncertainty of the environment in the construction of the background, to solve this problem, improved likelihood field model is proposed:

$$p(z_t^k\,|\,x_t,m)=z_{hit}\,p_{hit}+z_{rand}\,p_{rand}+p_{original} \tag{2}$$

$$p_{hit}(z_t^k\,|\,x_t,m)=e_{s_{hit}}\left(dist\right) \tag{3}$$

$$p_{rand}(z_t^k\,|\,x_t,m)=\frac{1}{z_{max}} \tag{4}$$

Where, $p_{original}$ represents the original likelihood value, it can be calculated using formula (1). p_{hit} is the likelihood value affected by noise, $dist$ is the Euclidean distance between the coordinate and the nearest obstacle on map m. It is considered that p_{hit} obeys the Gaussian distribution with the mean equal to 0 and variance equal to 1. p_{rand} is the likelihood value affected by objects randomly showing up. It is considered that the likelihood of obstacles detected by the measurement points obey an average distribution, which z_{max} is the maximum measurement distance. Based on the above two considerations, the likelihood of the objects detected in the map by the corresponding sensor, and z_{hit}, z_{rand} is the weight, respectively. Fig. 3 shows the relationship between likelihood and dist.

The value of likelihood represents the possibility of a surrounding point of the measurement point is a background, and it is related to the shortest distance $dist$ from the measurement point to the obstacle in the map. In this paper, a fixed threshold value $theta\_pk$ is set. For each $p(z_t^k \mid x_t, m)$, when it is bigger than the threshold value, the scanned region is the background. When the calculated likelihood value is smaller than the threshold, it is considered that the scanned region include foreground. $theta\_pk$ equal to 0.5, and the $dist$ equal to 0.12 in our experiment.

$$\begin{pmatrix} \mathrm{x}_{z_t^k} \\ y_{z_t^k} \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_{k,sens} \\ y_{k,sens} \end{pmatrix} + z_t^k \begin{pmatrix} \cos\left(\theta+\theta_{k,sens}\right) \\ \sin\left(\theta+\theta_{k,sens}\right) \end{pmatrix} \qquad (5)$$

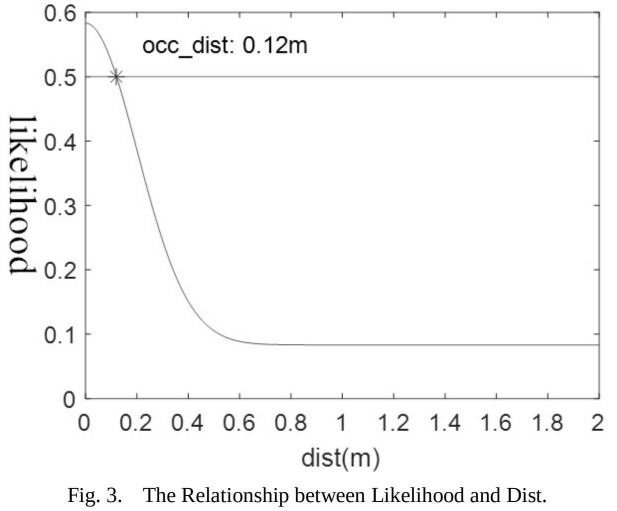


Fig. 3. The Relationship between Likelihood and Dist.

When the coordinates for each measurement point are obtained, the transformation for the world coordinate into the robot coordinate is needed. Where $\left(x_{k,sens} \quad y_{k,sens}\right)^{\mathrm{T}}$ represents the position of laser radar sensor in robot coordinate system and $\theta_{k,sens}$ is the angle of the measurement point.

IV. Pedestrian Detection and Tracking

A. Data Clustering

After obtained the foreground information taken from laser radar data, the clustering operation is then followed, the foreground is then classified into several class, and then is used to recognize. The state-of-the-art algorithms can be departed into three classes, they are hierarchical clustering algorithm, K-means clustering algorithm and DBSCAN clustering algorithm [13]. K-means clustering algorithm needs to know the number of clusters in advance, which is not suitable for the situation of unknown pedestrian and environment. DBSCAN is a density-based algorithm, which calculate the number of elements in a circle with a certain radius. As long as the number of elements exceeds the preset threshold, these elements in the circle are regarded as a class. The effect of DBSCAN algorithm depends on the radius of circle and the number of elements, and the calculation is huge. The hierarchical clustering algorithm first treats each data to be processed as a class, calculates the distance between classes, then compares the distance between adjacent classes, and merge points whose distance is less than the preset threshold value into a class.

The hierarchical clustering algorithm is selected in this paper, which does not need to specify the number of clustering results and has few restrictions. In traditional hierarchical clustering algorithm, the singular value would negatively affect the clustering results. To solve this problem, the laser radar foreground data is preprocessed firstly, directly removes the singular points which is far from the laser radar, and then conducts hierarchical clustering.

B. Pedestrian Detection

Although there are many extracted features related to human legs, [8] pointed out that only a few features will occupy a relatively large weight in the final trained AdaBoost detector, and most of the other features will have less weight. Adding these features may lead to over fitting.

On the basis of [8], the leg model is designed and can be seen in Fig. 4, in which two side points represent the cluster of each leg and center point represent the location of the pedestrian. D_{k} represents the distance between the head and tail elements of the corresponding cluster, and L_{k} represents the total length of the corresponding cluster. Finally, the detector gives the middle position of the pedestrian leg model as the pedestrian coordinate. The specific process of the pedestrian detector is shown in Fig. 5.

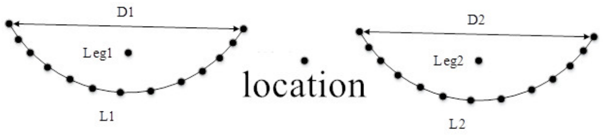


Fig. 4. The Leg Model used in our Paper.

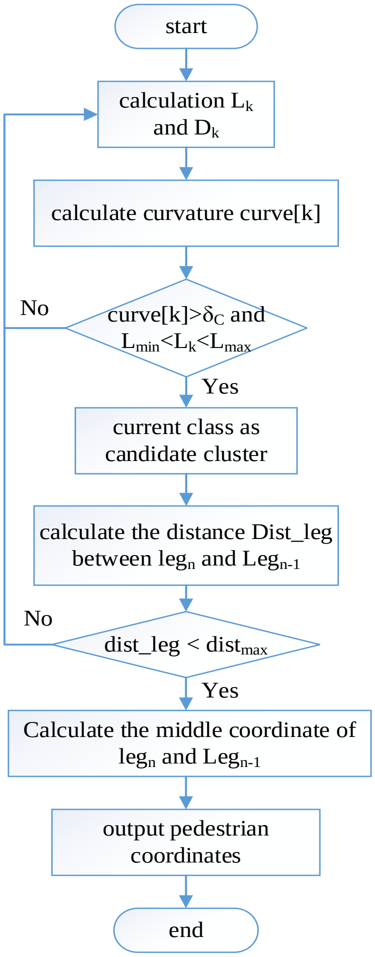

start
calculation L_k and D_k
calculate curvature curve[k]
No
curve[k]>δ_C and L_{min}<L_k<L_{max}
Yes
current class as candidate cluster
calculate the distance Dist_leg between leg_n and Leg_{n-1}
No
dist_leg < $dist_{max}$
Yes
Calculate the middle coordinate of leg_n and Leg_{n-1}
output pedestrian coordinates
end

Fig. 5. The Flowchart of Pedestrian Detection.

C. *Pedestrian Tracking*

In order to solve the problem of one of leg is occluded temporarily, the pedestrian tracking algorithm as mentioned in [17-18] is applied, and rectified the tracking intensity formula.

$$\mathrm{cf}_k = \begin{cases} cf_{k-1} + \log_{1.5}(cfi_k), & cfd = 0 \\ cf_{k-1} - 1.5^{cfd_k}, & cfi = 0, hide = 0 \\ cf_{k-1} - \log_{1.5}(cfd_k), & cfi = 0, hide = 1 \end{cases} \tag{6}$$

When a pedestrian suddenly lose in a scene, the history tracking information is used to judge whether the losing is caused by one of leg is occluded, if yes, the pedestrian is just occluded temporarily. In the actual experiment, it is found that the improved algorithm can effectively track multiple pedestrians, and robust in circumstances such as a pedestrian suddenly break in or leave or be occluded.

D. *Automatically Following*

The pedestrian is followed within 30cm in front of the robot as the following target until the tracking strength of the corresponding tracker is reduced to a certain extent or the following is given up artificially. After obtain the global coordinate of the objective pedestrian (x_p, y_p) and the robot coordinate (x_r, y_r), the next robot posture (x_g, y_g, yaw_g) is determined by direction of the line between the robot and the pedestrian, then the posture is sent to ROS (Robot Operating System) platform, and using the navigation function to complete the automatically following.

$$\begin{aligned} \mathrm{d}rp &= sqrt\left(\left(x_p - x_r\right)^2 + \left(y_p - y_r\right)^2\right) \\ \mathrm{d}rt &= \mathrm{d}rp - 0.2 \\ x_g &= x_r + \mathrm{d}rt * \left(\left(x_p - x_r\right) / \mathrm{d}rp\right) \\ y_g &= \mathrm{y}_r + \mathrm{d}rt * \left(\left(y_p - y_r\right) / \mathrm{d}rp\right) \end{aligned} \tag{7}$$

$$yaw_g = \arctan\left(\left(y_p - y_r\right), \left(x_p - x_r\right)\right) \tag{8}$$

V. Experimental Results and Analysis

The platform used in our experiment as can be seen in Fig. 6, is a wheel robot, which equipped with a Flash Lidar F4 laser radar which the scanning angle is 360, the angle resolution is 0.5 and the frame rate is 10FPS, two wheels with distance encoder, an industrial computer with Intel i7-3610 and using C++ in ROS to realize the algorithm.

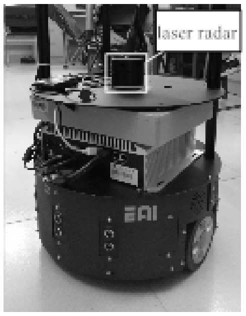
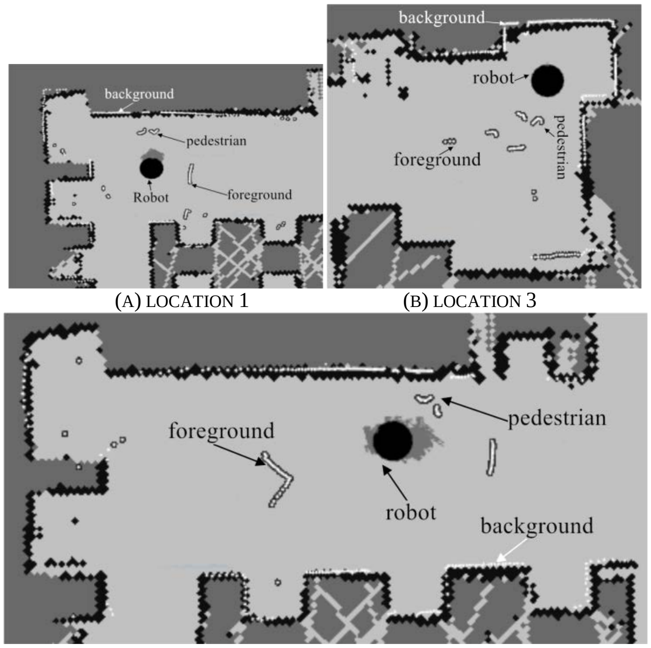


Fig. 6. Experimental Platform.

A. *The Experiment of Likelihood Field Background difference*

The experimental results are shown in Fig. 7. When the robot is in the three locations as shown in Fig. 2. The black outline in the figure is the original obstacle of the map, the white fine points on the obstacle are the laser radar background data separated from the likelihood field, the points with thick white and black edges are the laser radar foreground data extracted from the likelihood field method, the black circle in the middle is the location of the robot, and it can be seen that the location of the objective pedestrian can be figured out in each image.

Using the method in [19], when the region detected by laser radar close to the obstacle in background, they would be treated as a background. Only when the region is far away from the background, they would be regarded as a foreground. But in Fig. 7, which applies our method, the foreground and the background can be classified correctly.



(A) LOCATION 1 (B) LOCATION 3

(C) LOCATION 2

Fig. 7. The Result of Likelihood Field Background difference.

B. The Anti Interference Experiment

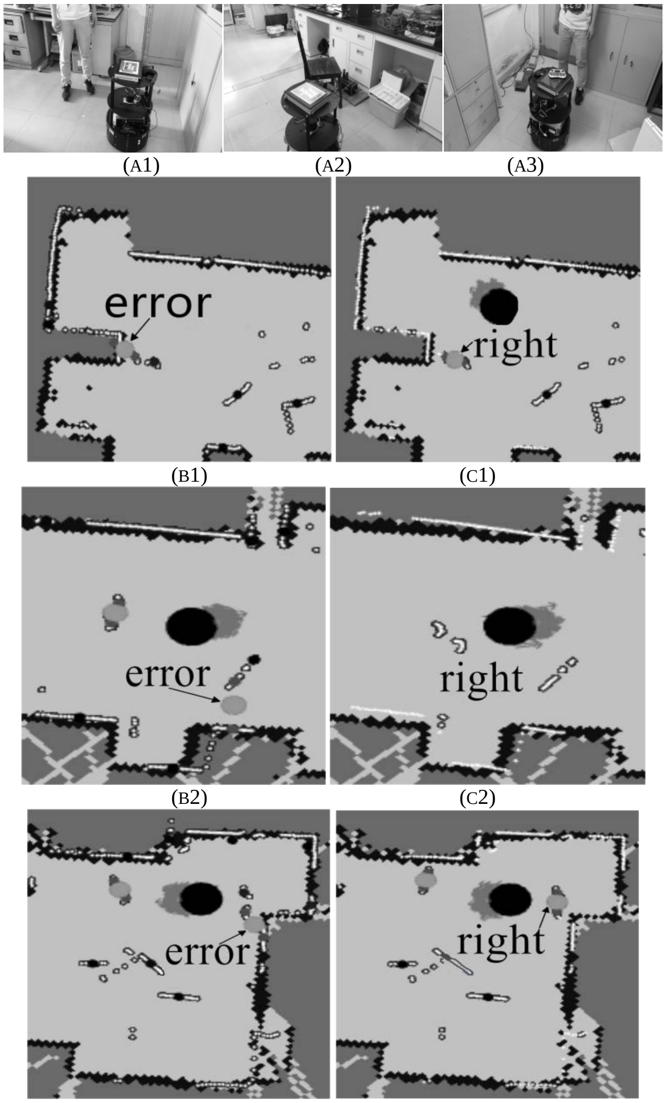
The method of [20] is used to pedestrian detection experiment. Compared the method with likelihood field background difference and the original method, it can be seen that our method can require better results in complex environments. Fig. 9 shows the experimental results, in which (a) is the actual scene, (b) is the result of our method (ours) and (c) is the result of original method (original).

Three scenes are used to test the algorithm, Fig. 8(a1) and Fig. 8(a3) is the circumstance that the pedestrian would easily be mistreated as background, Fig. 8(a2) is the circumstance that the leg of the chair would be mistreated as pedestrian. it can be seen in Fig. 8(b1), Fig. 8(b2), Fig. 8(b3), using original method, three wrong judgments are received in these three scenes. When using our method, we can receive the correct pedestrian detection results, in Fig. 8(c1) and Fig. 8(c3), the accurate location of the pedestrian is detected, and in Fig. 8(c2), the leg of the chair is not mistreated as pedestrian.

Experimental results show that after using likelihood field background difference, background and foreground can be classified correctly in complex environment, so as to reduce the interference of the complex environment to pedestrian detection.

C. The Experiment of Pedestrian Detection

The pedestrian experiments are done in two locations and with two different detection distances (50cm or 50-100cm). The experimental results can be seen in Fig. 9. The method can always detect the pedestrian in different locations and distances, which demonstrate the practicability of our method.



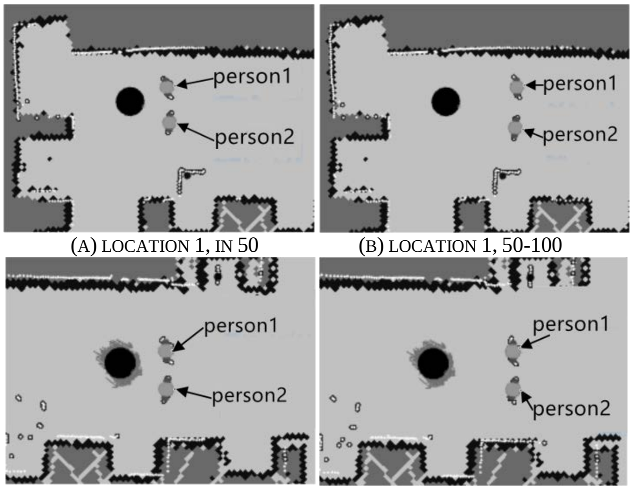
(A1) (A2) (A3)

(B1) (C1)

(B2) (C2)

(B3) (C3)

Fig. 8. The Result of Pedestrian Detection.



(A) LOCATION 1, IN 50 (B) LOCATION 1, 50-100

(C) LOCATION 2, IN 50 (D) LOCATION 2, 50-100

Fig. 9. The Result of Pedestrian Detection.

D. The Experiment of Pedestrian Tracking and Following

Our method is then used to pedestrian tracking and following, the experimental result can be seen in Fig. 10.

As shown in Fig. 10, black Pentagram and black circle represents the start and end points of the path of particle of the robot and the target pedestrian. The black thick line between the start point and the end point is the trajectory of the robot, and the white thick line is the trajectory of the target pedestrian. In order to test the stability of the proposed detection algorithm and tracking algorithm, a non-target pedestrian showed up, first move parallel to the target pedestrian, and finally surpass the target pedestrian and run counter to the target pedestrian. The track of non-target pedestrian is marked by white arrow, and the intermittent gray thick line represents the track of non-target pedestrian.

It can be seen that the trajectory of the robot and the target pedestrian basically coincides, and the emergence of non-target pedestrian does not affect the robot's follow-up to the target pedestrian. Therefore, the follow-up strategy in this paper can enable the robot to eliminate interference and track the target pedestrian continuously and stably.

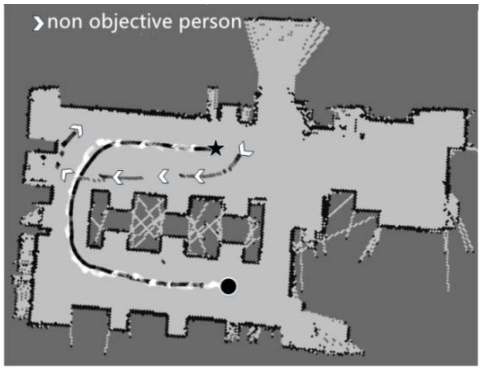


Fig. 10. The Result of Pedestrian Tracking and Following.

VI. CONCLUSION

This paper proposes a pedestrian detection, tracking and following algorithm in complex environment using a mobile robot with laser radar. The algorithm firstly maps the environment and then extracts the foreground data by the likelihood domain model to reduce the interference of complex background, the hierarchical clustering algorithm is used to cluster the foreground data, and then the improved Kalman tracking algorithm is used to effectively track the multi pedestrians. Finally, the automatic tracking strategy proposed in this paper is used to effectively follow the target pedestrians in the known map environment. Experiment result shows that the whole system has high real-time performance and is not interfered by complex background, and has certain practical value. The future work would focus on the combination of the laser radar and machine vision.

REFERENCES

[1] S. Liu, L. Song, M. Gen,et al. “Indoor Pedestrian Tracking Method of Dancing Robot Based on Laser Radar,” Computer Engineering, vol.43, no.6, pp. 247-252, 2017.

[2] G. Jiachen, X. Juan, Z. Hongfu, F. Hang, Z. Zhirong and X. Xiangqian, "Civil aircraft surface defects detection based on histogram of oriented gradient," 2019 IEEE 1st International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Kunming, China, 2019, pp. 34-38.

[3] P. Sombatpiboonporn, T. Charoenpong, A. Supasuteekul, C. Chianrabutra and K. Pattanaworapan, "Human Edge Segmentation From 2D Images By Histogram of Oriented Gradients and Edge Matching Algorithm," 2019 First International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP), Bangkok, Thailand, 2019, pp. 29-32.

[4] Wang G H, Liu Q, Zhuang J J. “Research on vehicle Infrared Pesestrian Detection method based on Local Features”, Acta Electronica Sinica, 2015, 43(7): 1444-1448.

[5] Zhu C C, Xiang Z Y. “Infrared pedestrian detection based on gradient direction and intensity histogram”, Computer engineering, vol. 40, no. 12. pp. 195-198. 2014.

[6] L. Li, F. Zhou, Y. Zheng and X. Bai, "Reconstructed Saliency for Infrared Pedestrian Images," in IEEE Access, vol. 7, pp. 42652-42663, 2019.

[7] T. Y. Han and B. C. Song, "Night vision pedestrian detection based on adaptive preprocessing using near infrared camera," 2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), Seoul, 2016, pp. 1-3.

[8] Beyer L, Hermans A, Leibe B. “DROW: Real-Time Deep Learning based Wheelchair Detection in 2D Range Data”, IEEE Robotics and Automation Letters, vol. 2, pp. 585-592, 2017.

[9] Y. Maalej, S. Sorour, A. Abdel-Rahim and M. Guizani, "Tracking 3D LIDAR Point Clouds Using Extended Kalman Filters in KITTI Driving Sequences," 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 2018, pp. 1-6.

[10] T. Tsai and C. Yao, "A Real-time Tracking Algorithm for Human Following Mobile Robot," 2018 International SoC Design Conference (ISOCC), Daegu, Korea (South), 2018, pp. 78-79.

[11] Q. Ren, Q. Zhao, H. Qi and L. Li, "Real-time target tracking system for person-following robot," 2016 35th Chinese Control Conference (CCC), Chengdu, 2016, pp. 6160-6165.

[12] J. Zou, W. Yin, E. X. Wang, J. Wang and Y. Lu, "Human Motion Prediction Based on Visual Tracking," 2019 4th International Conference on Robotics and Automation Engineering (ICRAE), Singapore, Singapore, 2019, pp. 39-44.

[13] M. Shiomi, K. Shatani, T. Minato and H. Ishiguro, "How Should a Robot React Before People's Touch?: Modeling a Pre-Touch Reaction Distance for a Robot's Face," in IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 3773-3780, Oct. 2018.

[14] H. Jo, H. M. Cho, S. Jo and E. Kim, "Efficient Grid-Based Rao–Blackwellized Particle Filter SLAM With Interparticle Map Sharing," in IEEE/ASME Transactions on Mechatronics, vol. 23, no. 2, pp. 714-724, April 2018.

[15] L. Chen, A. Yang, H. Hu and W. Naeem, "RBPF-MSIS: Toward Rao-Blackwellized Particle Filter SLAM for Autonomous Underwater Vehicle With Slow Mechanical Scanning Imaging Sonar," in IEEE Systems Journal, vol. 7, no. 14, pp. 3163-3170, 2019.

[16] F. Zhang, A. Cully and Y. Demiris, "Probabilistic Real-Time User Posture Tracking for Personalized Robot-Assisted Dressing," in IEEE Transactions on Robotics, vol. 35, no. 4, pp. 873-888, Aug. 2019.

[17] M. Hai. “A survey of big data clustering algorithms”, Computer Science, vol. 43. pp. 380-383, 2016.

[18] K. S. S. Reddy and C. S. Bindu, "A review on density-based clustering algorithms for big data analysis," 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, 2017, pp. 123-130.

[19] H. Kıvrak and H. Köse, "Social robot navigation in human-robot interactive environments: Social force model approach," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4.

[20] A. Garrell, L. Garza-Elizondo, M. Villamizar, F. Herrero and A. Sanfeliu, "Aerial social force model: A new framework to accompany people using autonomous flying robots," 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, 2017, pp. 7011-7017.

Comparative Study of EIGRP and OSPF Protocols based on Network Convergence

Ifeanyi Joseph Okonkwo[1]
Imperial College London
London, United Kingdom

Ikiomoye Douglas Emmanuel[2]
University of Salford
Manchester, United Kingdom

***Abstract*—Dynamic routing protocols are one of the fastest growing routing protocols in networking technologies because of their characteristics such as high throughput, flexibility, low overhead, scalability, easy configuration, bandwidth, and CPU utilization. Albeit convergence time is a critical problem in any of these routing protocols. Convergence time describes summary of the updated, complete, and accurate information of the network. Several studies have investigated EIGRP and OSPF on the internet; however, only a few of these studies have considered link failure and addition of new links using different network scenarios. This research contributes to this area. This comparative study uses a network simulator GNS3 to simulate different network topologies. The results are validated using Cisco hardware equipment in the laboratory. The network topology implemented in this research are star and mesh topology. The results are validated using Cisco hardware equipment in the laboratory. Wireshark is effectively used in capturing and analyzing the packets in the networks. This helps in monitoring accurate time response for the various packets. The results obtained from Wireshark suggest the EIGRP has a higher performance in terms of convergence duration with a link failure or new link added to the network than the OSPF routing protocol. Following this study EIGRP is recommended for most heterogeneous network implementations over OSPF routing protocol.**

Keywords*—OSPF (Open Shortest Path First); EIGRP (Enhanced Interior Gateway Routing Protocol); routing; protocol; network; convergence; topology; routers; packets; Wireshark

I. Introduction

Computer networking is now a fundamental part of life, especially the use of the internet. As new technologies emerge, the demand for wireless mobile computing is growing fast, thus the need for efficient routing protocols [1]. These protocols define the mechanism by which routers acquire information about the performance of the network topology, verify and identify the optimal route that a packet will take to arrive at its destination. Hence, routing algorithms are crucial because they select the best path for communication in a heterogeneous network. Routing is the entire process of selecting the optimal route for the transmission of data packets from source to destination [2]. The process includes routers advertising their known IP networks, the administrative cost to its neighbor or adjacent routers, in this way the neighbor's routers gain knowledge of the characteristics and the topology of the network, then update the routing table. The administrative cost is the number of hops, link speed and latency [3].

There has been different research about routing protocols, especially on EIGRP and OSPF routing protocol in terms of convergence time. This research in routing protocols has been predominant because of the increasing demand of data transmission over a reliable network connection amongst enterprise companies, therefore network resilience/redundancy has been the key in curbing link failures. If there is a link failure in the network, the routing protocols are expected to identify the failure and converge to form a new topology for the continuous flow of packets in the network. Despite the wide research and interest made, many problems about routing protocols are yet to be solved in terms of convergence rate, which can yield optimal routing to deliver high throughput in heterogeneous networks. Hence, this work focuses on analyzing several scenarios of link failures, measuring their convergence rate, and identifying changes in the network topology when using EIGRP and OSPF routing protocol [4]. The rate of convergence occurs when all the routers in the network have an updated, complete, and accurate information on the network. The convergence rate includes the total time required by all the routers to calculate the optimal path, update their routing tables, and share the routing information with neighboring routers in the network.

The remainder of this study is planned as follows. In Section II, present a brief literature of recent work and background knowledge of Routing Information Protocol (RIP), OSPF and EIGRP. Section III describes the methodology, where the network topology such as star and mesh are designed. Section IV presents the design parameters and metrics, which includes the Hello interval, hop count and interface cost. Section V are the Wireshark results and Section VI is the presentation and analyses of the results. Finally, the work is concluded in Section VII.

II. Background

Dynamic routing protocols allow changes in the network topology because of the update in routing tables. Dynamic routing protocol is divided into distance vector routing protocols and link state routing protocols. Distance vector routing protocols calculate the administrative cost of a packet arriving at a destination based on the number of routers the packet passes through, these include Routing Information Protocol (RIP) and EIGRP) [5]. Whilst, Link state routing protocol is building a complete topology of the network and calculates the optimal path from the topology for all interconnected networks, these include Intermediate System to

Intermediate System (IS-IS) and Open Shortest Path First (OSPF) [3].

There are series of research in convergence time, packet loss and throughput of OSPF, EIGRP and RIP routing protocol. Each of this research work has a unique role they play in network efficiency. The first research work was the Advanced Research Projects Agency Network (ARPANET) in 1969, which is the foundation of most routing algorithms [6]. Furthermore, [7] designed a star network topology consisting of a switch, eight cisco routers and 14 hosts using the cisco packet tracer to determine the convergence time. In a similar context [8] determines a comparative study of RIP, OSPF and EIGRP using ring topologies on GNS3 network simulator. The design comprises five routers connected in a ring topology with a personal computer that makes use of networking management tools to manage the information in the network. Also, [9] evaluates dynamic routing protocols for real time applications such as voice, video based on convergence time, end to end delay by using Cisco Packet Tracer[1] and OPNET simulator. The design is implemented using ten routers with two switches and ten personal computers using a mesh topology. Author in [10] analyze EIGRP and OSPF protocol with OPNET for real time application with a focus on large, realistic and scalable networks. Lastly [11], did a comparison of OSPF and EIGRP in a small IPv6 Enterprise Network. Hence, within the best of our knowledge of the literature presented. There is yet to be a deep analysis of OSPF and EIGRP considering the scalability, resilience, and validation of simulator results with Cisco active devices.

A. Routing Information Protocol (RIP)

RIP is the first routing protocol implemented in the TCP/IP and uses the variants of the Bellman-Ford algorithm that was designed by Richard Bellman and Lester Ford in 1958[2]. They perform three functions, discovering the neighbor router addresses, downloading of the routes, and updating the routing table and the cost associated with each route. The first RIP was designed in 1969; it has three versions RIPv1, RIPv2, and RIPng. The latest version of RIPv2 and RIPng works in IPV6 autonomous based systems. The implementation of the exchange of information through the User Datagram Protocol (UDP) and each router is limited to several routers in the network around it. The RIP applies a hop count mechanism to determine the optimal path for packet routing and a maximum of 16 hops is applied to avoid routing loops in the network[2].

B. Open Shortest Path First (OSPF)

OSPF is one of the widely used link state routing protocols. It operates by routing network packets by gathering link state information from neighboring routers thus, computing a map of the network. OSPF sends different messages, which include the hello messages, link state request, updates, and database description packets[3]. OSPF operates with Dijkstra's algorithm, which focuses on the distribution of routing information in a single autonomous system. There are different versions of OSPF; the first version was designed in 1989, which is known as OSPFv1 published in RFC 1131, in 1998 the second version OSPFv2 published in RFC 2328 and in 1999, the OSPFv3 is designed specifically to accommodate the IPv6 published in RFC 5340[4]. OSPF calculations are computed periodically on the link state advertisement (LSA) received in the network and protocol information [12]. A change in the topology is detected quickly; hence, it is fast, flexible, and scalable in terms of configuration parameters. The metric represents the path cost between interfaces in OSPF and that define the speed, bandwidth from nodes to another in the network [3].

C. Enhanced Gateway Routing Protocol (EIGRP)

The Enhanced Gateway Routing Protocol is a hybrid routing protocol developed in 1994. EIGRP focuses on Classless Inter-Domain Routing/Variable length Subnet Mask, route summarization with discontinuous networks and supports load balancing across six routes to a single destination. The EIGRP is designed based on the DUAL (Diffusing Update Algorithm) algorithm and uses multicast for routing updates [13]. The DUAL algorithm is used in obtaining route freedom every time throughout different routing computation and uses the reliable transport protocol to ensure the successful delivery of each packet [13].

III. Methodology

In the design of network scenarios, there are two network topologies implemented. These topologies are used in determining the convergence time of EIGRP and OSPF routing protocol. In the analyses, design of four, six, eight till twenty routers are implemented for both Star and Mesh topologies for EIGRP and OSPF routing protocol using a network simulator and Cisco hardware equipment.

A. Star Topology

In this topology, all the devices are connected to a central hub or switch in a point-to-point connection. The advantage of this topology, it is easy to troubleshoot and isolate problems. It is easily expanded without disruption of the network topology. In this design, the use of loop backs is implemented because a star topology is based on a single network, but since routing applied to a heterogeneous network, it allows hop-to-hop transmission of data. Hence, loop back helps in creating a virtual subnet in the network and each virtual subnet has a network ID as a result making the network to be heterogeneous [14]. Fig. 1 shows a simple design of six routers with the loopback network as virtual subnets.

B. Mesh Topology

Mesh topology is a topology where all devices are connected to each other. Hence, they have a high level of redundancy. They are rarely implemented in today's networks because of the cabling cost, wiring which is complicated and the problem faced in troubleshooting the network at failure. There are two variations of mesh topology, full and partial mesh topology, in this design a partial mesh topology is implemented because of the number of ports in the routers [15]. Fig. 2 shows a diagram showing partial mesh topology designed to be used in the network.

[1] Cisco Packet Tracer: https://www.netacad.com/courses/packet-tracer
[2] RFC 2453, RIP Version 2 https://tools.ietf.org/html/rfc2453
[3] RFC 2328, OSPF Version 2 https://tools.ietf.org/html/rfc2328
[4] RFC 5340, OSPF for IPv6 https://tools.ietf.org/html/rfc5340

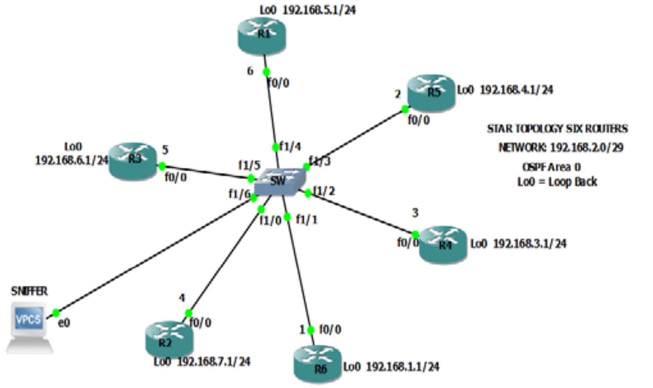


Fig. 1. Design of Star Topology for Open Shortest Path First Routing Protocol using Loop Back.

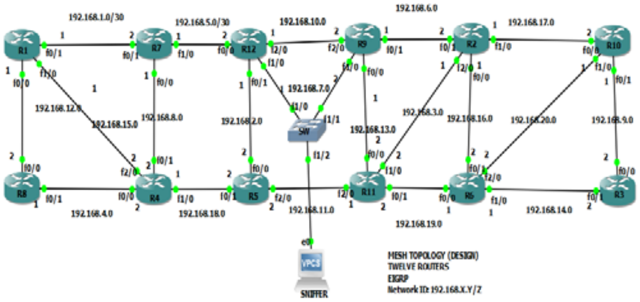


Fig. 2. Design of Mesh Topology for Enhanced Interior Gateway Routing Protocol.

IV. Design Parameters and Metrics

In the design of the EIGRP and OSPF Routing Protocols, there are parameters that are considered in the design such as the interface cost, hello interval, and maximum hops. These parameters are key in the implementation of the EIGRP and OSPF Routing Protocols in both the network simulator and hardware implementation.

A. Open Shortest Path First Routing Protocol

In the OSPF routing protocol, the cost associated with the interfaces depends on the network cables used in the design. In each of the topology designs, the interface cost is equal to one (1) because Fast Ethernet is implemented and for the hello interval ten seconds is used. The router dead interval and transmission delay are set to be 40 and one seconds respectively [16]. OSPF routing protocol does not have a maximum number of hops. See Table I for setup.

B. Enhanced Interior Gateway Routing Protocol (EIGRP)

In EIGRP, the cost associated with the interfaces is one (1) for both software and hardware implementation. Since Fast Ethernet link is used so the cost is equal to one (1) while in the hello interval is ten seconds and the hold time is three times the hello interval. The split horizon is enabled to avoid advertisement of route to the neighbor from which route was learned [17]. See Table II for setup.

C. Convergence Duration

The convergence duration occurs when all the routing tables in all the routers in each network are consistent. In OSPF, convergence duration involves the total time taken for all the routers to exchange the database description packets among the routing tables on the network. These include determining the best path and sharing the complete information in all the routers in the network. In EIGRP, the convergence time is the total time taken for the updates packets and acknowledgment packets to distribute the routing information among the different routers in the network. Furthermore, the complete time required for each router in the network to have complete information on the neighboring routers defines the convergence time. In addition, it comprises the speed of transmission and calculation of the optimal paths taken [18].

D. Convergence Startup Time

This is a measure of how fast and precise individual routers in a group or network are connected dynamically to exchange their routing among themselves for the first time in the network. This is very important because the faster the network converges the faster it can start its routing process.

E. Convergence Failure

This is a measure of how fast and precise time taken for individual routers in a network to converge dynamically or adapt to changes in the network such as node failure, loop back or any other factor that causes a network to fail.

F. Convergence New Link

This is to measure how fast and precise time taken for individual routers in a network to converge dynamically or recover from changes in the network as a result of adding a new link or nodes. Hence, in the addition of a new link or node, the convergence time will be changed [19], owing to determination of the new convergence time.

TABLE I. Design Parameters value for OSPF Routing Protocol

Parameters	Implemented
Interface Cost	1.00
Hello Interval	10.00 seconds
Router Dead Interval	40.00 seconds
Transmission Delay	2.00 seconds
Retransmission Interval	5.00 seconds
Number of Hops	Unlimited

TABLE II. Design Parameters Value for Enhanced Interior Gateway Routing Protocol

Parameters	Implemented
Interface Cost	1.00
Hello Interval	10.00 seconds
Hold Time	30.00 seconds
Split Horizon	Enabled
Number of Hops (limited)	100

G. Graphical Network Simulator (GNS3)

GNS3[5] (1.5.4) is an open source software with no limitation of the number of devices that will be used in the environment. It mimics a real-time network scenario simulation for pre-deployment without the need for hardware. Omnet++, NS2, and OPNET are also powerful open source software that can be used in designing network models. They are built on the platform of a discrete event simulator. Which is used in networking research and provides a comprehensive development environment to support user-defined models [18]. However, GNS3 comes with an inbuilt Wireshark for packet capturing and monitoring.

H. Wireshark

Wireshark[6] is a powerful computer software used for network packet analyzer. The network packet analyzer verifies and captures network packets and displays comprehensive information of the packet data. Some useful features include capturing live packet data, displaying packets with detailed information, filter packets, search, and create various statistics about the packet data. In this work, Wireshark comes inbuilt with GNS3, making the capture of the packet data to be more realistic and effective than using extra hardware to capture the information of the packets [20]. Wireshark is proposed ahead of other monitoring devices because it captures network failures, recovery, and jitter performance of the two protocols. Therefore, Wireshark is used in the hardware implementation and importantly, Wireshark does not require any external component for the capturing of packets [20].

I. Cisco Packet Tracer

The Cisco Packet[7] Tracer is an innovative network and technology tool developed by Cisco Networking Academy. It provides a combination of realistic simulation and visualization experiences for different user's collaborations. In this work, it is used in designing the network topology that will be used for analysis or reference models.

V. Wireshark Result

The Wireshark results helps in checking the network configuration, a design implemented using the EIGRP and OSPF routing protocol. The results monitored or obtained are expected to mimic most of the characteristics highlighted in the background knowledge of OSPF and EIGRP.

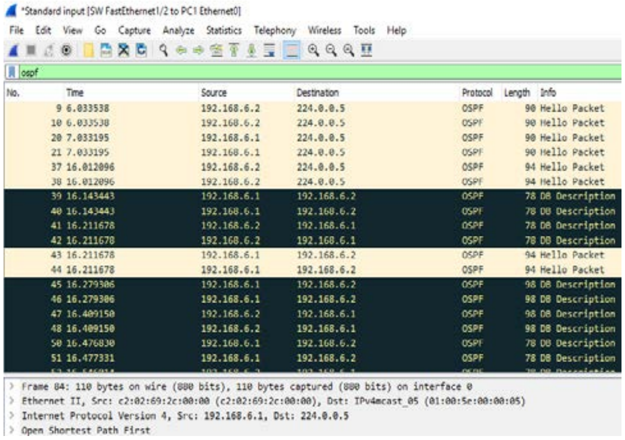
A. Open Shortest Path First Routing Protocol Result

The convergence startup time is the duration of the first Database description till the last link-state acknowledgement packets are displayed in Fig. 3.

The database description provides information of each router in the network. Wireshark monitors the entire exchange of the packets in the different topology designed in the network. The results show the Hello, Database Description packets when implementing OSPF (Simulator). The link-state request, updates, and acknowledgment operate synonymously. The link-state request sends a specific request to nodes in the network when the request is sent. The network updates itself to identify the changes in topology and reply with an acknowledgment.

B. Enhanced Interior Gateway Routing Protocol Result

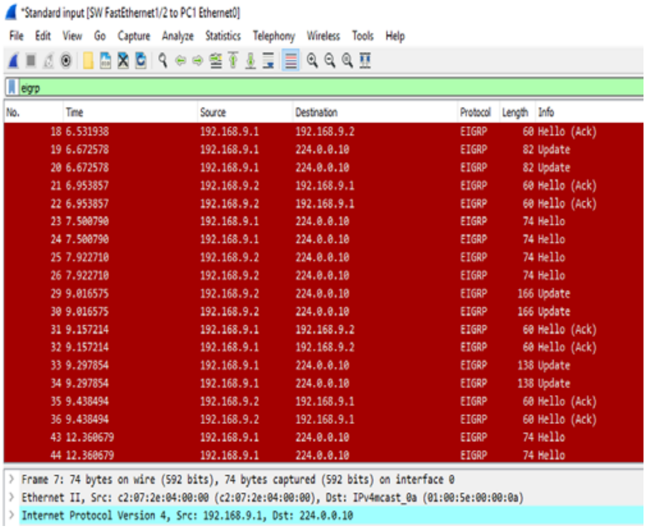
The Enhanced Interior Gateway Routing Protocol packets are different from the ones obtained in Open Shortest Path First Routing Protocol. The convergence duration occurs when the hello packets have been distributed in the entire nodes in the network followed by the updates and acknowledgement packets. The monitoring software (Wireshark) captures all the Hello, updates and acknowledgment packets in the network that leads to convergence between the routers. Enhanced Interior Gateway Routing Protocol operates such that whenever there are changes in the link or nodes in the network, it will send out a query packet that will have an equivalent reply. This occurs because of a shutdown or failure in any of the links or nodes (routers) in the network. The result is shown in Fig. 4.



```
ospf
No.  Time        Source        Destination   Protocol  Length  Info
  9 6.033538     192.168.6.2   224.0.0.5     OSPF        90 Hello Packet
 10 6.033538     192.168.6.2   224.0.0.5     OSPF        90 Hello Packet
 20 7.033195     192.168.6.1   224.0.0.5     OSPF        90 Hello Packet
 21 7.033195     192.168.6.1   224.0.0.5     OSPF        90 Hello Packet
 37 16.012096    192.168.6.2   224.0.0.5     OSPF        94 Hello Packet
 38 16.012096    192.168.6.2   224.0.0.5     OSPF        94 Hello Packet
 39 16.143443    192.168.6.1   192.168.6.2   OSPF        78 DB Description
 40 16.143443    192.168.6.1   192.168.6.2   OSPF        78 DB Description
 41 16.211678    192.168.6.2   192.168.6.1   OSPF        78 DB Description
 42 16.211678    192.168.6.2   192.168.6.1   OSPF        78 DB Description
 43 16.211678    192.168.6.1   192.168.6.2   OSPF        94 Hello Packet
 44 16.211678    192.168.6.1   192.168.6.2   OSPF        94 Hello Packet
 45 16.279306    192.168.6.1   192.168.6.2   OSPF        98 DB Description
 46 16.279306    192.168.6.1   192.168.6.2   OSPF        98 DB Description
 47 16.409150    192.168.6.2   192.168.6.1   OSPF        98 DB Description
 48 16.409150    192.168.6.2   192.168.6.1   OSPF        98 DB Description
 50 16.476830    192.168.6.1   192.168.6.2   OSPF        78 DB Description
 51 16.477331    192.168.6.1   192.168.6.2   OSPF        78 DB Description

> Frame 84: 110 bytes on wire (880 bits), 110 bytes captured (880 bits) on interface 0
> Ethernet II, Src: c2:02:69:2c:00:00 (c2:02:69:2c:00:00), Dst: IPv4mcast_05 (01:00:5e:00:00:05)
> Internet Protocol Version 4, Src: 192.168.6.1, Dst: 224.0.0.5
> Open Shortest Path First
```

Fig. 3. Wireshark Results of updates and Acknowledgement Packets for OSPF.



```
eigrp
No.  Time        Source        Destination   Protocol  Length  Info
 18 6.531938     192.168.9.1   192.168.9.2   EIGRP       60 Hello (Ack)
 19 6.672578     192.168.9.1   224.0.0.10    EIGRP       82 Update
 20 6.672578     192.168.9.1   224.0.0.10    EIGRP       82 Update
 21 6.953857     192.168.9.2   192.168.9.1   EIGRP       60 Hello (Ack)
 22 6.953857     192.168.9.2   192.168.9.1   EIGRP       60 Hello (Ack)
 23 7.500790     192.168.9.1   224.0.0.10    EIGRP       74 Hello
 24 7.500790     192.168.9.1   224.0.0.10    EIGRP       74 Hello
 25 7.922710     192.168.9.2   224.0.0.10    EIGRP       74 Hello
 26 7.922710     192.168.9.2   224.0.0.10    EIGRP       74 Hello
 29 9.016575     192.168.9.2   224.0.0.10    EIGRP      166 Update
 30 9.016575     192.168.9.2   224.0.0.10    EIGRP      166 Update
 31 9.157214     192.168.9.1   192.168.9.2   EIGRP       60 Hello (Ack)
 32 9.157214     192.168.9.1   192.168.9.2   EIGRP       60 Hello (Ack)
 33 9.297854     192.168.9.1   224.0.0.10    EIGRP      138 Update
 34 9.297854     192.168.9.1   224.0.0.10    EIGRP      138 Update
 35 9.438494     192.168.9.2   192.168.9.1   EIGRP       60 Hello (Ack)
 36 9.438494     192.168.9.2   192.168.9.1   EIGRP       60 Hello (Ack)
 43 12.360679    192.168.9.1   224.0.0.10    EIGRP       74 Hello
 44 12.360679    192.168.9.1   224.0.0.10    EIGRP       74 Hello

> Frame 7: 74 bytes on wire (592 bits), 74 bytes captured (592 bits) on interface 0
> Ethernet II, Src: c2:07:2e:04:00:00 (c2:07:2e:04:00:00), Dst: IPv4mcast_0a (01:00:5e:00:00:0a)
> Internet Protocol Version 4, Src: 192.168.9.1, Dst: 224.0.0.10
```

Fig. 4. Wireshark Results of updates and Acknowledgement Packets for EIGRP.

[5] https://gns3.com/

[6] https://www.wireshark.org/

[7] https://www.netacad.com/courses/packet-tracer

VI. Result and Discussion

In the design of the EIGRP and OSPF Routing Protocols, two topologies are examined (Star and Partial Mesh) which are widely used in today's networking for both software and hardware implementation.

A. OSPF Routing Protocol (Star Topology) Software

The average convergence duration at the start of the network and when one of the links fail of star topology using OSPF increases as the number of resources increases. The results shows that when the number of routers is less than ten (10), it takes less than 15.5 milliseconds for the routers to converge, the same occurs when one of the links fails in any of the designs. Meanwhile, when a new link is installed in the network, it requires a longer time to converge, though this does not transpire in all cases, see Fig. 5.

B. OSPF Routing Protocol (Mesh Topology) Software

The results show that when a new link is added to the network, it requires less time for the network to converge. More than 90% of the time, it requires less than 20 milliseconds for the network to converge when new nodes are added. The time required for the network to converge at a start and when a link fails are relatively the same. Moreover, it takes less than 9% of the time for the difference between the convergence time at the start and when a link fails in each of the numbers of resources. Furthermore, the result obtained shows that mesh topology takes a longer time to converge at the start and when there is a link failure especially as the number of routers increases in the network, see Fig. 6.

C. OSPF Routing Protocol (Star Topology) Hardware

The result is the same as the simulated result obtained in (A) above, just a slight difference which is negligible. When the number of routers is twelve (12) the convergence time is 19.85 milliseconds which is slightly higher than the simulator results with 1.2 milliseconds. Whereas when a link fails, the time required for it to re-converge is 20.23 milliseconds which is approximately 20.0 milliseconds. When a new node is added to the network the time difference between the simulated and hardware result is 3 milliseconds. The GNS3 result is slightly different from the hardware result with about 10% which might be from errors of configuration or computer bugs that is associated with GNS3, see Fig. 7.

D. OSPF Routing Protocol (Mesh Topology) Hardware

The time taken for the convergence duration is high when compared with the simulator results obtained. When a new link is added to the network, it takes 17.85 milliseconds for the network to converge while in the simulator is 8.98 milliseconds. Furthermore, the results obtained in the hardware simulation have a stable slope and consistent trend than the results obtained using GNS3. In each of the network scenarios or number of resources the convergence duration, the time when a link fails, and new links are added to the network is higher in the hardware implementation than the simulator results obtained, see Fig. 8.

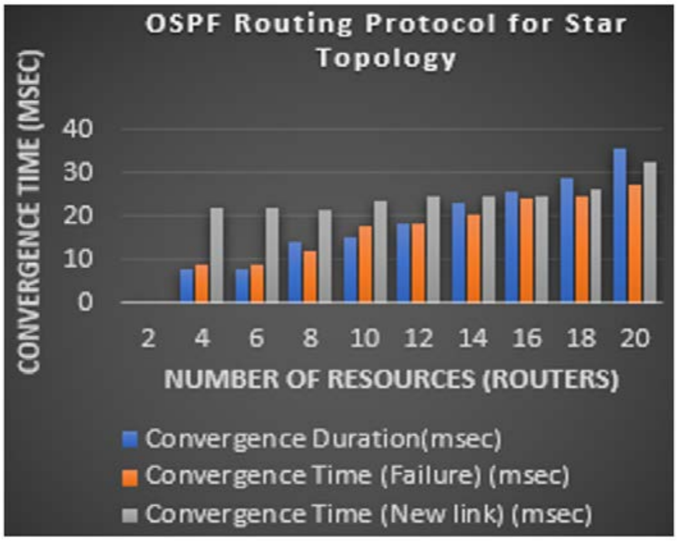


Fig. 5. OSPF Results for Star Topology – Software.

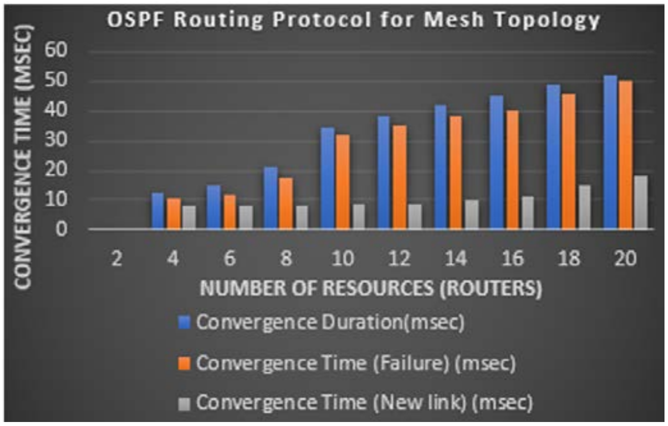


Fig. 6. OSPF Results for Mesh Topology – Software.

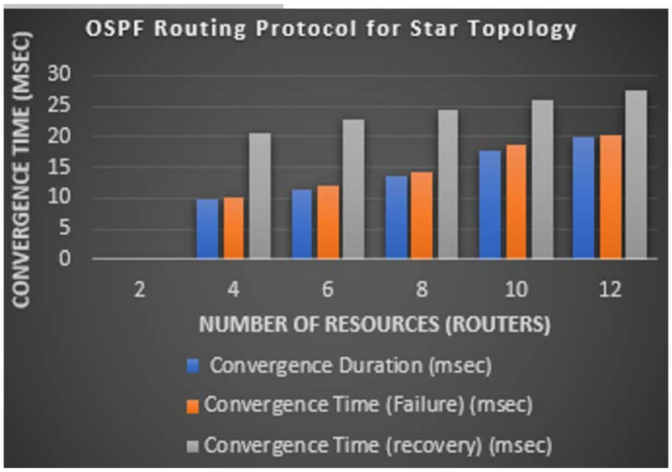


Fig. 7. OSPF Results for Star Topology (Hardware).

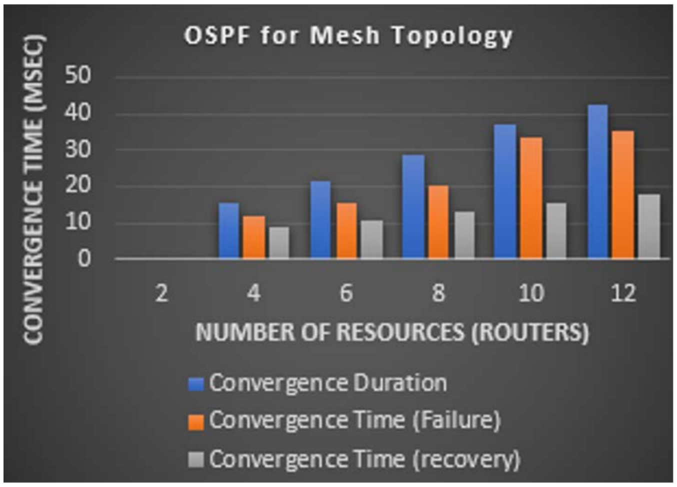


Fig. 8. OSPF Results for Mesh Topology (Hardware).

E. EIGRP (Star Topology) Software

The result obtained is consistent all through the different convergence time. It takes an average of 5.25 milliseconds for the network to converge, the same time it requires when a link fails, or a new link is added to the network. Furthermore, it takes an average of 26.25 milliseconds for the network to converge when a link fails or when a new link is added to the network when the number of resources is twenty 20 (maximum). The results indicate that the higher the resources the slower the network takes to converge in each network scenario. The average convergence time in EIGRP is faster compared to the OSPF routing protocol when using the same number of resources, settings, and devices, see Fig. 9.

F. EIGRP (Mesh Topology) Software

The results described the mesh topology to have the best convergence duration, time when a link failure and new links are added to the network. It takes an average of 1.8milliseconds for the network to converge when the number of resources is 4. The same time (1.8milliseconds) is required when a link fails, or a new link is added to the network. The results describe that as the number of resources increases, the convergence time increases representing a straight-line graph, see Fig. 10. Furthermore, the results described that EIGRP has a higher convergence period or performance in all the network scenarios than any of the topologies implemented in simulated and hardware devices.

G. EIGRP (Star Topology) Hardware

Considerably, it takes a longer time for the network to converge compared to when a link is shut, or a new link is added to the network. It requires an average of 17.54milliseconds for the network to converge at the beginning while it requires less than 14.00milliseconds to converge when a link failure or a new link is added to the network, see Fig. 11. This might be because of the implementation of virtual subnets (loop back) in the star topology. On the average, the results obtained from the hardware implementation are better than the simulator with about 10% in terms of convergence duration, the time when a link fails, and a new link is added to the network.

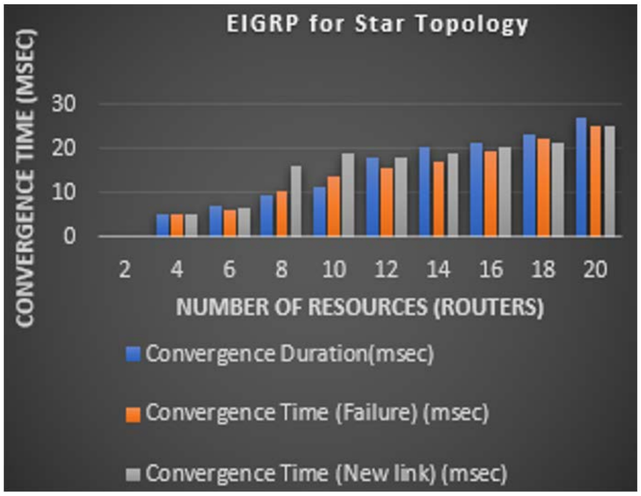


Fig. 9. EIGRP Result for Star Topology (Software).

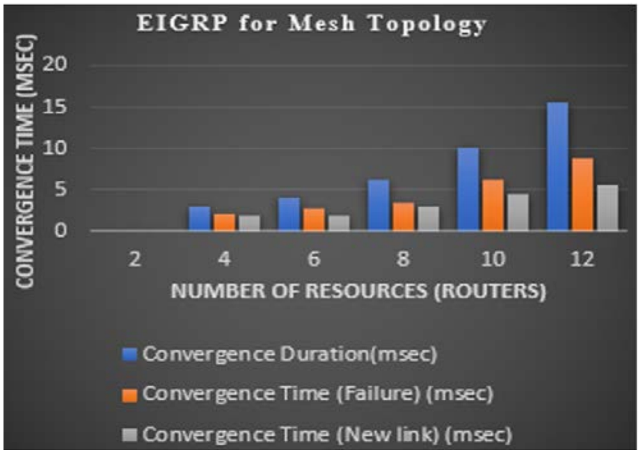


Fig. 10. EIGRP Result for Mesh Topology (Software).

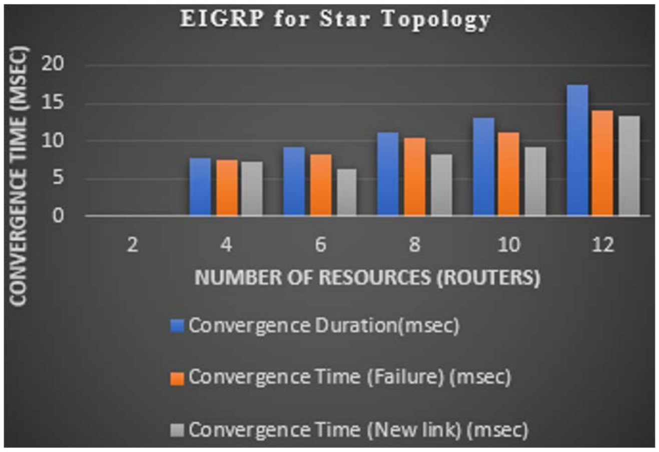


Fig. 11. EIGRP Result for Star Topology (Hardware).

H. EIGRP (Mesh Topology) Hardware

The mesh topology using the EIGRP provides the best performance for convergence duration, the time when a link fails, and new links are added to the network. The hardware results obtained are not different from the simulator results. The slight difference occurs in the convergence duration with about 3.0 milliseconds when the network is flooded with twelve routers. Because the convergence time of the failure of a link and when a new link is added does not change. The results indicate that EIGRP performs better in convergence time since both software and hardware implementation provides less than 10.0milliseconds for the network to converge when a link fails, and a new link added to the network, see Fig. 12.

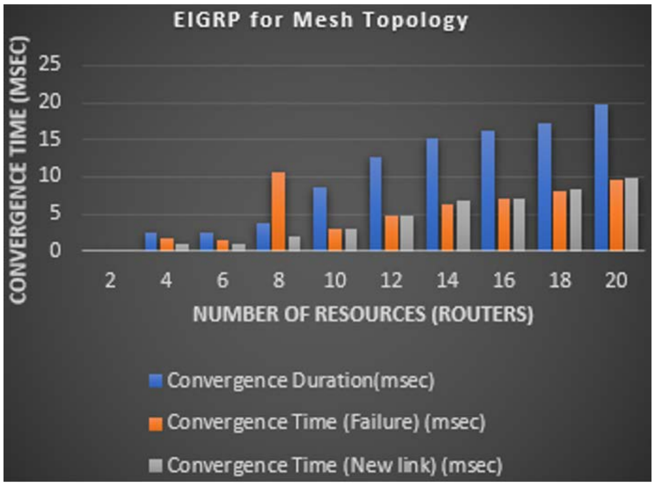


Fig. 12. EIGRP Result for Mesh Topology (Hardware).

VII. Conclusion

A reflective summary of these experiments enables the justification and analyses of EIGRP and OSPF routing protocol using GNS3 and Cisco IOS devices using different network scenarios. The EIGRP uses DUAL which helps in recalculating a given route globally to avoid routing loop, so it has the attributes of a link state and distance vector routing protocol. This ensures a faster convergence time in all the topologies when using GNS3 and Cisco IOS devices. This experiment contributes to the existing knowledge by identifying that: mesh topology has the best topology for convergence time ahead of star topology. Based on the result obtained, it clearly states that hardware implementations of routing protocol are better than using a network simulator. Because the network simulator has computer bugs, runtime failure, updates and simulation errors which influence the results obtained when implementing EIGRP and OSPF routing protocol. The conclusion described in the network scenarios indicates that EIGRP has a higher performance in convergence duration, the time when a link fails, and new links added to the network than OSPF routing protocol. This is because EIGRP does not perform routing updates that require longer time compared to the OSPF routing protocol.

Also, this research cannot be limited to only OSPF and EIGRP, further analysis to BGP comparison with the above protocol to see their different performance will be a good research. Also, with the transition from IPV4 to IPV6, research on how the protocol changes or adaptation in terms of convergence time with the versions of IPV4 and IPV6 can be examined. Finally, the Latency and Quality of Service are vital areas of research in both EIGRP and OSPF routing protocol.

References

[1] Nefkens P-J. Phase Three: Design, Deploy, and Extend. In: Transforming Campus Networks to Intent-Based Networking. Cisco Press; 2019.

[2] Lamle T. CCNA® Cisco Certified Network Associate. Sybex; 2011.

[3] Hill S. Distance Vector Routing Protocols. Greater Manchester; 2016.

[4] Lacoste R, Edgeworth B. Chapter 2: EIGRP. In: CCNP Enterprise Advanced Routing ENARSI 300-410 Official Cert Guide. Cisco Press; 2020.

[5] Tanenbaum AS, Wetherall DJ. Computer Networks. 2010.

[6] Evans J, Schneider G, Pinard K. The Internet Illustrated. 6th edition. London: Course Technology Inc; 2009.

[7] Dey GK, Ahmed MM, Ahmmed KT. Performance analysis and redistribution among RIPv2, EIGRP & OSPF Routing Protocol. In: 1st International Conference on Computer and Information Engineering, ICCIE 2015. Institute of Electrical and Electronics Engineers Inc.; 2016. p. 21–4.

[8] Larrea Luzuriaga R, Jimenez J, Sendra S, Lloret J. Comparative Study of Routing Protocols in Ring Topologies using GNS3. Valencia; 2016. 38–44 p.

[9] Sirika S. Performance Evaluation of Dynamic Routing Protocols for Real time application. Int J Eng Trends Technol. 2016 Feb;Volume 32:328–37.

[10] Anibrika BS, Adamu M, Franklin A, Asante M. Performance Analysis of Enhanced Interior Gateway Routing Protocol (EIGRP) Over Open Shortest Path First (OSPF) Protocol with Opnet. Int J Adv Comput Sci Appl. 2016 May;7.

[11] Whitfield R, Zhu SY. A comparison of OSPFv3 and EIGRPv6 in a small IPv6 enterprise network. Int J Adv Comput Sci Appl [Internet]. 2015;6(1):162–7. Available from: http://hdl.handle.net/10545/620915.

[12] Hellberg C, Greene D, Boyes T. Designing a Triple-Play Backbone. In: Broadband Network Architectures: Designing and Deploying Triple-Play Services. 2007.

[13] Diaz L. Diffusing Update Algorithm or DUAL. In: CCNA Routing and Switching 200-125 Certification Guide. Packt Publishing; 2018.

[14] Usman A. Cisco Packet Tracer Overview [Internet]. 2014 [cited 2020 Jun 21]. p. 9–14. Available from: https://www.slideshare.net/AliUsman10/cisco-packet-tracer-overview.

[15] Henry T. Network Topologies. In Rhodes Island; 2017. Available from: https://homepage.cs.uri.edu/~thenry/csc414/72_NetworkTopo_TOC.pdf.

[16] Mirzahossein Michael Nguyen Sarah Elmasry K. Analysis of RIP, OSPF, and EIGRP Routing Protocols using OPNET [Internet]. 2013 [cited 2020 Jun 21]. Available from: www.sfu.ca/~mtn9/Group5.html.

[17] Wallace K. Split Horizon | CCNA Routing and Switching 200-125 [Internet]. Pearson IT Certification 2016; 2016 [cited 2020 Jun 21]. Available from: https://learning.oreilly.com/videos/ccna-routing-and/9780134580715/9780134580715-CCNA_03_03_05.

[18] Kaur S, Roohie NM. (PDF) Performance Analysis of Interior Gateway Protocols. Adv Res Electr Electron Eng [Internet]. 2014 [cited 2020 Jun 21];01(01):59–63. Available from: https://www.researchgate.net/publication/303812574_Performance_Analysis_of_Interior_Gateway_Protocols.

[19] Coleman A, Bombal D, Duponchelle J. Getting Started with GNS3 - GNS3 [Internet]. 2020 [cited 2020 Jun 20]. Available from: https://docs.gns3.com/1PvtRW5eAb8RJZ11maEYD9_aLY8kkdhgaMB0wPCz8a38/index.html.

[20] Sharpe R, Warnicke E, Lamping U. Wireshark User's Guide Preface Foreword [Internet]. 2020 [cited 2020 Jun 20]. Available from: https://www.wireshark.org/docs/.

Estimate the Total Completion Time of the Workload

Muhammad Amjad*[1], Waqas Ahmad[2], Zia Ur Rehman[3], Waqar Hussain[4], Syed Badar Ud Duja[5], Bilal Ahmed[6], Usman Ali[7], M. Abdul Qadoos[8], Ammad Khan[9], M. Umar Farooq Alvi[10]
College of Information and Computer Taiyuan University of Technology
Taiyuan, Shanxi, China

***Abstract*—The business intelligence workload is required to serve analytical process. The data warehouses have a very large collection of digital data. The large collection of digital data is required to analytical process within the perplexing workload. The main problem for perplexing workload is to estimate the total completion time. Estimate total completion time is required when workload is executed as a batch of queries. To estimate the queries according to their interaction aware scheme because queries are run in batches. The database administrators often require to perceive how much longer time for business intelligence workloads will take to complete. This question ascends, when database administrator entails to accomplish workloads within existing time frame. The database system executes mixes of multiple queries concurrently. We would rather measure query interactions of a mix than practiced approach to consider each query separately. A novel approach as a estimate framework is presented to estimate running time of a workload based on experiment driven modeling coupled with workload simulation. An estimation framework is developed which has two major parts offline phase and online phase. Offline phase collects the experiments sampling of mixes which has different query types. To find the good accuracy for estimating the running time of the workload by evaluation with TPC-H queries on PostgreSQL.**

Keywords—Query interactions; estimate time; running time

I. Introduction

OLAP workload has long-running queries that has to execute database system at different time period repeatedly. These batches of queries execution time range sometimes from minuts to hours. The database administrator wants to accomplish business intelligent workloads within time frame but due to indeterminacies of queries execution time can't fulfill her requirement. Resource contentions are a reason which effects response time of a query. Therefore, to measure query interactions in a mix is essential, a phenomenon that query execution might be hastened or hindered by parallel queries [23]. For example the performance of Q_{18} and Q_5 describe in the three mixes m_2, m_3 and m_4. The m_2 presents the positive interaction for Q_{18}. Q_{18} has the average response time 609 seconds while run alone in the system is 624 seconds. On the others hand Q_{18} suffers in mix m_3 due to negative interaction. Q_{18} has the average response time 707 seconds in mix m_3. m_4 is also a positive interaction for Q_5. We need to capture these interactions.

If query interaction in a mix is measured then a database administrator can adjust business intelligent workload within time-bound without flawlessly. The state of the art does not provide any method to estimate the total completion time of a workload. To predict query execution time that is not only use for estimate the total completion time, it is also useful for database other management tasks, sizing, progress monitoring, admission control and query scheduling [2-5]. Recently most of the work focuses on estimating the time for independent query [6-9]. whereas a very little work studies to predict the time for simultaneously running queries [10]. The database systems most often allow simultaneously running queries. Therefore, to estimate execution time for concurrent queries are required. To estimate simultaneously running queries are more important than for independent queries.

The approaches are investigated to building estimation model for estimating the response time of a query running concurrently with other queries. Specifically, the model focuses to the following scenario. The database systems constantly run a mix of *M* queries simultaneously. Whenever a query is finished execution and exits, the database systems arbitrary manner select a query from the queue to form a new mix. The model is required for estimating the response time of a query at any time point of its running.

In this paper experiment driven approach is used to take into account the query interactions. Experiment driven approach is attainment to build performance static model which estimate the query response time. A dynamic model is manipulate such performance static model for estimating the response time of the newly form mixes. A relevant work uses machine learning technique to estimate performance metrics for queries [7], but authors focus at the single query running in the database system and our motive is concurrently running queries.

Our contributions can be concluded as follows:

- A performance static model is proposed to estimate the query response time in a mx.
- A dynamic model is proposed to manipulte the performance static model for estimating newly formed mixes.

To meaure the impact of the query interactions in a mix, queries are used from TPC-H benchmark with a database system extent of 10GB operating on PostgreSQL. Table I shows average running time of the TPC-H queries in the database system. Table II shows the number of each type of queries in a mix.

The rest of paper organized as follow. Section II present related work . A framework is developed in Section III. Section IV presents evaluation of this approach and conclude in Section V.

TABLE I. RUNNING TIME t_j OF SINGLE QUERY TYPES

Query type	Q_1	Q_4	Q_5	Q_6	Q_8	Q_{10}	Q_{12}	Q_{14}	Q_{18}
Runtime tj (sec)	5085	1414	585	578	598	71	866	573	624

TABLE II. THE AVERAGE RUNNING TIME FOR SIMULTANEOUSLY QUERIES

	Mix	$\mathbf{M_1}$	$\mathbf{M_2}$	$\mathbf{M_3}$	$\mathbf{M_4}$
$\mathbf{Q_1}$	A_{ij}	1	0	0	0
	N_{ij}	5193	0	0	0
$\mathbf{Q_2}$	A_{ij}	1	0	0	1
	N_{ij}	1620	0	0	1620
$\mathbf{Q_5}$	A_{ij}	0	0	0	1
	N_{ij}	5193	0	0	443
$\mathbf{Q_6}$	A_{ij}	1	0	0	0
	N_{ij}	602	0	0	0
$\mathbf{Q_8}$	A_{ij}	0	1	0	1
	N_{ij}	5193	707	0	707
$\mathbf{Q_{10}}$	A_{ij}	0	0	1	0
	N_{ij}	5193	0	78	0
$\mathbf{Q_{12}}$	A_{ij}	0	1	0	0
	N_{ij}	5193	1133	0	0
$\mathbf{Q_{14}}$	A_{ij}	0	0	1	0
	N_{ij}	5193	0	607	0
$\mathbf{Q_{18}}$	A_{ij}	0	1	1	0
	N_{ij}	0	609	609	0

II. REALTED WORK

A. Literature Review

The main method of the analytical model is to obtain the query response time by performing a detailed analysis of the query process of the database system. The core research content is (1) the speed at which the system executes the query (2) the amount of data that the query needs to process. For the analytical model of a single query, the response time of the query can be obtained by clarifying the above two points. For the analytical model of parallel query, additional research is needed (3) the contention pattern of resources between parallel queries, and the mode is explicitly described. The research process of the analytical model is also the research progress of the above three core contents. In 2004, Chaudhuri et al. [16] and Luo et al. [24] published a paper on query progress indicators at SIGMOD 2004, and studied the progress indicator and the query response time prediction model based on the analysis method.

The workload prediction has to do with query interaction in a mix that is known as building query progress indicator, statistical prediction models and analytical prediction models. Progress indicator is a tool which shows the percentage progress of a running query [15, 16]. The progress indicators fundamentally partition a query plan do on different phases and updated the query progress information based on the execution information consistently.

Ahmad et al. [10] study the problem for estimating simultaneously query response time. In [13], the authors propose an experiment driven approach for sampling. In the paper the authors use Gaussian process as the particular statistical model. The key restraint of the work is assumed static workload that is not practically. According to our reading, we indicate the concurrent query response time estimation problem under dynamic workload.

In database system the research communities have gained substantial interest to predicting query execution time [11,12, 13,14]. The current query response time prediction model is divided into analytical type and statistical type. Analytical modeling predicts response time by studying the query execution process. Statistical modeling uses machine learning methods to model query response time. This method can strike a balance between model complexity and usability. The static modeling technique is employed to estimate running time for a query in a database system [17,18,19]. We employed our developed Gscheduler by solving a linear programming problem and also for the estimation model to work [1].

III. THE FRAMEWORK

Estimation framework is described in this Section. In first instance, estimation problem is described, then solution is described and would be provided some analysis.

We present problem definition in Subsection A. Subsection B presents the structure of our defined workflow. Subsection C presents performance static model. Subsection C presents dynamic model.

A. The Problem Definition

In order to meet and maintained the peak performance of our proposed scheduler in our prior work, that decrease total completion time of a business intelligent (BI) workload [1]. Now to solve another workload management problem that estimate the total running time of a workload. The database administrators want to knowing how much the workload will expect to run.

The database consistently runs query mixes according to multiprogramming level M queries draw as $W = \{< q_i, w_i > | i = 1,2, \ldots, N\}$, w_i represents the number of queries q_i. Whenever a query is finished execution, the database system selects the query from workload W to form a new mix, and estimate leftover running time of newly formed mix. We require to find an approach which give us estimate running time of newly form mixes.

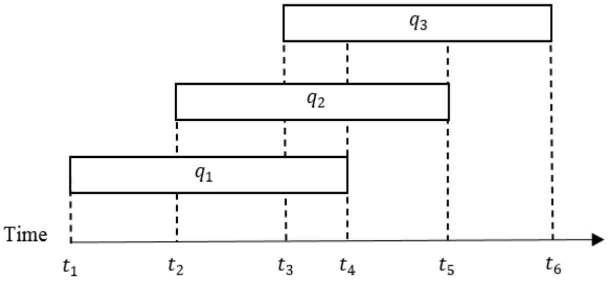


Fig 1. Illustration the Estimation Running Time Problem.

An example is employed to explain the estimation problem in Fig. 1, three queries q_1 , q_2 and q_3 that are concurrently running and arrive at time denoted as t_1 , t_2 and t_3. In this scenario, there are three estimation problems. At t_1 we need to estimate the running time of q_1. The estimation requires the information of the upcoming q_2, q_3 which is not available at t_1. At t_2 , q_2 join and need to make an estimation for both q_1 and q_2. Actually q_1 that has been running for some time, we require its remaining running time. The estimation requires the knowledge that q_3 will arrive which is unavailable at t_2. The same argument can be further applied to the estimation for q_1 , q_2 and q_3 at t_3. For example, let M be a mix of n queries $M = \{q_1, q_2, \ldots\ldots, q_n\}$ which are concurrently running. Let me know, s_0 is a start time of n queries and e_i is the end time. $T_i = e_i - s_0$ to be the execution time of q_i is defined. An estimation model is required to concern this problem. The estimation problem in Fig. 1 is generated by setting $M = \{q_1, q_2, q_3\}$ and $s_0 = t_3$.

B. The Structure of the Framework

The structure of the framework is initiated for batch workload. The structure contains a dynamic model that can estimate response time of the mixes. This nontrivial task is accomplished with dynamic model whosoever can estimate the running time of queries to manipulating performance static model.

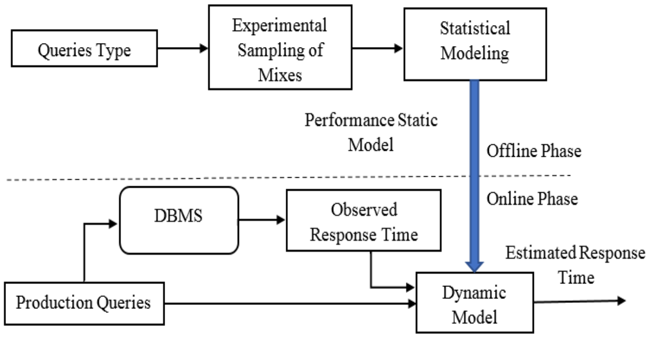


Fig 2. An Overview of the Framework.

Fig. 2 defines the overall workflow of the framework which consists of two major components, that are offline phase, and online phase. The workflow is used by a database administrator who executes the batch workloads repeatedly and need to estimate the total completion time of a workload. The database administrator provides a queries type q_i as input to the offline phase. Next, the detail description is described of the performance static model and the dynamic model.

C. The Performance Static Model

The impact of query interactions have been defined capturing by experimentally measuring the average completion time of various queries type. A new approach is proposed which designed and conducted experiments for possible query mix sampling. The experiment is run to chose query mix sample. Collecting the data from all experiments as a query mix model. No prior assumptions are required for this method about the working of the database system. For example, the database consistently executes queries concurrently drawn from the set of N queries in a mix, $m = \{q_1, q_2, \ldots\ldots, q_n\}$. Every executed query mix in the sample called a known mix. The response time of each query is recorded as the sample data set. A model is developed to estimate the newly formed unknown query mixes.

Through experiment driven approach, sampling experiments are required to collect data for static model. The sampling policy gives feasible query mixes that provides a feasible point for an offline phase. The method perks that the instances can be updated incessantly and improved model performance. Our workload generator generates query mixes sample. For collecting the samples, the workload setting is generated by client coordinator, the MPL=3. 19 hours is taken to run these experiments for 504 different query mixes of workload.

For the samples data, to develop a sampling algorithm is required, and then an appropriate regression model is employed, which gives accuracy in estimating query completion time in a mix. Now, a sampling algorithm is described for the efficiency of the sample data.

There are many types of sampling techniques. To choose the possible sampling from the 504 set of query mixes. The random sampling is employed, but according our modeling perspective, it is inefficient. The drawback of the random sampling is wide-ranging when to require to learn a good model. When we increase the queries type, the same sample data may be repeated unnecessarily that's why the sample data may be larged. For save the processing time, minimize the samples data is required for a large number of queries type.

For this purpose, Latine Hypercube Sampling (LHS) is used. McKay et al. [20] propose LHS that is suitable for designing of computer experiments. Stein et al. [21] prove that LHS is a powerful and useful method. LHS gives better coverage than random sampling. LHS comes from the family of space-filling designs and performs well in practice [22].

The constraints are harded that the number of concurrent query instances in a sampled mix would not be exceeded from the fixed multiprogramming level. The requirements can't be fulfilled by the simple LHS technique. For making it interactions aware and fixed conditions on MPL, an algorithm is needed to develop for collecting those sample data which have minimum estimation error and also to satisfy fixed instances of query types as fixed MPL.

The Task: A set of samples n mixes for the given query types T, MPL= M, interaction level = k, and permutation matrix = P_k.

The method:

- Let's P_k of size $n \times T$ by LHS, $P_k$is a query mix.
- P_k at random set values of $T - k$, the column set as 0 to k make the level of its interactions. The column does not exceed as fixed M.
- We want to cover the interactions level of k within n samples data. The n mixes that are allowed for k. We pick mixes from P_k.

The algorithm fulfills our requirements to sample mixes, which covers interactions level and makes the fixed Multiprogramming level. When a query finish execution and exits the new query comes and forms new mix constantly in the database, therefore a model is required which uses this performance static model and estimates the time of the newly formed mix. Next, we present detail of the model namely called dynamic model for estimating query response times.

D. The Dynamic Model

The dynamic model is very powerful for database system administration. The important feature of the dynamic model is a plugable incorporating of the scheduling method. The dynamic model is used for typical scheduling methods First-Come, First-Served, Shortest-Job-Next scheduling, and a variety of sophisticated scheduling methods.

Dynamic estimation problem: The workload, $W = \{< q_i, w_i > | i = 1,2,3, \ldots, N\}$, is run according to multiprogramming level M queries. w_i represents the number of queries q_i in a mix. Whenever a query is finished and exits, the database system selects the query from workload W to form a new mix, the model is estimated the remaining execution time of the mix according to the new mix.

For example, the query q_1 in Fig. 3, when m_1 start at time t_0. The model will estimate the leftover time of q_1 according to the query mix (q_1, q_3, q_6), from t_0 to t_4. When q_3 exits and q_4 comes in to form new mix m_2, the model estimates the remaining execution time of q_1 according to mix (q_1, q_6, q_4), this is the time from t_1 to t_4. The estimation process for all other queries is the same.

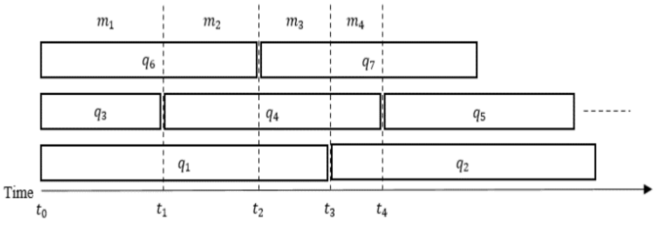


Fig 3. Different Mixes of Three Queries Running Continuously.

The dynamic model is similar to the progress indictors [5, 19, 26], but the dynamic model uses different modeling approach. (i) The dynamic model is activated at the exits of a query, but progress indicators predict periodically. (ii) The progress indicators are used in a single query scenario. The existing parallel progress indicators [5] is a simple extension of the classic progress indicators [17, 24]. The interactions between queries are not considered for the parallel progress indicators.

When a query exits, the dynamic model estimates the remaining execution time of remaining running queries using Equation as follows:

$$\hat{t}^{r}_{<q,m_s>} = t^I < q, m_s > \left(1 - tanh(\textstyle\sum_{l=1}^{s-1} \frac{t^e_{<q,m_l>}}{t^I_{<q,m_l>}})\right) \qquad (1)$$

$\hat{t}^{r}_{<q,m_s>}$, is shows the estimated remaining time of q with mix m_s. $t^I < q, m_s >$, is shows the running time of query q running in m_l. $t^e_{<q,m_l>}$, is the elapse time of query q in m_l. Equation 1 estimates the time needs to exit query q running in m_s . m_s estimates the starting time of the query instance. According to Fig. 3, suppose we need to estimate the remaining time of q_1 at t_2. The elapsed time of q_1 is t_0 to t_2, and the work done for q_1 is $\tanh\left(\frac{t^e_{\langle 1,\{6,3.1\rangle}}{t^I_{\langle 1,\{6,3.1\rangle}} + \frac{t^e_{\langle 1,\{6,4.1\rangle}}{t^I_{\langle 1,\{6,4.1\rangle}}\right)$. To estimate the remaining time of q_1 in mix (7, 4, 1), we multiply $t^I_{\langle 1,\{7,4.1\rangle}$ with $1 - tanh\left(\frac{t^e_{\langle 1,\{6,3.1\rangle}}{t^I_{\langle 1,\{6,3.1\rangle}} + \frac{t^e_{\langle 1,\{6,4.1\rangle}}{t^I_{\langle 1,\{6,4.1\rangle}}\right)$.

The dynamic model refreshes the state of unfinished queries at the end of the mix. When scheduling picks new query for making new mix, the dynamic model change the time state with the help of perfromance static model. That procedure continues running til all queries in the workload are completed.

$$L_w = \sum_{i=1}^{|W|-M+1} I_i$$

The workload estimated completion time is add the lengths for all mixes come upon throughout the imitation.

IV. Experiment Evaluation

This section presented experiment evaluation of the proposed approach. The estimation accuracy we measure in terms of mean relative error. Which is defined as:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{|T_i^{est} - T_i^{act}|}{T_i^{act}}$$

The number of testing queries is *N*. The estimated and the actual execution time for q_i are T_i^{est} and T_i^{act}. We measured the estimation approaches as well. In subsection 5.7.1 gives out the experimental settings. In subsection 5.7.2 define the overall accuracy. In subsection 5.7.3 define the scheduling performance.

A. Experimental Setup

Environments. The software and hardware is described for using in our experiments. A machine with Intel E3500 2.7GHz CPU and 4GB RAM is used for experiment. PostgreSQL database is run under Window 2007x64. whole configurations of Postgres are the default and turned off entire the statistics and tuning tools. TPC-H scale factor 10, denoted by 10GB is used. The configuration advisor of the PostgreSQL ensures the configuration parameters are well acquainted.

Workload: Table III shown the workload of eighty instances of nine types of queries, which are chosen pursuant to the *TPC-H* and run on 10 GB database system. We limit the workload size and MPL by virtue of the long-term of queries in database system. Recall from Fig. 4 that 80 query workloads can take more than 6 hours to complete.

Methodology: We generated 80 workloads due to choice of queries type, use our Gscheduler algorithm [1] with other two scheduling algorithms as FCFS, SJF. The completion times limit from 6 hours to more than 8 hours. We comparison acutal time as *act* and estimated completion time as *est.* the estimation error is computed as.

$$estimation\ error = \frac{|est - act|}{act} \times 100$$

B. Overall Accuracy

The overall accuracy of our estimations. The relative estimation error in all 80 workloads in our experiment. In one case we use performance static model for *First-Come-First-Served*, second, we use *Shortest-Job-Next* and third we use *Gschedule*r. Amjad et.al. defined Gscheduler is to schedule query mixes for a given query workload in order to minimize *W*'s total completion time. When query finished in query mix the Gscheduler chose query from the workload which will take minimum time for execution. The estimation errors in all cases are less than 20% of the 80% time. Aginst the 80 workload queries, these results show that our framework is accurate for estimating total completion time of the workload.

The database administrator can accurate estimations for future workloads in a database with the help of one time samples.

TABLE III. WORKLOAD OF QUERIES

SF	Q_1	Q_4	Q_5	Q_6	Q_8	Q_{10}	Q_{12}	Q_{14}	Q_{18}
10	3	12	6	13	8	15	6	13	4

C. Scheduling Performance

Here, our Gscheduler is compared with two other state of the art scheduling approaches from the literature. The scheduling approaches are compared without any assistance from queueing models.

The performance of scheduling algorithms in this section and compare with each other, *First-Come-First-Served*, *Shortest-Job-Next* and *Gscheduler* scheduling algorithms. We represent the number of similar mixes used for the performance static model to estimate. The workload *W* contains 80 queries, as shown in Table III.

First-Come-First-Served is sensitive to the arrival order of queries in *W*, so we sequentially generate a sequence of queries and report the completion time of *W*. *Shortest-Job-Next* is not sensitive to the arrival order we select the shorted query based on query response time t_q in isolation. *Gscheduler* is different from *First -Come-First-Served*, *Shortest-Job-Next* scheduler, we randomly select a query from *W* whenever a query finish. The performance of all schedulers shows in Fig. 4, 5, 6 and 7, respectively.

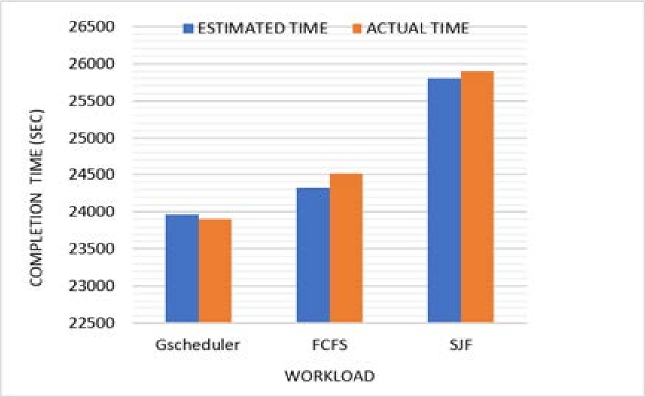


Fig 4. Completion Times for Gscheduler, FCFS and SJN.

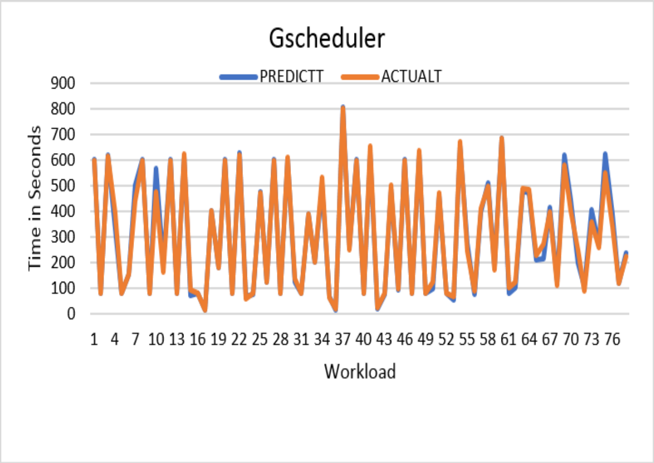


Fig 5. Predicted and Actual Time of the Gscheduler.

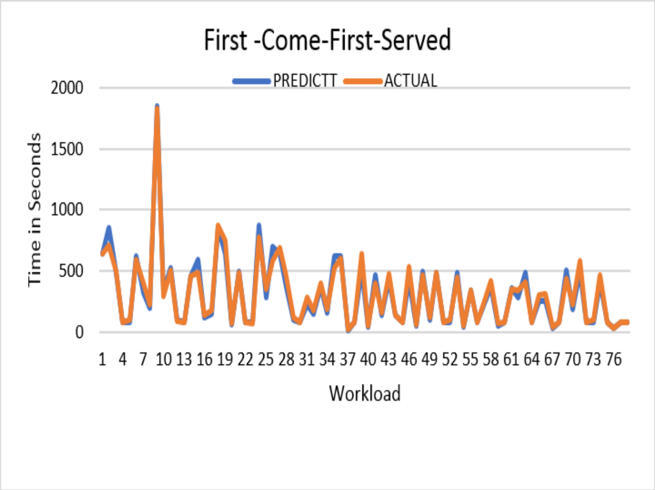


Fig 6. Predicted and Actual Time of the FCFS.

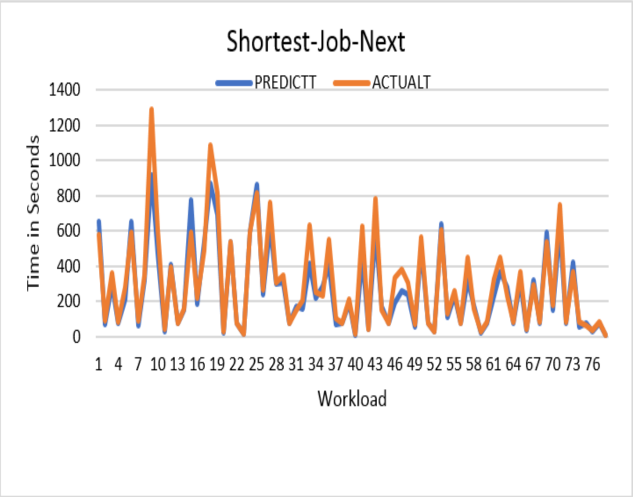


Fig 7. Predicted and Actual Time of the SJN.

V. Conclusion and Future Work

In a business intelligence setting, it is substantial for database administrators to estimate the total completion time of the workload. The state of the art methods don't tender any procedure for the database administrators to estimate the completion time of the workload. An approach, interaction into account is presented in this paper. Interaction aware experiment driven model assumbled with workload simulation for estimating completion time of the workload. an experimental evaluation with TPC-H benchmark running on PostgreSQL has proven that our approach can estimate workload completion time with high degree of accuracy across a broad spectrum of workloads.

A full research agenda is moving forward. We want to conduct research to form an integrated characterizing framework for query mixes, which may help us learn more about to measure query interactions method so that we could improve the models. We would make plan to study the adaptable similarity model which could make more accurate prediction because the DBMS is changing over time.

References

[1] Amjad, M. and J. Zhang, Gscheduler: A Query Scheduler Based on Query Interactions. In proceeding of the Web Information Systems and Applications, Lecture Notes in Computer Science, Springer, Cham, China: WISA, 2018: 11242.

[2] Wasserman, T, J. Martin, P. Skillicorn, D, B. and H. Rizvi, Developing a characterization of business intelligence workloads for sizing new database systems. In DOLAP, 2004.

[3] Mishra, C. and N. Koudas, The design of a query monitoring system. ACM Trans. Database Syst., 34(1), 2009.

[4] Guirguis, S. Sharaf, M. A. and P. K. Chrysanthis, A. Labrinidis, and K. Pruhs. Adaptive scheduling of web transactions. In ICDE, 2009.

[5] Tozer, S. Brecht, T. and A. Aboulnaga, Q-Cop: Avoiding bad query mixes to minimize client timeouts under heavy loads[C]. Proc. of the 26th International Conference on Data Engineering. Long Beach, California, USA: IEEE, 2010:397-408.

[6] Akdere, M. Çetintemel, U. Riondato, M. et al., Learning-based Query Performance Modeling and Prediction [C]// Proc. of the 25th International Conference on Data Engineering. Washington, DC, USA: IEEE, 2012: 390-401.

[7] Ganapathi, A. Kuno, H. Dayal, U. and J. L. Wiener, et al., Predicting multiple metrics for queries: Better decisions enabled by machine learning[C] // Proc. of the Int Conf Data Eng, 2009: 592–603.

[8] Li, J. K ¨ onig, A. C. Narasayya, V. R and S. Chaudhuri, Robust estimation of resource consumption for sql queries using statistical techniques. PVLDB, 5(11):1555–1566, 2012.

[9] Wu, W. Yun, C. Shenghuo, Z. and T. Jun'ichi, et al., Predicting Query Execution Time: Are Optimizer Cost Models Really Unusable? Proc. of the 29th International Conference on Data Engineering, Brisbane Computer Society, Australia: IEEE, 2013: 1081-1092.

[10] M. Ahmad, M. Duan, S. Aboulnaga, A. and S. Babu, Predicting completion times of batch query workloads using interaction-aware models and simulation. In EDBT, pages 449–460, 2011.

[11] Duggan, J. Cetintemel, U. Papaemmanouil, O. et al., Performance prediction for concurrent database workloads. In Proceeding of the ACM SIGMOD International Conference on Management of Data. Athens, Greece: ACM, 2011:337-348.

[12] Ahmad, M. Aboulnaga, A. Babu, S. and K. Munagala, Interaction-aware scheduling of report-generation workloads. The VLDB Journal, 20:589–615, 2011.

[13] Ahmad, M. Duan, S. Aboulnaga, A and S. Babu, Predicting completion times of batch query workloads using interaction-aware models and simulation. In EDBT, pages 449–460, 2011.

[14] Wu, W. Chi, Y. Zhu, S. Tatemura, J. Hacıg¨ um ¨ us, H. and J. F. Naughton, Predicting query execution time: are optimizer cost models really unusable? In ICDE, 2013.

[15] Li, J. Nehme, R.V. and JF. Naughton, GSLPI: a Cost-based Query Progress Indicator. Proc. of the 28th International Conference on Data Engineering, 2012, Washington DC, USA, IEEE Computer Society, pp. 678-689.

[16] Chaudhuri, S. Narasayya, V. R. and R. Ramamurthy, Estimating Progress of Execution for SQL Queries. Proc. of the 2004 ACM SIGMOD/PODS Conference ,2004, Paris, France, ACM, pp. 803-814.

[17] Gupta, C. Mehta, A. and U. Dayal, PQR: Predicting query execution times for autonomous workload management. In Proceedings of the 5th International Conference on Autonomic Computing (ICAC), 2008.

[18] Babu, S. Borisov, N. Duan, S. Herodotou, H and V. Thummala, Automated experiment-driven management of (database) systems. In Proceedings of the 12th Workshop on Hot Topics in Operating Systems (HotOS), 2009.

[19] Suri, R. Sahu, S. and M. Vernon, Approximate mean value analysis for closed queuing networks with multiple-server stations. In proceeding. Of Industrial Engineering Research Conference, 2007.

[20] McKay, M. D. Conover, W. J. and R. J. Beckman, A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code, Technometrics, 1979, 21,239-245.

[21] M. Stein, Large Sample Properties of Simulations Using Latin Hypercube Sampling, Technometrics, 1987 29:2, 143-151, DOI: 10.1080/00401706.1987.10488205.

[22] Hicks, C. R. and K. V. Turner, Fundamental Concepts in the Design of Experiments. Oxford University Press, 1999.

[23] Ahmad, M., Aboulnaga, A., Babu, S., and Munagala, K. Modeling and Exploiting Query Interactions in Database Systems . Proc. of the 17th ACM Conference on Information and Knowledge Management, 2008, Napa Valley, California, USA, ACM, pp. 183-192.

[24] Luo, G. Naughton, J. F. Ellmann, C. J, et al. Toward a progress indicator for database queries[C] // Proc. of the ACM SIGMOD International Conference on Management of Data. Paris, France: ACM, 2004:791-802.

Improving Disease Prediction using Shallow Convolutional Neural Networks on Metagenomic Data Visualizations based on Mean-Shift Clustering Algorithm

Hai Thanh Nguyen[1]
College of Information and
Communication Technology
Can Tho University
Can Tho, Vietnam

Toan Bao Tran[2]
Center of Software Engineering, Duy Tan University,
Da Nang, 550000 Vietnam
Institute of Research and Development, Duy Tan University,
Da Nang, 550000 Vietnam

Huong Hoang Luong[3]
Department of Information Technology
FPT University
Can Tho, Vietnam

Trung Phuoc Le[4]
Department of Information Technology
FPT University
Can Tho, Vietnam

Nghi C. Tran[5]
National Central University
Taoyuan, Taiwan

***Abstract*—Metagenomic data is a novel and valuable source for personalized medicine approaches to improve human health. Data Visualization is a crucial technique in data analysis to explore and find patterns in data. Especially, data resources from metagenomic often have very high dimension so humans face big challenges to understand them. In this study, we introduce a visualization method based on Mean-shift algorithm which enables us to observe high-dimensional data via images exhibiting clustered features by the clustering method. Then, these generated synthetic images are fetched into a convolutional neural network to do disease prediction tasks. The proposed method shows promising results when we evaluate the approach on four metagenomic bacterial species abundance datasets related to four diseases including Liver Cirrhosis, Colorectal Cancer, Obesity, and Type 2 Diabetes.**

***Keywords*—Clustering algorithm; metagenomic; visualization; disease prediction; mean-shift; personalized medicine; species abundance; bacterial**

I. Introduction

Human healthcare has been moving towards step by step to personalized medicine which using genetic insights and technologies. In 2020, the outbreaks of SARS-CoV-2 raises questions about the advantages of personalized medicine in general and metagenomics in particular. Personalized medicine also commonly referred to as precision medicine is the most promising approach for effective medical treatment of the individual patients based on their genetic information and medical symptoms. By combining with the traditional approaches which are based upon a policy of "one size fits all" applying the same treatments to whom with the same diseases, personalized medicine may be used to analyze and treat the disease by personalizing medicines to make them more specific, effective, and thereby improving treatment outcomes. Summarily, personalized medicine is a new approach in disease management, focusing on four essential premises: prediction, prevention, personalization, and participation [1]. Following the premises of personalized medicine, the appearance of SARS-CoV-2 may be explored based on acting on risk factors, cultures, and social determinants ($prediction$), constrain on evolution of the virus ($prevention$), analyzing the genetic and molecular of each the patient and giving them their personalizing medicines ($personalization$), requiring the investment for the infrastructure, human resource training, and the cooperation of the patients ($participation$) [2]. Several studies have indicated that many diseases are originally from genotypic so that personalized medicine is an effective treatment and can reduce the disadvantages of side effects. Many of the advantages of personalized medicine within healthcare detect and diagnose diseases, prevention of disease, and reduction of trial-and-error prescriptions.

Metagenomics that is the study of the metagenome, an application of modern genomic techniques, explores directly the communities of microbial in their natural habitats [3]. The emergence of high-throughput sequencing technology such as deep metagenomic sequencing has generated an amount of data that allowed the researchers to study both taxonomic and functional effects of microbiota on hosts [4]. The uncultured microorganisms represent the huge majority of organisms in most habitats on this planet proving by the analysis of 16rRNA sequences, it is the beginning for the development of metagenomics and led to the discovery of vast new lineages of microbial life [4], [5], [6]. The importance of understanding the microbiome has been repeatedly emphasized, thousands of human microbiome projects that have focused on the bacterial cell structure of the microbiome. The metagenomic analysis revealed variations in niche-specific abundance within the microbiome. Several studies presented the advantages of metagenomics in diagnostics and evidence-based medicine. Analyzing of Big data play a specific role in determining the causality of clinical diseases by bacteria and treating by a suitable medicine. Therefore, in the Personalized Medicine field, metagenomics is an efficient tool to deal with numerous pressing issues and the relatives [7], [8].

II. Related Work

Metagenomic analysis has become an exciting subject for the scientific community, the primary effort on the analysis of the microbiome is the identification of microbial communities for disease or host phenotype prediction [9], [10].

Diagnostic metagenomics can be used to identify pathogens on clinical samples, outbreaks of disease or novel variant viruses. Recently, the first genome sequence of SARS-CoV-2 was conducted with metagenomic RNA sequencing, an unbiased and high-throughput method of sequencing multiple genomes [8]. As an indicator of the benefits and problems of broad screens in clinical microbiology, the well-developed blood culture contamination literature has numerous researches to conduct clinical utility studies of diagnostic metagenomics, and demonstrate associations with increased hospital costs, hospitalizations, antibiotics, surgeries, and laboratory tests [11], [12], [13], [14], [15], [16].

The study in [17] proposed a method that can detect the overlapping clusters on metagenomic sequencing data by the Bayesian multinomial matrix factorization model. The authors stated under the Bayesian framework, the number of clusters is determined by the algorithm and improving the interpretability of their detection from the available information gained from a rank tree of microbes. The cluster structures are built hierarchically based on Dirichlet-multinomial mixtures with the purpose to indicate the relative abundance of taxa through a set of latent variables. By given the binary matrix, the priors are assorted hierarchically to characterize the heterogeneity via latent features. Summarily, this approach can handle the natural microbiome data and describes the generating process of data by the Bayesian model.

DeepMicrobes is described as a state-of-the-art metagenomics tool and the first deep learning architecture that incorporates self-attention mechanisms for DNA sequence analysis. *DeepMicrobes* facilitates taxonomic classification for cohorts of interest using newly discovered species in large-scale metagenomic assembly studies. The DNA sequence was encoded into numeric matrices, these are one-hot encoding as and k-mer embedding. The convolution model, hybrid convolutional, and recurrent model take DNA sequence one-hot encoding as an input layer whereas the other as the first layer of deep neural networks. For one-hot encoding, DNA was converted into $4 \times L$ matrix. For k-mer embedding, a DNA sequence of length L was split into a list of substrings of length K with a stride of S. The authors used a stride of none for their final model, ending up with $L-K+1$ substrings. The length of K was chosen to reach a balance between the model's fitting capacity and computational resources [18].

The different approach, phylogenetic tree embedded is an interesting approach for metagenomics data analysis. Essentially, the phylogenetic tree is a $2D$ matrix populated with the relative abundance of microbial taxa in a metagenomic sample, then, to be used as an input for the CNN [19]. With this method, the constructed matrices provide better spatial and quantitative information in the metagenomic data. Besides, the authors also proposed the convolutional neural networks, namely the PopPhy-CNN and Cytoscape-a visualization method used to facilitate the examination and interpretation of the retrieved taxa on the phylogenetic tree. The authors demonstrated the feasibility of extracting features can improve the performance of SVMs compared to the other models. They also indicated the conventional vector input $1D - CNN$ does not take advantage of the biological knowledge in the phylogenetic tree. The phylogenetic information was also utilized in sparse linear discriminant models with the simultaneous use of intermediate nodes and leaves on a phylogenetic tree [20].

PhyloPhlAn 3.0 is a framework for large-scale microbial genome characterization and phylogenetic analysis on a large number of features, it scales to large phylogenies comprising $> 17,000$ microbial species and assign genomes from isolate sequencing or MAGs to species-level genome bis built from $> 230,000$ publicly available sequences. Generally, this framework is to use available references genomes, retrieve the phylogenetic markers, perform taxonomic assignment and refinement, adopt specific choices for very large scale phylogenies, and provide additional information obtained from the resulting phylogenies [21].

The data is the most limitation in machine learning, many learning algorithms require large amounts of data for the training section. However, with the data augmentation method, the performance and generalization can be improved. The authors in [22] proposed an approach for generating microbiome data by using a conditional generative adversarial network (CGAN). Additionally, synthetic datasets generated using GAN models have shown to be able to boost the performance of prediction based tasks through data augmentation [23]. CGANs are an extension of the GAN and allow the generation of samples that have certain conditions or attributes. The authors in [22] have shown this approach can improve the performance of logistic regression and MultiLayer Perceptron in predicting host phenotype. They also stated the selecting CGAN model is the limitation of this approach, it is a subjective and may miss the optimal model.

The identification based on statistical analysis to detect the different abundant taxa between disease. The authors in [24] presented a new deep learning approach, namely PopPhy-CNN, a novel convolutional neural networks (CNN) learning architecture that effectively exploits phylogenetic structure in microbial taxa. The microbial taxonomic abundance profiles have been transformed into a structured data by using a phylogenetic tree, their approach is using “Operational Taxonomic Units” (OTUs) then converting OTU vector into an input matrix for their model. OTUs are generated by clustering sequences according to a computed distance between two similar sequences and a threshold. OTUs clustering can produce high quality groups precisely due to amplicon sequences are by definition taxa-specific and different between species [25]. The clustering performance depends on the choice of threshold due to sequencing errors. Furthermore, the analysis and biologically meaningful can be problematic [26].

In this study, we present a metagenomic data visualization-based Mean-Shift algorithm to cluster features in images prepared for prediction tasks, the contributions include:

- We present a features clustering approach with Mean-Shift algorithm and compare to the other visualization methods including Fill-up with phylogenetic ordering [27] and t-Distributed Stochastic Neighbor Embedding (t-SNE) [27], [28].
- The efficient of the proposed visualization methods is evaluated on four diseases including Liver Cirrhosis (CIR), Colorectal Cancer (COL), Obesity (OBE), Type 2 diabetes (WT2) [9], [27]. The performance on the datasets with the considered diseases obtains better results in prediction tasks comparing to the-state-of-the-art such as MetAML [9], Fill-up with phylogenetic ordering and t-SNE using transparent rates with $alpha = 0.5$ and $alpha = 1$.
- We also test visualizations with a vast of colormaps including jet, rainbow, gray and customized colormaps. These color spaces exhibit various results. Color images perform the best on Liver Cirrhosis dataset and samples of Colorectal Cancer while gray scale reveals good results for Obesity and Type 2 Diabetes samples.

The remaining of this study, we introduce the visualization approaches for metagenomic data including Fill-up approach and Fill-up with Mean-Shift clustering algorithm for arranging features in Section III. In Section IV, metagenomic bacterial species abundance datasets used in the experiments are described in detail. Moreover, we present the Convolutional Neural architecture for the proposed visualization method and settings for the learning. The performance of our approach is compared to the state-of-the-art in this section. We discuss and summarize the experimental results in Section V.

III. Features Arrangement based on Mean-shift Clustering in Fill-up method

Data visualization is a strong method to interpret data. Each visualization method will be used to represent the abundance or presence of data. In this study, we propose a visualization method based on Mean-shift clustering algorithm to arrange features in images, thereby making it easier for observing the distribution of the features. Therefore, we expect to improve the performance of disease prediction task with the proposed visualizations.

A. Metagenomic Visualization by Fill up Approach

Fill-up [27] is an effective solution for visualizing metagenomic data into images. The main idea of this method is to arrange features into a square matrix which has minimum size to fit all features and contains arranged abundance or presence values in a right-to-left order by row top-to-bottom. The order to arrange species can follow the phylogenetic-sorting or another type of ordering.

t-SNE is also a technique for visualizing metagenomic data. t-SNE not only captures the local structure of the higher dimension but also preserves the global structures of the data like clusters.

In order to convert continuous values into discrete values (for coloring features on images), we use a binning technique. Binning is a data pre-processing method, the key goal is to reduce the effects of minor observation errors, it has a smoothing effect on the input data and may also reduce the chances of overfitting in case of small datasets. In this study, Species Bin (SPB) [27] is implemented and investigated to pre-process values before visualizing them onto an image. With the binning technique, the features were visualized into images by Fill-up with phylogenetic-sorting or t-Distributed Stochastic Neighbor Embedding (t-SNE) in [27].

B. Mean-shift Clustering in Fill-up Method

Algorithm 1 Algorithm for features clustering based on Mean-shift algorithms

```
Input:
  ○ $D$: a data matrix where each row is a sample and each column represents a feature
Output:
  ○ $B$: an array containing a list of strings combining between the labels of generated clusters and order of features sorted by phylogenetic ordering.
Begin
    Step 1: Sort $D$ so that features along with their data follow phylogenetic ordering. Save the list containing the order of features according to phylogenetic ordering to $P$.

    Step 2: Transpose $D$: $D_1 = t(D)$. Because we want to group features into clusters, we transpose $D$ so that each feature at this time is considered as a "data point" for clustering.

    Step 3: Run Mean-shift clustering algorithm on $D_1$ to indicate clusters for features. Each feature is assigned to a cluster. A cluster can contain one or more features.

    . The labels of clusters contained features are saved to $L$. $L$ includes information on clusters which each feature belongs to. For example, the 1st feature belongs to cluster 5, the 2nd feature is labeled to cluster 1, and so on.

    Step 4:
    . We concatenate labels of clusters for features $L$ and their phylogenetic ordering:
        $B[i] = string(L[i]) +' \_' + string(P(i))$
        $With : i = 0..\#features$

    Return $B$
End
```

Mean-shift [29] is an unsupervised learning algorithm. In principle, the algorithm iteratively assigns each data point towards the closest cluster centroid and direction to the closest cluster centroid is determined by where most of the points nearby are at. So each iteration each data point will move closer to where the most points are at, which is or will lead to the cluster center. When the algorithm stops, each point is assigned to a cluster. Assume we have:

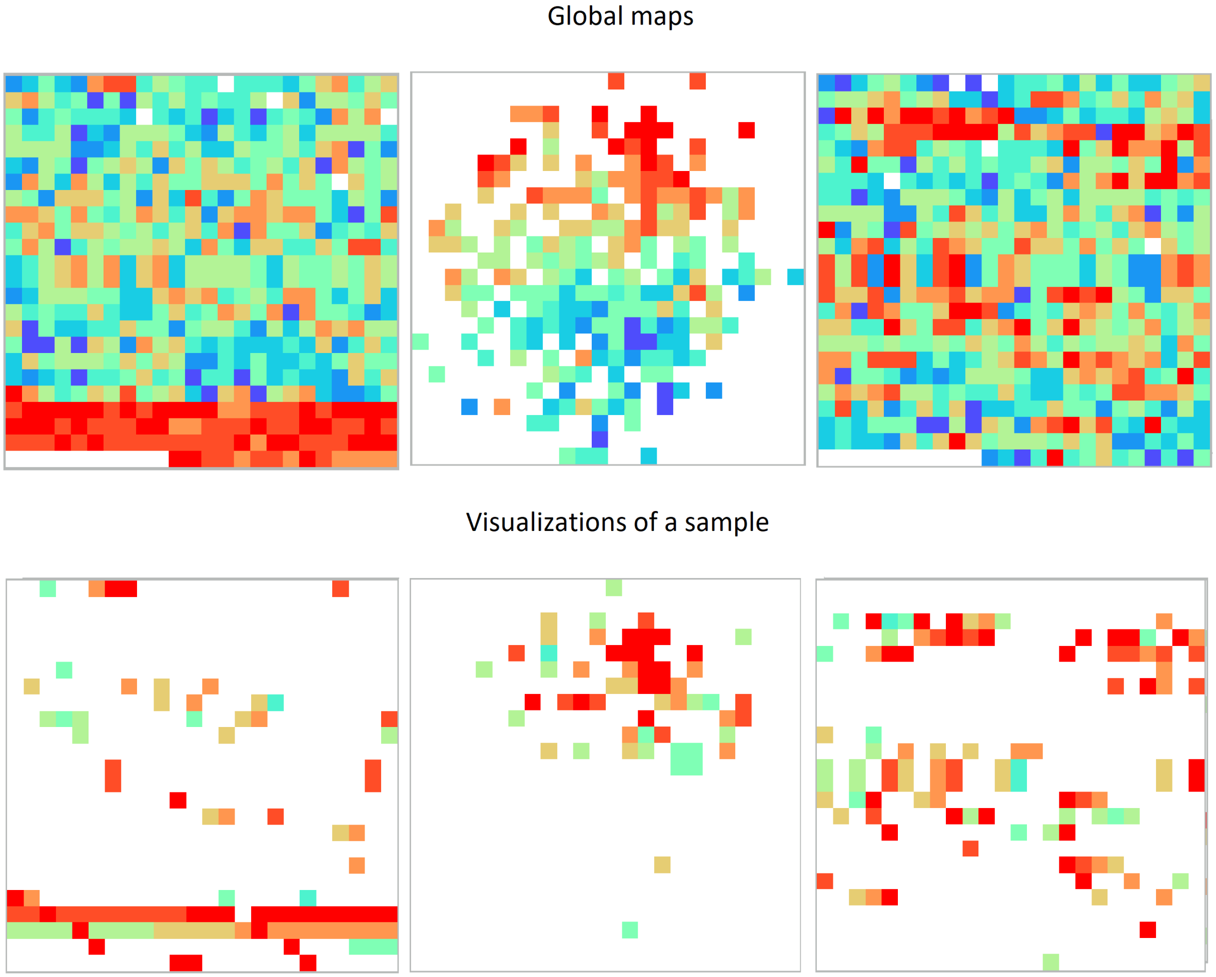


Fig. 1. Liver Cirrhosis samples (CIR dataset, details in Table I) Visualization comparison using various features arrangements including (left-to-right) Mean-shift, t-SNE, and features ordered based on phylogenetic information. The first row: global images. The second row: visualizations of a sample.

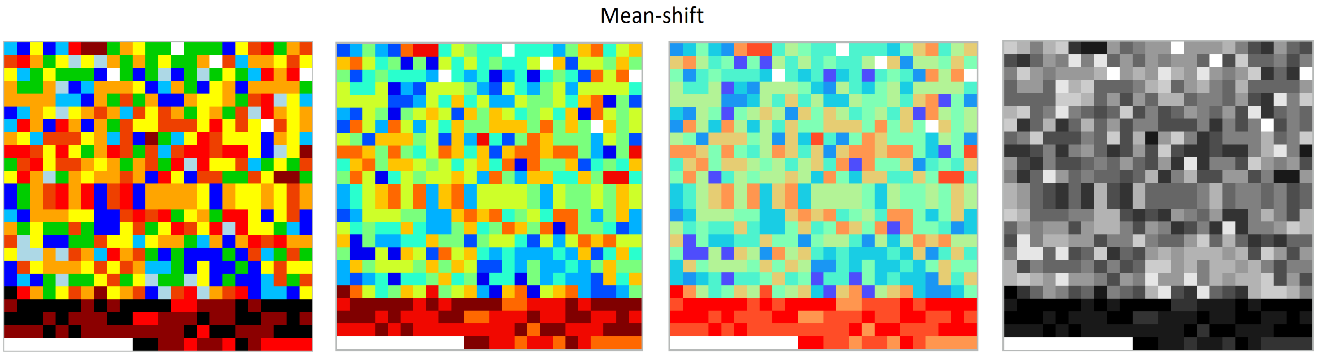


Fig. 2. Visualization of the global maps from Liver Cirrhosis samples (CIR dataset) using various color spaces including custom, jet, rainbow and gray scale with Mean-shift clustering.

- ○ Initial estimate x.
- ○ Gaussian kernel function:

$$K(x_i - x) = e^{-c\|\mathbf{x_i} - \mathbf{x^2}\|}$$

This function determines the weight of nearby points for re-estimation of the mean.

The weighted mean of the density in the window determined by K is:

$$m(x) = \frac{\sum_{x_i \in N_{(x)}} K(x_i - x)x_i}{\sum_{x_i \in N_{(x)}} K(x_i - x)}$$

Where:

- ○ $N_{(x)}$ is the neighborhood of x.

$m(x) - x$ is called mean shift [29] and $x \leftarrow m(x)$, and repeats the estimation until $m(x)$ converges.

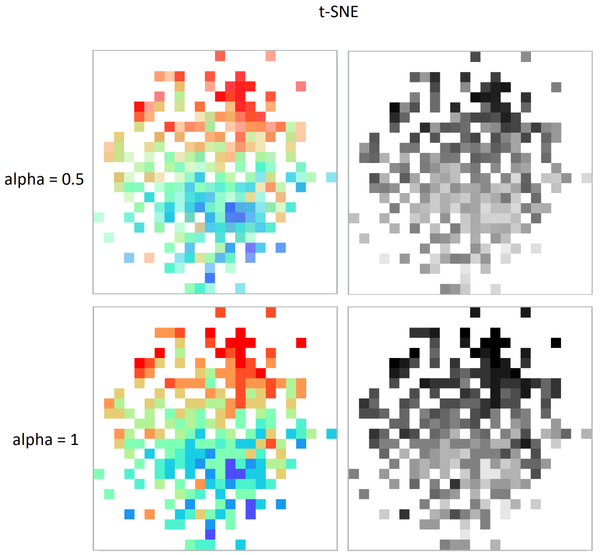


Fig. 3. Visualization comparison between rainbow colormap and gray scale images using t-SNE on Liver Cirrhosis samples. Top: t-SNE with $alpha = 0.5$. Bottom: t-SNE with $alpha = 1$.

The synthetic metagenomic images are generated by Fill-up and t-SNE method in [27]. In this study, use of Mean-shift algorithm, we expect to improve the performance by finding regions containing a high density of data and group them into a cluster with smallest non-overlapping boundaries. This approach is performed as shown in Algorithm 1. After clustering, we obtain an array B which contains the arranged features order by the labeled clusters along with information on phylogenetic ordering. Information on phylogenetic embedded in synthetic metagenomic images is based on the alphabetical order as described in [27]. In our method, if features are in the same cluster, we consider the alphabetical order of features to place them close together. By combining between order of features sorted by cluster labels and phylogenetic ordering, we expect to improve the quality of visualizations as well as the prediction performance of deep learning algorithm on the proposed visualizations.

In order to visualize features, we use 10 colors in gray-scale, rainbow, jet, and custom colormap [27]. In Fig. 1 displays the comparison between clustered features on global and sample images from Liver Cirrhosis samples (CIR dataset, see details in Table I) based on mentioned visualization methods in rainbow colormap. The global map which is an image visualizing average value of each feature of all samples in training set. From left-to-right and top-to-bottom, the first two images in Fig. 1 shows the global and sample image visualized by Fill-up combining the clustering method, in the middle contains images represent the global map and a sample visualization of t-SNE embedding. The last ones are visualized by Fill-up with phylogenetic ordering. We only use samples from training set to cluster features and build coordinates for all features. These coordinates are carried out to generate all images for samples of both training set and test sets.

Fig. 2 illustrates the representation of clustered features in different colors. The images in Fig. 2 are global images from CIR dataset which mentioned above with Fill-up and clustering method, from left-to-right custom, jet, rainbow, and gray colormaps. More specific, the custom colormap is built based on jet combined to black with distinctive colors. Furthermore, we also visualize the global images with t-SNE exhibited in Fig. 3, the images on the top are generated by t-SNE with $alpha = 0.5$ while the second row shows the images with $alpha = 1$. The first column, we use rainbow colormap while gray scale is applied for images in the other. The difference between t-SNE with $alpha = 0.5$ and $alpha = 1$ is the problem of the overlapped points. t-SNE suffers overlapped issues where the visualization exists numerous points which are hidden by other points. In order to reduce this negative affect, the alpha value is deployed in the RGBA color space to indicate the transparency of a colour. The alpha value ranges from 0 to 1 where 0 is completely transparent while alpha value of 1 is not transparent at all. By choosing $alpha = 0.5$, the futures are mixed-up if they are overlapped. Otherwise, with $alpha = 1$, some features can be hidden by other features.

IV. Experimental Results

A. Benchmark Datasets

We evaluated our approach on four bacterial species abundance datasets [9], [27] which are related to four diseases including Liver Cirrhosis (CIR), Colorectal Cancer (COL), Obesity (OBE), and Type 2 diabetes samples from western women (WT2). Details are in Table I. For each sample, species abundance (feature) is represented as a real number and the total abundance of all species in a sample sums to 1:

$$\sum_{i=1}^{k} f_i = 1$$

With:

- ○ k is the number of features for a sample.
- ○ f_i is the value of the i-th feature.

Table I presents the details of all considered datasets including the numbers of features, samples, and some extra

TABLE I. Four considered bacterial species abundance datasets description

Diseases	**Liver Cirrhosis**	**Colorectal Cancer**	**Obesity**	**Type 2 diabetes**
Datasets name	**CIR**	**COL**	**OBE**	**WT2**
#Features	542	503	465	381
#Samples	232	121	253	96
#Patients	118	48	164	53
#Controls (healthy)	114	73	89	43
Ratio of patients	0.51	0.40	0.65	0.55
Ratio of Controls (healthy)	0.49	0.60	0.35	0.45
Minimum size of images	24 × 24	23 × 23	22 × 22	20 × 20

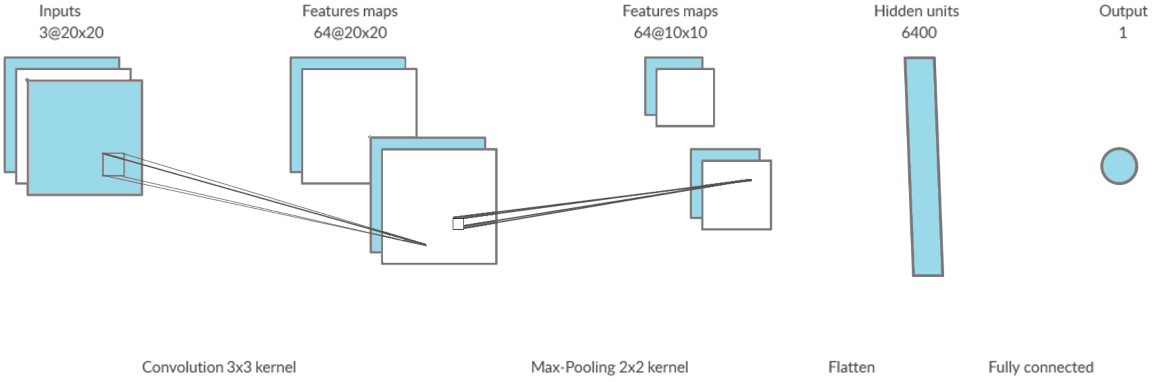


Fig. 4. A shallow convolutional Neural Network Architecture for metagenomic images on color images of WT2 samples.

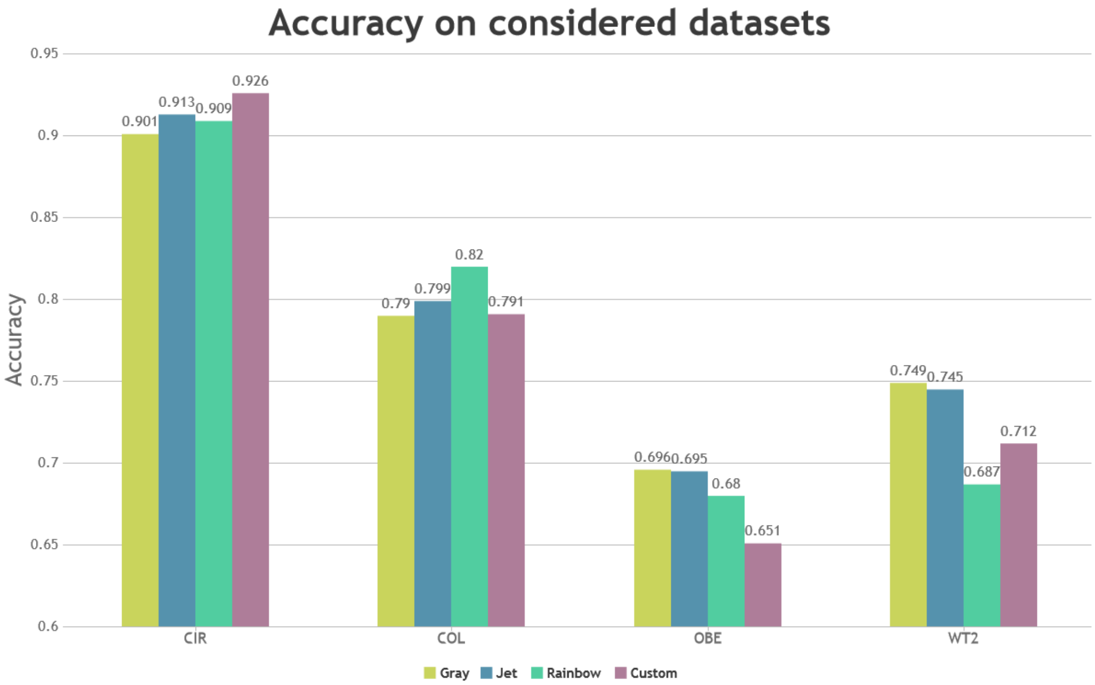


Fig. 5. Performance Comparison of different colormaps on all considered metagenomic datasets using Mean-shift for features arrangement in metagenomic visualization

information. We calculate the ceiling of Square Root of the numbers and then of features to feed into a 2D matrix. For instant, on CIR dataset we have 542 features, so the 2D matrix shape should be 24×24 to contain 542 features because $\sqrt{542} = 23.28$ and the ceiling of 23.28 is 24.

B. Learning Model and Settings

Our classification tasks are carried out by a shallow deep learning network, a Convolutional Neural Network (CNN) as illustrated in Fig. 4. The architecture contains one Convolutional layer with 64 kernels of 3×3, followed by a Max-Pooling layer of 2×2 (stride 2) and a Fully Connected layer. CNN is implemented with Adam optimizer, the default learning rate is 0.001, and the network uses a batch of size 16. The architecture is suggested from [27]. To avoid overfitting issues, if the $loss$ is not improved after every consecutive 5 epochs, we will stop the training section by using the Early Stopping method. In the opposite case, training can run up to 500 epochs. To evaluate the performance, we compute average accuracy (ACC) on 10-fold-cross-validation. The same folds are used for all classifiers. We calculate the accuracy which is the fraction of true predictions by the following formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Where:

- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative

C. Disease Classification of Mean-shift Clustering with Fill up on Various Diseases

The efficiency of arranging features based on Mean-shift is evaluated in various colormaps, namely gray, jet, rainbow, and custom. The last one is combined between black and jet colormap. Fig. 5 illustrates the average accuracy of the methods on four considered datasets. Generally, each colormap gives a satisfying result each individual dataset. The the jet colormap exhibits a quite good and reaches the highest average performance on four datasets while gray scale works well on OBE and WT2 and the rainbow achieves the highest performance on COL while the custom colormap gives exceptional results on CIR with the performances of 0.926.

D. State-of-the-art Comparison

The performance comparison of Mean-shift clustering, t-SNE, and phylogenetic ordering [27] are illustrated in Fig. 6 and Fig. 7. The chart in Fig. 7 reveals the accuracy on four considered datasets using rainbow colormap while the results with gray images are shown in the other. As observed, the in Fig. 6 features arrangements based on Mean-shift clustering demonstrates its advantages on 3 out of 4 datasets using both rainbow in comparing to phylogenetic ordering.

Furthermore, we also summarize the result with results of the jet and custom colormaps, and compare to MetAML [9], a computation framework for metagenomic analysis based on classic machine learning algorithms such as Random Forests and Support Vector Machines in Table II. On CIR dataset, Mean-shift clustering method reaches the accuracy of 0.926 while MetAML, t-SNE ($alpha = 1$) and Fill-up using phylogenetic ordering reveal the results of 0.877, 0.853 and 0.897 respectively. The color images which are jet, rainbow, and custom give quite better results than gray images on Liver Cirrhosis and Colorectal Cancer samples while the results with gray scale are slight better for OBE and WT2 datasets.

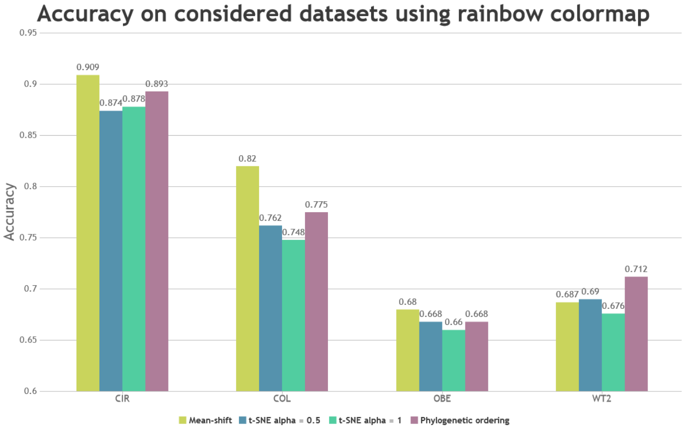


Fig. 6. Performance of different visualization approaches using rainbow colormap on four considered datasets (details in Table II).

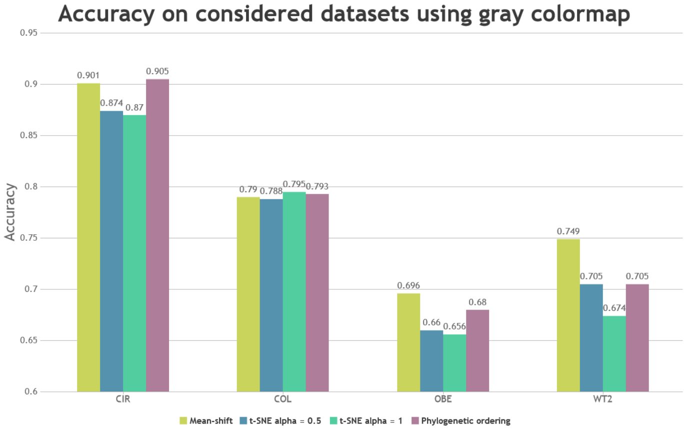


Fig. 7. Visualization methods Comparison in ACC on the considered datasets using gray scale images (details revealed in Table II).

We also compute the average accuracy on four investigated datasets for the comparison in the last column of Table II. In general, as shown in the table, visualization methods with Fill-up based on Mean-shift clustering algorithm (including average values of 0.771, 0.777, 0.784, 0.788 with customized, rainbow, gray scale, and jet colormaps, respectively) outperform MetAML, t-SNE and Fill-up using phylogenetic ordering with the values of 0.757 of MetAML, 0.774 and 0.741 being the best results of Fill-up with phylogenetic ordering and t-SNE, respectively. Jet colormap appears to be the most efficiency while custom colormap with customized distinctive colors shows the worst among the considered color spaces. However, we noted that the best accuracy is on CIR dataset with an average accuracy of 0.926 using custom colormap.

TABLE II. Comparison with the-state-of-the-art. **Bold results** are better performance than the method of fill-up with phylogenetic ordering.

Approaches	Color space	CIR	COL	OBE	WT2	AVG
MetAML [9]	-	0.877	0.805	0.644	0.703	0.757
t-SNE with $alpha = 1$ [27]	gray	0.870	0.795	0.656	0.674	0.749
Fill-up phylogenetic ordering [27]	gray	0.905	0.793	0.680	0.705	0.770
Our approach	gray	0.901	0.790	**0.696**	**0.749**	**0.784**
t-SNE with $alpha = 1$ [27]	jet	0.879	0.748	0.661	0.660	0.737
Fill-up phylogenetic ordering [27]	jet	0.903	0.798	0.681	0.713	0.774
Our approach	jet	**0.913**	**0.799**	**0.695**	**0.745**	**0.788**
t-SNE with $alpha = 1$ [27]	rainbow	0.878	0.748	0.660	0.676	0.741
Fill-up phylogenetic ordering [27]	rainbow	0.893	0.775	0.668	0.712	0.762
Our approach	rainbow	**0.909**	**0.820**	**0.690**	0.687	**0.777**
t-SNE with $alpha = 1$ [27]	custom	0.853	0.771	0.660	0.661	0.736
Fill-up phylogenetic ordering [27]	custom	0.897	0.782	0.673	0.707	0.765
Our approach	custom	**0.926**	**0.791**	0.656	**0.712**	**0.771**

V. Discussion and Conclusion

We presented an approach to visualize high-dimensional data using features arrangement with Mean-shift and compare the results to the state-of-the-art. The method reveals encouraging results. We obtain better results on all considered datasets compared to Fill-up with phylogenetic ordering and t-SNE images classified by deep learning algorithm and MetAML with a classic machine learning algorithm. As seen from the experiments, features which are clustered to arrange close together show benefits to improve the performance both in visualizations and in classification tasks. Although t-SNE also groups similar features, it suffers the issue of overlapped points. However, for gray images on Colorectal cancer samples, t-SNE achieves a slightly better result comparing to others. Further research can work on t-SNE to investigate approaches to enhance performance.

Various colormaps are carried out to compare different methods. The results depend on different datasets for the classification tasks. Customised colors obtain the highest average accuracy with 0.926 on CIR dataset while it shows only an average accuracy of 0.656 on OBE dataset. It is clear that visualization methods are good solutions for Liver Cirrhosis, Colorectal Cancer prediction but Predicting Obesity and Type 2 diabetes is still facing challenges with metagenomic data. However, the performance on Cirrhosis samples and Colorectal cancer samples also reveal great potentials of metagenomic in disease prediction with personalized medicine.

Our study only runs the classification tasks with shallow deep learning architectures. Advancements in deep learning techniques have been increasing their efficiency on numerous fields. In the future, further research should investigate on deeper architectures and more sophisticated techniques to improve the performance on synthetic metagenomic visualizations classification tasks.

References

[1] Sagner M, McNeil A, Puska P, Auffray C, Price ND, Hood L, et al. The P4 health spectrum - a predictive, preventive, personalized and participatory continuum for promoting healthspan. Prog Cardiovasc Dis. 2017;59:506–521. doi: 10.1016/j.pcad.2016.08.002. 2016.

[2] Crisci, Carlos D et al. “A Precision Medicine Approach to SARS-CoV-2 Pandemic Management.” Current treatment options in allergy, 1-19. 8 May. 2020, doi:10.1007/s40521-020-00258-8. 2020.

[3] Chen K, Pachter L. Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput Biol. 2005;1(2):106-112. doi:10.1371/journal.pcbi.0010024

[4] Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev. 2004;68(4):669-685. doi:10.1128/MMBR.68.4.669-685.2004

[5] Ma, Bing & France, Michael & Ravel, Jacques. (2020). Meta-Pangenome: At the Crossroad of Pangenomics and Metagenomics. doi10.1007/978-3-030-38281-0_9.

[6] Jang SJ, Ho PT, Jun SY, Kim D, Won YJ. Dataset supporting description of the new mussel species of genus Gigantidas (Bivalvia: Mytilidae) and metagenomic data of bacterial community in the host mussel gill tissue. Data Brief. 2020;30:105651. Published 2020 Apr 29. doi:10.1016/j.dib.2020.105651. 2020

[7] Alfredo D. Guerron et al. “Performance and Improvement of the DiaRem Score in Diabetes Remission Prediction - A Study with Diverse Procedure Types”, May. 2020, doi:https://doi.org/10.1016/j.soard.2020.05.010. 2020.

[8] Hongyu Chen, Sanjeev Kumar Awasthi, Tao Liu, Zengqiang Zhang. Mukesh Kumar Awasthi, “An assessment of the functional enzymes and corresponding genes in chicken manure and wheat straw composted with addition of clay via meta-genomic analysis”, Industrial Crops and Products, vol. 153, 2020, doi:https://doi.org/10.1016/j.indcrop.2020.112573

[9] Pasolli et al. “Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights”. PLoS Comput. Biol. 2016;12(7):e1004977. Published 2016 Jul 11. doi:10.1371/journal.pcbi.1004977. 2016.

[10] Syed Hamid Jalal Shaha, Aamir Humayun Malik, Bing Zhang, Yiming Bao, Javaria Qazi, “Metagenomic analysis of relative abundance and diversity of bacterial microbiota in Bemisia tabaci infesting cotton crop in Pakistan”, May 2020, doi:https://doi.org/10.1016/j.meegid.2020.104381

[11] Hasman, Henrik et al. “Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples.” Journal of clinical microbiology vol. 52,1 (2014): 139-46. doi:10.1128/JCM.02452-13. 2014.

[12] Self WH, Speroff T, Grijalva CG, et al. Reducing blood culture contamination in the emergency department: an interrupted time series quality improvement study. Acad Emerg Med. 2013;20(1):89-97. doi:10.1111/acem.12057

[13] Hall KK, Lyman JA. Updated review of blood culture contamination. Clin Microbiol Rev. 2006;19(4):788-802. doi:10.1128/CMR.00062-05

[14] Gander RM, Byrd L, DeCrescenzo M, et al. Impact of blood cultures drawn by phlebotomy on contamination rates and health care costs in

a hospital emergency department. J Clin Microbiol. 2009;47:1021–1024. 2009. DOI:10.1128/JCM.02162-08

[15] Bates DW, Goldman L, Lee TH. Contaminant blood cultures and resource utilization. The true consequences of false-positive results. JAMA. 1991;265(3):365-369.

[16] van der Heijden YF, Miller G, Wright PW, Shepherd BE, Daniels TL, Talbot TR. Clinical impact of blood cultures contaminated with coagulase-negative staphylococci at an academic medical center. Infect Control Hosp Epidemiol. 2011;32(6):623-625. doi:10.1086/660096

[17] Fangting Zhou, Kejun He, Qiwei Li, Robert S. Chapkin, Yang Ni, "Bayesian biclustering for microbial metagenomic sequencing data via multinomial matrix factorization," arXiv:2005.08361, May 2020

[18] Qiaoxing Liang, Paul W Bible, Yu Liu, Bin Zou, Lai Wei, DeepMicrobes: taxonomic classification for metagenomics with deep learning, NAR Genomics and Bioinformatics, Volume 2, Issue 1, March 2020, lqaa009, https://doi.org/10.1093/nargab/lqaa009

[19] D. Reiman, A. Metwally, J. Sun and Y. Dai, "PopPhy-CNN: A Phylogenetic Tree Embedded Architecture for Convolutional Neural Networks to Predict Host Phenotype from Metagenomic Data," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.2993761. 2020.

[20] Fukuyama J, Rumker L, Sankaran K, et al. Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. PLoS Comput Biol. 2017;13(8):e1005706. Published 2017 Aug 18. doi:10.1371/journal.pcbi.1005706. 2017.

[21] Asnicar, F., Thomas, A.M., Beghini, F. et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat Commun 11, 2500 (2020). https://doi.org/10.1038/s41467-020-16366-7. 2020.

[22] Reiman, Derek and Dai, Yang, "Using Conditional Generative Adversarial Networks to Boost the Performance of Machine Learning in Microbiome Datasets," bioXiv:2020.05.18.102814, https://doi.org/10.1101/2020.05.18.102814, May 2020.

[23] Che, Z., Cheng, Y., Zhai, S., Sun, Z., & Liu, Y. (2017). Boosting Deep Learning Risk Prediction with Generative Adversarial Networks for Electronic Health Records. 2017 IEEE International Conference on Data Mining (ICDM), 787-792.

[24] D. Reiman, A. Metwally, J. Sun and Y. Dai, "PopPhy-CNN: A Phylogenetic Tree Embedded Architecture for Convolutional Neural Networks to Predict Host Phenotype from Metagenomic Data," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.2993761. 2020.

[25] Soueidan, Hayssam and Macha Nikolski. "Machine learning for metagenomics: methods and tools." arXiv: Genomics (2015): n. pag.

[26] (2014). Molecular Markers in Phylogenetic Studies-A Review. Journal of Phylogenetics & Evolutionary Biology. 02. 10.4172/2329-9002.1000131.

[27] Thanh Hai Nguyen, Edi Prifti, Nataliya Sokolovska, Jean-Daniel Zucker. Disease Prediction using Synthetic Image Representations of Metagenomic data and Convolutional Neural Networks. The 13th IEEE-RIVF International Conference on Computing and Communication Technologies 2019, Da Nang 20-22/03/2019; pp 231-236; 2019; ISBN 978-1-5386-9313-1. IEEE Xplore. 2019.

[28] Maaten, Laurens van der and Geoffrey E. Hinton. "Visualizing Data using t-SNE." (2008).

[29] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," in IEEE Transactions on Information Theory, vol. 21, no. 1, pp. 32-40, January 1975, doi: 10.1109/TIT.1975.1055330.

Hybrid Memory Design for High-Throughput and Low-Power Table Lookup in Internet Routers

Hayato Yamaki
Graduate School of Informatics and Engineering, The University of Electro-Communications
1-5-1 Chofugaoka, Chofu 182-8585, Japan

***Abstract*—Table lookup is a major process to decide the packet processing throughput and power efficiency of routers. To realize high-throughput and low-power table lookup, recent routers have employed several table lookup approaches, such as TCAM (Ternary Content Addressable Memory) based approach and DRAM (Dynamic Random Access Memory) based approach, depending on the purpose. However, it is difficult to realize both ultrahigh throughput and significant low power due to the trade-off between them. To satisfy both of the demands, this study proposes a hybrid memory design, which combines TCAM, DRAM, PPC (Packet Processing Cache), CMH (Cache Miss Handler), and IP Cache, to enable a high-throughput and low-power table lookup. The simulation results using an in-house cycle-accurate simulator showed that the proposed memory design achieved nearly 1 Tbps throughput with similar power of the DRAM-based approach. When compared to the approach proposed in a recent study, the proposed memory design can realize 1.95x higher throughput with 11% power consumption.**

Keywords*—Inter routers; packet processing; table lookup; hybrid memory architecture; Packet Processing Cache (PPC)

I. Introduction

A demand for high-throughput and low-power packet processing is becoming serious in routers year by year due to an increase in internet traffic. For example, to achieve 400 Gbps, routers must process packets every 1.28 nano second if the shortest packets arrive continuously. Moreover, because the power consumption of routers increases depending on the number of packets processed, the power efficiency is also an important factor for recent routers. According to the reports [1]–[4], the total power consumed by the network devices will reach several percentages of the total power generated in the world. Thus, routers must process packets at high throughput with low power.

Table lookup is a key process to decide the packet processing throughput and power efficiency of routers. When a packet arrives at a router, the router searches tables, such as a routing table and access control list, to decide the next-hop IP address and whether filtering. Because this process requires memory accesses per packet, table lookup is a bottleneck of the packet processing throughput and a power-hungry operation in routers.

To increase the throughput or reduce the power consumption, several table lookup approaches have been employed. A DRAM (dynamic random access memory) based approach is the most standard approach of the table lookup [5], [6]. This approach stores the tables into DRAM, which can be implemented with a large capacity at low expense and low power. Thus, the DRAM-based approach has been employed in enterprise edge routers, which are required to be introduced at low expense, and application routers, which process packets with fine grained services. However, because the access latency of DRAM is large, this approach cannot satisfy the demand of ultrahigh throughput, such as 400 Gbps and 1 Tbps.

In contrast to the DRAM-based approach, TCAM (ternary content addressable memory) based approach has been used for high-throughput table lookup [7]–[10]. The TCAM is a specialized memory for fast data lookup and can retrieve data at one cycle. Because the access latency of TCAM is smaller than that of a DRAM, this approach can achieve larger throughput than the DRAM-based approach. Thus, the TCAM-based approach has been employed in core routers and data-center routers, which are required to process packets at high throughput. However, to retrieve data at one cycle, TCAM consumes significant large power in comparison to DRAM. Accordingly, it was reported that TCAM consumed 40% of the total power consumed in a router [11]. The paper [9] also indicated that TCAM consumed 150 times larger power than same-sized DRAM.

Due to the trade-off between the access latency and power consumption of memories, as mentioned above, satisfying the demands of both throughput and power efficiency is difficult for these approaches. In this study, a novel hybrid memory design, which combines the DRAM- and TCAM-based approaches and further adds PPC (packet processing cache), CMH (cache miss handler), and IP Cache, is proposed for high-throughput and low-power table lookup in routers.

This study also builds an in-house cycle-accurate table lookup simulator which can simulate not only the proposed memory design but also the other conventional table lookup approaches (e.g., the DRAM-based or TCAM-based approaches). This simulator enables to measure the table lookup throughput of a router considering the hardware behavior (e.g., stalling and queuing) while most previous studies evaluated the throughput based on a mathematical analysis using a throughput model without considering the concrete hardware. Thus, this study newly reveals the impact of the hardware constraints (i.e., memory stalls, hash conflicts, and the number of buffer entries) on the table lookup throughput.

The major contributions of this study are summarized below:

- This study proposed a novel hybrid memory design for high-throughput and low-power table lookup in routers. The simulation results showed the proposed memory design can realize nearly 1 Tbps throughput with similar power of the DRAM-based approach.

- This study revealed the concrete hardware design of various table lookup approaches and their connection. Most previous studies did not consider the concrete hardware design of them, and thus, the impact of the hardware design on the table lookup performance was not revealed.
- This study is a first study of evaluating the throughput of the table lookup with caches considering the hardware constraints (i.e., memory stalls, hash conflict, and the number of buffer entries) by measuring them based on cycle-accurate simulations.

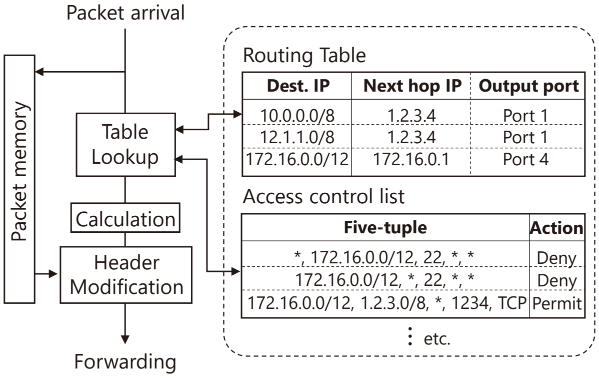
The remainder of this paper is organized as follows: Section 2 describes the table lookup process in a router. Section 3 introduces details of table lookup approaches used in this study as related works. We propose the efficient memory design for the high-throughput and low-energy table lookup in Section 4 and evaluate it in Section 5. Finally, Section 6 concludes this paper.

II. Table Lookup

Routers have several tables to correctly decide how to process packets. In general, typical routers have four tables: a routing table, ARP (address resolution protocol) table, ACL (access control list). and QoS (quality of service) table. Note that several router vendors have different named tables, such as a FIB and adjacency table in Cisco routers, instead of the routing table and ARP table. However, they make no difference in this study, and thus, we suppose that routers have the four tables mentioned above in this study hereafter.

Fig. 1 illustrates a basic flow of packet processing in a router. When a packet arrives at a router, the whole data of the packet is stored to a packet memory, and only the header part is sent to next processing flow (i.e., a table lookup module). It is because the payload is not required to process the packet. The table lookup module searches the tables using the packet header information. For example, the routing table lookup is done using a destination IP address while the ACL lookup is done using the five-tuple (i.e., source and destination IP addresses, source and destination port numbers, and protocol number). After searching the tables, the router recalculates the ttl (time to live) and checksum value from the packet header. Based on the table lookup results and recalculation values, the router modifies the packet header and concatenates the modified header with the packet payload read from the packet memory. Finally, the packet is forwarded to a next hop.

In the packet processing, the table lookup is a major throughput bottleneck due to the large access latency. As it is often discussed, memory-wall, which represents the large gap between CPU operation frequency and memory access latency, is the most serious problem of computer architecture, and it is also a problem in routers. Make matters worse, each table requires a large number of entries. For example, all BGP entries in the routing table exceeded 1M entries. Consequently, the table lookup conventionally takes from dozens to a hundred cycles although other processes (i.e., the calculation and header modification) take several cycles. Thus, improving the table lookup is the most important issue to increase the packet processing throughput.



Dest. IP	Next hop IP	Output port
10.0.0.0/8	1.2.3.4	Port 1
12.1.1.0/8	1.2.3.4	Port 1
172.16.0.0/12	172.16.0.1	Port 4

Five-tuple	Action
*, 172.16.0.0/12, 22, *, *	Deny
172.16.0.0/12, *, 22, *, *	Deny
172.16.0.0/12, 1.2.3.0/8, *, 1234, TCP	Permit

Fig. 1. Outline of table lookup in a router.

TABLE I. Summary of each approach.

Approach	Throughput	Energy	Notes
DRAM only (e.g., [5], [6])	Low	Low	Low expense
TCAM only (e.g., [7]–[10])	High	High	Lookup at one cycle
IP Cache + DRAM (e.g., [12]–[15])	Middle	**Very low**	Performance depends on the cache miss rate
PPC + TCAM (e.g., [16]–[19])	Very high	Middle	Same as above
PPC + CMH + TCAM (e.g., [20], [21])	**Very high+**	Middle	Non-blocking

III. Related Works

As explained in Section 2, the table lookup is the most important factor for routers to determine the packet processing throughput. Consequently, many studies are conducted to improve the table lookup. This section introduces major approaches which have been employed in routers. Table I summarized characteristics of each approach.

A. DRAM-based Approach

A DRAM-based approach is the most standard approach for the table lookup [5], [6]. In this approach, the tables are stored into DRAM. The DRAM is a typical memory and can be implemented with a large capacity at low expense. Thus, this approach can realize the table lookup at lowest expense compared to other approaches.

Conventionally, there are several methods to find data in DRAM. The most simple method is to search entries from the head to the end by linear search until the data are found. However, this method significantly increases the DRAM accesses because the tables require a large number of entries, as mentioned in Section 2. Using hash values as addresses of data in DRAM is another method. However, this way causes conflicts of addressing data.

Tree-based methods, such as Radix trees and Patricia trees, are the most standard methods for the table lookup in routers [5], [6]. Fig. 2 depicts the structure of the Radix tree as a representative. In the tree-based methods, binary trees are traced based on each bit of the destination IP address. Conclusive stopped nodes indicate the routes for the destination

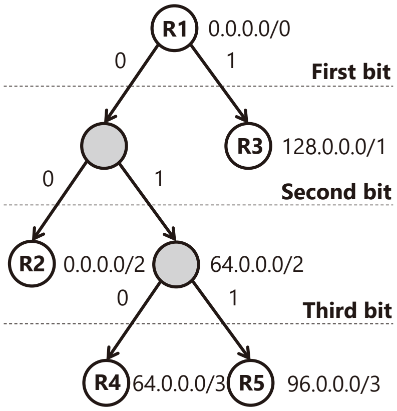


Fig. 2. Structure of Radix tree.

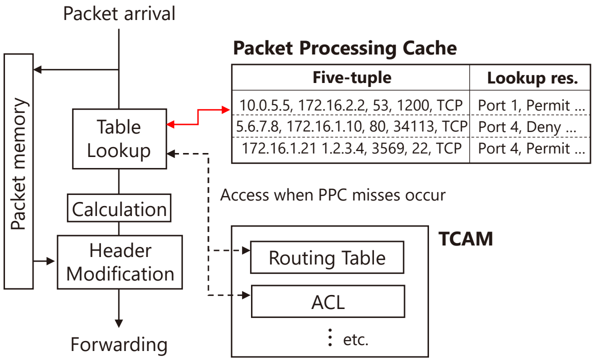


Fig. 3. Outline of table lookup with PPC.

IP address. This tracing process can be done through pipeline and thus, achieved high throughput.

A problem of the DRAM-based approach results from the large delay of DRAM accesses. The DRAM takes a long time to retrieve data due to the DRAM structure. Although routers must process packets each several nano seconds to achieve 400+ Gbps, as mentioned in Section 1, approximately 50 nano seconds are required for a DRAM access. The DRAM-based approach is comparatively slow in comparison to other approaches, while it can process packets with significant low power.

B. TCAM-based Approach

A TCAM-based approach is employed in routers which require high throughput, such as core routers [7]–[10]. This approach stored the tables into TCAM. The TCAM is a specialized memory for fast data lookup and can search data at one cycle. Because the access latency of TCAM is also small (approximately several nano seconds), the TCAM-based approach can achieve high throughput.

A problem of the TCAM-based approach is that TCAM consumes significant large power for data lookup. It is because TCAM compares all stored bits simultaneously to obtain data at one cycle. The papers [8] reported that TCAM consumes 150 times larger energy than a same sized RAM. Therefore, this approach achieves high throughput at the sacrifice of power consumption. According to the papers [11], TCAM consumes 40% of total power consumed in a router with TCAM.

C. IP Cache

IP Cache is a supplemental approach of the DRAM- and TCAM-based approaches and accelerates the table lookup with reducing the power consumption [12]–[15]. IP Cache is placed before accessing the DRAM or TCAM. It stores table lookup results of the routing table and ARP table per destination IP address into a small SRAM (static RAM) and references them to process subsequent packets. Because the packets of the same destination IP address indicate the same routing table lookup result and ARP table lookup result, IP Cache can reduce the number of DRAM or TCAM accesses if IP Cache has the table lookup results of the packets.

A small SRAM used as IP Cache shows low access latency and low power consumption. For example, the access latency and energy consumption of a 32KB SRAM, which is conventionally used in IP Cache, are 0.5 ns and 0.0539 nJ per access, respectively, while those of TCAM are 5 ns and 30 nJ per access, respectively [18]. Thus, it is important for IP Cache to process packets using only SRAM as many as possible.

The entries of IP Cache are constructed by 4 bytes tag (i.e., a destination IP address) and 6+ bytes data (i.e., table lookup results of the routing table and ARP table). Thus, a typical 32KB IP Cache has approximately 3K entries. Conventionally, IP Cache is configured by a 4-way set associative cache, and the entries are mapped using the CRC hash values calculated from destination IP addresses. The paper [22] showed that a 32KB 4-way IP Cache can achieve cache hit rates from 80% to 90%. Moreover, IP Cache has a possibility to achieve the cache hit rate of up to 98% when increasing the capacity. It indicates that IP Cache has possibility to process most packets with only a SRAM.

The throughput and power consumption of the table lookup produced by IP Cache mainly depend on two factors: the cache hit rate and cache performance (i.e., the access latency and power consumption). The cache hit rate is the most important factor to determine the table lookup performance because it represents the rate of packets processed using the SRAM. To achieve a high cache hit rate, increasing the SRAM capacity is the most simple solution. However, it is not suitable considering the SRAM performance. For example, the latency of a 1MB SRAM is almost the same as that of TCAM. Thus, achieving a high cache hit rate with small capacity SRAM is required to exhibit further performance.

D. Packet Processing Cache

Similar to IP Cache, PPC has been proposed as a supplemental approach of mainly the TCAM-based approach to realize both accelerating the table lookup and reducing the power consumption [16]–[18]. PPC solves the problem of the TCAM-based approach that TCAM consumes large power. Unlike IP Cache, PPC caches lookup results of all kinds of tables in a router (i.e., four tables) per flow into a SRAM. If

PPC has the table lookup results of a flow, subsequent packets of the same flow can be processed without accessing TCAM. Consequently, PPC can substitute one PPC access for four TCAM accesses while IP Cache substitutes one cache access for two TCAM accesses (i.e., the routing table access and ARP table access).

Fig. 3 shows the outline of the table lookup with PPC. PPC stores table lookup results per five-tuple (i.e., source and destination IP addresses, source and destination port numbers, and protocol number), which is called a flow in PPC. It is because most tables in routers store data based on the some or all the five-tuple. PPC can process subsequent packets of flows using only PPC by caching the table lookup results of the first packets of the flows into PPC. In comparison to IP Cache, although flows have a smaller temporal locality than packets of the same destination IP address, PPC can process packets with one PPC access. Thus, PPC has a possibility to further improve the throughput and power consumption of the table lookup in comparison to IP Cache if it achieves a high PPC hit rate.

By combining PPC with the TCAM-based approach, it can achieve ultrahigh table lookup throughput. However, the power consumption is still high due to remaining TCAM accesses. The paper [19] indicated that the dynamic energy of TCAM was still dominant in a router even if PPC achieves the PPC hit rate of 95%.

The entries of PPC are constructed by 13 bytes tag (i.e., five-tuple) and 15+ bytes data (i.e., table lookup results of the four tables). Thus, a typical 32KB PPC has approximately 1K entries. Similar to IP Cache, conventional PPC is configured by a 4-way set associative cache, and the entries are mapped using the CRC hash values calculated from five-tuples. Because the PPC entry size is larger than the IP Cache entry size, the PPC hit rate tends to become lower than IP Cache. A 32KB PPC shows the PPC hit rate of approximately 70% [23].

E. Cache Miss Handler

CMH was proposed in [20], [21] to assist PPC and enable to process packets without blocking. When a packet of a flow misses in PPC, the router must block the table lookup operation of subsequent packets of the same flow until the entry of the flow is prepared in PPC. It is because subsequent packets may continuously miss in PPC before completing the table lookup of the flow and updating it in PPC. Thus, CMH queues subsequent packets of the flow until the PPC update of a former packet of the flow is completed.

Fig. 4 illustrates the overview of CMH. It is placed between PPC and TCAM and accessed if packets miss in PPC. CMH is composed of two modules: CMT (cache miss table) and CMQ (cache miss queues).

CMT is implemented by a small full-associative cache and manages the flows which are being processed in TCAM. When a packet missed in CMT, CMT stores the five-tuple and set the valid bit on if a CMT entry is empty. The subsequent packets with the same five-tuple are sent to CMQ, which is simple FIFOs, by hitting in CMT. At this time, the queue number of CMQ is determined based on the hit address of CMT. After the CMT miss packet is processed in TCAM, the release signal and CMT address (i.e., the queue number) are sent to CMT and CMQ. Finally, the CMT entry is disabled by setting the valid bit off, and the packets queued in CMQ are released with the TCAM lookup results.

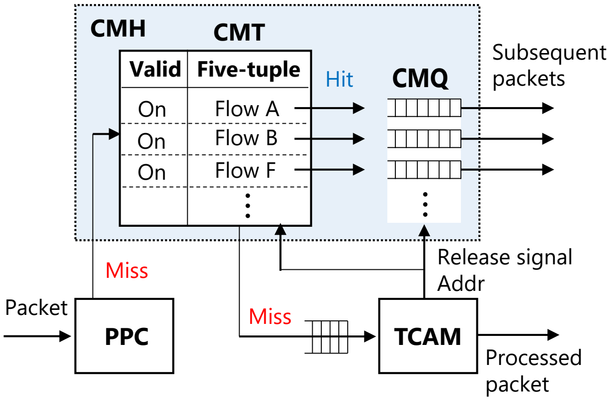


Fig. 4. Overview of CMH.

IV. Hybrid Memory Design for High-throughput and Low-energy Table Lookup

As explained in Section 3 and summarized in Table I, each table lookup approach contributes to increasing the throughput or reducing the power consumption. However, there are no approaches to realize ultrahigh throughput at significant low power consumption. For example, the combination of PPC, CMH, and TCAM shows the highest throughput in the approaches shown in Table I. However, it still consumes large power due to the remaining TCAM accesses. Our targets of the throughput and power consumption are shown as bold in Table I. To meet these requirements, this study proposes a novel hybrid memory design for high-throughput and low-power table lookup in routers.

Fig. 5 depicts an overview of the proposed memory design. In this design, the five approaches introduced in Section 3 are combined using four buffers, called PPC buffer, Victim buffer, Table buffer, and DRAM buffer, in suitable order. Because the latter memory lookup is slower in the proposed memory design, the buffers are required. Each buffer enables to operate the former memory lookup independently of the latter memory lookup. If a buffer is filled, the former memory lookup is stopped until the buffer becomes available. Details of each combination are explained hereinafter.

1) Combination of PPC and CMH: The combination of PPC and CMH was already considered in [20], [21], and thus, the processing follows these papers. Packets missed in PPC are sent to CMH and judged whether hit or missed in CMT. When a packet hits in CMT, the packet is sent to CMQ and queued until the former packet of the same flow is processed by DRAM or TCAM. On the other hand, when a packet misses in CMT, the packet is sent to IP Cache, and the new CMT entry is registered, as mentioned in Section 3.

In the proposed memory design, the number of CMT

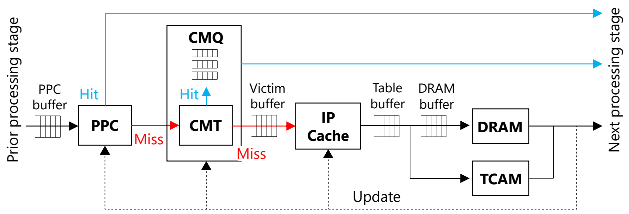


Fig. 5. Overview of proposed memory design.

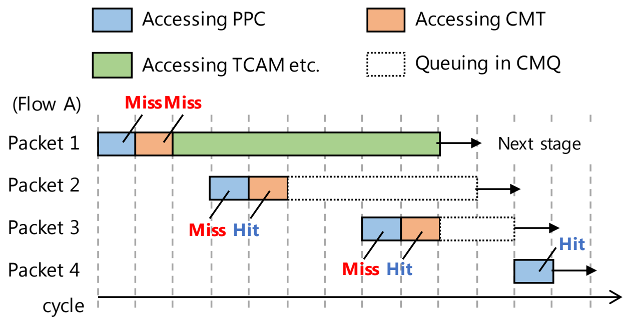


Fig. 6. Time chart of processing with our memory design.

entries becomes the same as the summation of the number of buffer entries of Victim buffer, Table buffer, and DRAM buffer to manage all flows being processed in the latter memories. Larger buffer sizes enable to process a larger number of packets without blocking and achieve higher table lookup throughput. However, there is a trade-off between the buffer sizes and the implementation costs of the buffers, CMT, and CMQ. Consequently, deciding the number of CMT entries and buffer sizes considering both the throughput and implementation costs is important.

We newly discuss the behavior of the combination of PPC and CMH. First, CMH guarantees the order of packets processed in the proposed memory design at flow level. Fig. 6 shows this behavior using a time chart of the processing. Packets of the same flow are sent to the next processing stage in the arrival order, and out of order due to the latency gap between PPC and DRAM or TCAM does not occur in this memory design. It is because CMH keeps subsequent packets of a flow waiting until the former packet of the flow is processed in DRAM or TCAM. Second, CMH does not require a buffer between PPC and CMT. It is because the CMT access is always faster than PPC due to the small capacity.

2) Combination of CMH and IP Cache: The combination of PPC and IP Cache was also considered in [22]. In this study, IP Cache is placed after CMH, and packets which missed in CMT are sent to IP Cache after queuing in Victim buffer. Because IP Cache enables to process packets based on the destination IP address, it has a possibility to achieve a significant higher hit rate than PPC. Moreover, the proposed memory design allows IP Cache to make the capacity larger because packets which sent to IP Cache is significantly small in comparison to PPC, and large IP Cache latency is permissible. It also induces an increase in the cache hit rate. Note that packets are sent to Table buffer regardless of the hits or misses in IP Cache because packets must access DRAM or TCAM even if they hit in IP Cache.

3) Combination of DRAM and TCAM: In the proposed memory design, the DRAM and TCAM are combined to reduce the power consumption with increasing the throughput. To meet the requirement of the power efficiency, packets should be processed by the DRAM as many as possible, especially when packet arrival is a slow. The TCAM is used in the case that the DRAM lookup is too late for processing packets.

To realize these behaviors, a simple method using two buffers is proposed in this study. As depicted in Fig. 5, packets which missed in IP Cache are first sent to Lookup buffer. The packets queued in Lookup buffer are next sent to DRAM buffer, which placed before the DRAM, until DRAM buffer is filled up. If DRAM buffer is full, the packets queued in Lookup buffer are sent to the TCAM and processed using the TCAM. Thus, the TCAM are utilized only when the DRAM is busy, and the power consumed by the table lookup can be reduced as large as possible. In this study, we consider that one entry is enough as the DRAM buffer size. It is because increasing the DRAM buffer size causes the increase in CMT and CMQ entry sizes. Moreover, the DRAM buffer size does not significantly impact on the throughput and power consumption.

After a packet is processed using the DRAM or TCAM, the corresponding CMT entry and packets queued in CMQ are released, and the table lookup results are cached into IP Cache and PPC. The processed packet and queued packets are sent to the next processing stage in the arrival order.

V. Evaluation

This section shows the evaluation of the proposed memory design. In this evaluation, the table lookup operation in a router is simulated using an in-house cycle-accurate table lookup simulator and packet traces (i.e., pcap files) captured in real networks. The throughput and power consumption of the table lookup can be measured using this simulator. In this study, the evaluation was done based on following points.

- Effect of the combination of DRAM and TCAM
- Comparison to other approaches

A. Simulation Environment

1) Cycle-accurate Table Lookup Simulator: To evaluate the proposed memory design, an in-house table lookup simulator, written in C++, was used. This simulator can simulate the table lookup operation in a router including the queuing and stalling at cycle level. The architecture of the simulator was modeled in Fig. 5.

Table II shows the parameters set in the simulator and the reference values. In the following simulations, the reference values are used if the values are not written clearly. These values were mainly decided based on the previous studies such as [18], [20], [22]. The latency, dynamic energy, and static power of PPC, CMH, IP Cache, and DRAM were estimated using CACTI 7.0 [24], which was a major tool for estimating the latency and power consumption of various

TABLE II. Simulator parameters and reference values.

	Item	Reference value
	Operation frequency	2 GHz (0.5 ns / cycle)
PPC	Associativity	4 ways
	Entries	1,024 entries
	Replacement	LRU
	Latency	1 cycle
	Dynamic energy	0.0342 nJ
	Static power	40 mW
CMH	Associativity	full associative
	CMT entries	32 entries
	CMQ entries / queue	10 entries
	Latency	1 cycle
	Dynamic energy	0.0346 nJ
	Static power	4.53 mW
IP Cache	Associativity	4 ways
	Entries	4,096 entries
	Replacement	LRU
	Latency	1 cycle
	Dynamic energy	0.0208 nJ
	Static power	12.8 mW
DRAM	Latency	10 cycles
	Number of ports	4 ports
	Dynamic energy	1.26 nJ
	Static power	13.9 mW
TCAM	Latency	100 cycles
	Number of ports	4 ports
	Dynamic energy	166 nJ
	Static power	17 mW
Buffers	PPC buffer	unrestricted
	Victim buffer	10 entries
	Lookup buffer	10 entries
	DRAM buffer	1 entries

TABLE III. Simulator parameters and reference values.

Trace Name	Bandwidth	Number of packets	Duration
WIDE	1 Gbps	22,172,838 packets	900 s
Academic	10 Gbps	2,411,883 packets	90 s

types of memories, while those of TCAM were estimated from recent CAM's studies [25], [26]. We also note that the numbers of ports in a DRAM and TCAM were set to four because they can search the four tables in a router independently.

2) Packet Traces: As workloads, two pcap-format packet traces captured in real networks were used. Details of them are summarized in Table III. WIDE (Widely Integrated Distributed Environment) trace contains communication traffic at the 1-Gbps transit link of WIDE to the upstream ISP and can be obtained from [27]. Academic trace contains communication traffic at a 10-Gbps core link in a institute which mixed various University traffic and is not opened to the public. The pcap files include information of a packet per line, namely, the arrival time, each header information (i.e., Ethernet header, IP header, and TCP/UDP headers), and some payload. The simulator can obtain packets by reading each line in the pcap files.

3) Three Simulation Conditions: To evaluate the throughput and power consumption based on the practical use, this study conducted simulations under the three conditions: the full load, 400-Gbps load, and 100-Gbps load.

The full-load simulation was conducted to measure the achievable throughput and maximum power consumption. It starts the table lookup simulation without considering the arrival time of packets. It means that the simulation is started under the situation that all packets are queued in PPC buffer. In routers, the throughput in the situation that packets are queued in buffers is the most important because packets are dropped if the throughput is insufficient. Note that this study measured the throughput on the assumption that all packets were constructed of 64 bytes (i.e., the shortest packet length). It is because routers conventionally show this worst-case throughput as an important barometer of the packet processing throughput.

TABLE IV. Comparison of table lookup performance with three different designs.

		DRAM-only	TCAM-only	Proposed
WIDE	Throughput	65 Gbps	660 Gbps	727 Gbps
	Power (100-Gbps)	89.9 mW	377 mW	142 mW
	(400-Gbps)	124 mW	1,430 mW	386 mW
	(Full-load)	152 mW	33,700 mW	33,800 mW
Academic	Throughput	84 Gbps	854 Gbps	929 Gbps
	Power (100-Gbps)	108 mW	936 mW	166 mW
	(400-Gbps)	189 mW	3,560 mW	1,390 mW
	(Full-load)	153 mW	33,700 mW	33,400 mW

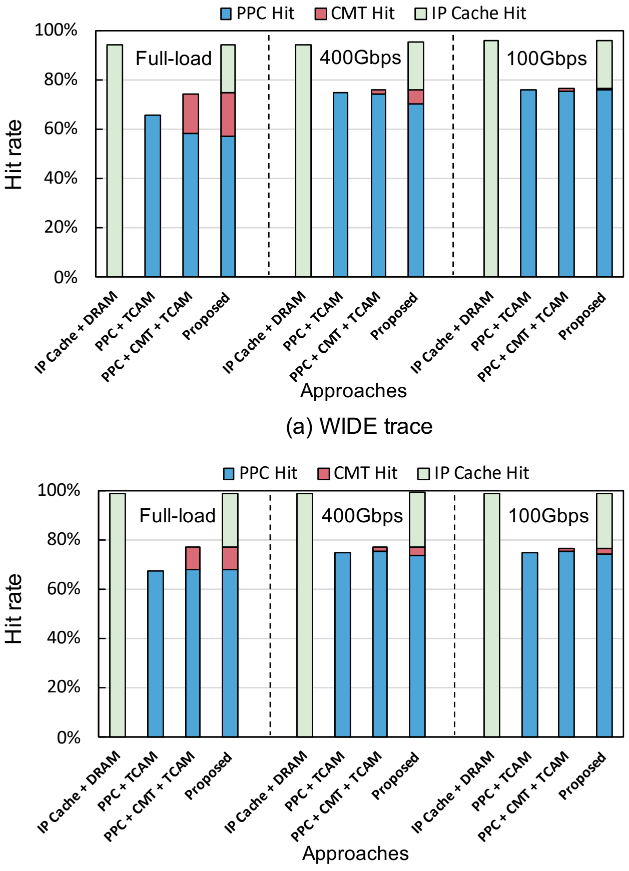
On the other hand, the 400- and 100-Gbps simulations were also conducted to measure the power consumption under the specific traffic loads. In these simulations, the arrival times of each packet were modified to satisfy the bandwidths of 400- or 100-Gbps considering the packet length, and packets were sent to the simulator in accordance with the modified arrival time. Note that the throughput measured in these simulations was not meaningful because the systems obviously had the capability to achieve 400-Gbps throughput (not the shortest-packet-length throughput). Beside the throughput, the power consumption measured in these simulations is important because routers are not always operating under a full load. Consuming the power under the full-load condition is rare for routers.

B. Effect of the Combination of DRAM and TCAM

First, effect of the combination of the DRAM and TCAM was evaluated. To reveal it, this study implemented the approaches that all packets which missed in IP Cache were assigned to only the DRAM or TCAM (referred to as DRAM-only design and TCAM-only design, respectively) for comparison.

Table IV summarizes the throughput and power consumption of the table lookup measured in the simulations. As shown in the table, the proposed memory design achieved significant higher throughput than the DRAM-only design, and it slightly overcome the TCAM-only design because the proposed memory design can process packets using both the DRAM and TCAM. According to the full-load simulations, the proposed memory design achieved 11.0x and 1.09x higher throughput than the DRAM-only and TCAM-only designs, respectively, on average. In Academic trace, the proposed memory design achieved nearly 1-Tbps throughput, and this result substantiated the previous studies which analyzed the throughput based on the mathematical model.

Table IV also shows usefulness of the proposed memory design from the aspect of the power consumption. In the full-load condition, the proposed memory design showed almost the same power of the TCAM-only design because most packets were assigned to the TCAM to speed up the table lookup. However, in the 100- and 400-Gbps conditions, the proposed memory design can significantly reduce the power in comparison to the TCAM-only design. The results showed



(a) WIDE trace

(b) Academic trace

Fig. 7. Comparison of hit rates and the breakdown.

that the proposed memory design can reduce the power consumption by 72.3% and 67.0% in the 100- and 400-Gbps conditions, respectively, in comparison to the TCAM-only design. These results indicate that the proposed memory design can significantly reduce the power consumption of the table lookup with keeping the throughput to the same level as the TCAM-only design.

C. Comparison to Other Approaches

This section reveals superiority of the proposed memory design to other approaches from the perspectives of the throughput and power efficiency. For comparison, this study implemented five conventional approaches: the TCAM-based approach, DRAM-based approach, combination of IP Cache and DRAM [12], combination of PPC and TCAM [16], and combination of PPC, CMT, and TCAM [23].

First, the cache hit rates achieved by each approach were evaluated. Fig. 7 showed the cache hit rates and their breakdown under the three load conditions. According to the results, the proposed memory design can achieve significant high hit rate (94% in WIDE trace and 99% in Academic trace) by combining PPC and IP Cache. The combination of IP Cache and DRAM also achieved high hit rate; however, it did not significantly impact on the table lookup performance because it required DRAM accesses even if packets hit in IP Cache.

In addition, the larger the load of a router increased, the more CMH assisted the cache hit rate. This is because the number of packets still waiting to be processed by the DRAM or TCAM becomes large when the load of a router increases. If a router does not employ CMH (i.e., in the case of *PPC + TCAM* in Fig. 7), this situation causes a large number of PPC misses because it takes time to update PPC entries. This behavior also reveals the reason that the PPC hit rate of the proposed memory design was a little lower than that of the *PPC + CMT + TCAM*. In the proposed memory design, there is a possibility that packets are waited for a longer time compared to TCAM-based approaches because packets may assign to the DRAM. Thus, the PPC hit rate may become a little lower; however, it is no problem for the table lookup performance because it can be saved by CMH, as shown in Fig. 7.

Second, the table lookup throughput and power efficiency were evaluated. Table V summarized them. In the table, the parenthesis values represent the power efficiency calculated from the power consumption divided by the throughput [mW/Gbps], which were often used in routers as a barometer of the router performance. The smaller power efficiency is more suitable for routers; however, considering not only this power efficiency but also the achievable throughput is important for routers.

As shown in Table V, the proposed memory design achieved 1.95x throughput on average compared to *PPC + CMT + TCAM*, which showed the largest throughput in recent studies. It realized up to nearly 1-Tbps table lookup throughput. The proposed memory design also had the advantage of the power efficiency in comparison to other TCAM-used approaches. It reduced the power consumption per Gbps by 86%, 89%, and 44% in 100-Gbps, 400-Gbps, and full-load conditions, respectively. In the full-load condition, the proposed memory design consumes the power as the same level of the other TCAM-used approaches because most packets are assigned to the TCAM. However, the power efficiency is better due to the high achievable throughput. In addition, the full-load condition is rare in routers in practical use. Consequently, it was showed that the proposed memory design can achieve significant high table lookup throughput with low power consumption as the same level of the DRAM-based approach.

VI. Conclusion

The table lookup is the most important operation in routers to determine the packet processing throughput and power consumption. Thus various approaches, such as DRAM-based approaches, TCAM-based approaches, and cache-based approaches, have been proposed and employed. However, there are no approaches to satisfy the requirements of both ultrahigh throughput and significant low power consumption.

For realizing them, this study proposed a novel hybrid memory design, which combines five conventional approaches (i.e., PPC, CMH, IP Cache, DRAM, and TCAM) in the appropriate order. The effectiveness of the proposed memory design was evaluated using an in-house simulator which can simulate the table lookup in a router at cycle level. The simulation results indicated that the proposed memory design achieved

TABLE V. COMPARISON OF TABLE LOOKUP PERFORMANCE WITH VARIOUS CONVENTIONAL APPROACHES.

		DRAM-based	TCAM-based	IPCache+DRAM	PPC+TCAM	PPC+CMT+TCAM	Proposed
WIDE	Throughput	10 Gbps	102 Gbps	19 Gbps	297 Gbps	401 Gbps	727 Gbps
	Power (100-Gbps)	91 mW	2,550 mW	41 mW	651 mW	638 mW	95 mW
		(12)	(333)	(5)	(85)	(83)	(12)
	(400-Gbps)	115 mW	10,320 mW	103 mW	2,640 mW	2,550 mW	287 mW
		(11)	(329)	(5)	(84)	(81)	(9)
	(Full-load)	115 mW	33,600 mW	103 mW	33,300 mW	33,700 mW	33,800 mW
		(11)	(328)	(5)	(112)	(84)	(47)
Academic	Throughput	10 Gbps	102 Gbps	20 Gbps	313 Gbps	444 Gbps	930 Gbps
	Power (100-Gbps)	115 mW	6,900 mW	103 mW	1,770 mW	1,660 mW	145 mW
		(11)	(330)	(5)	(84)	(80)	(7)
	(400-Gbps)	115 mW	29,800 mW	103 mW	7,510 mW	6,780 mW	1,407 mW
		(11)	(329)	(5)	(83)	(75)	(16)
	(Full-load)	115 mW	33,600 mW	103 mW	33,700 mW	33,700 mW	33,400 mW
		(11)	(328)	(5)	(108)	(76)	(36)

nearly 1 Tbps throughput with similar power consumption of the DRAM-based approach. If compared to the table lookup approach with PPC, CMT, and TCAM, which shows the largest throughput in recent studies, the proposed memory design achieved 1.95x throughput with 11% power consumption at 400-Gbps condition.

As future works, the increase in the PPC hit rate is one of the important issues to further improvement in the table lookup performance. Although the proposed memory design shows a high cache hit rate (more than 95%), it does not largely impact on the performance due to the remaining DRAM and TCAM accesses. To reduce the number of DRAM and TCAM accesses, improving the PPC hit rate is the most effective approach. Thus, more efficient PPC design should be considered.

ACKNOWLEDGMENT

This work was supported by the JSPS KAKENHI Grant Number JP18K18022.

REFERENCES

[1] H. Kawase, Y. Mori, H. Hasegawa, and K.-c. Sato, "Dynamic router performance control utilizing support vector machines for energy consumption reduction," *IEEE Transactions on Network and Service Management*, vol. 13, no. 4, pp. 860–870, 2016.

[2] T. Song, Z. Jiang, Y. Wei, X. Shi, X. Ma, O. Ormond, M. Collier, and X. Wang, "Traffic aware energy efficient router: Architecture, prototype and algorithms," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3814–3827, 2016.

[3] Y. Li, L. Zhu, S. K. Bose, and G. Shen, "Energy-saving in ip over wdm networks by putting protection router cards to sleep," *Journal of Lightwave Technology*, vol. 36, no. 14, pp. 3003–3017, 2018.

[4] P. Liu, S. R. Chaudhry, X. Wang, and M. Collier, "Progres: A programmable green router with controlled service rate," *IEEE Access*, vol. 7, pp. 143 792–143 804, 2019.

[5] K. Sklower, "A tree-based packet routing table for berkeley unix," in *Proc. of USENIX Winter Conference*, 1991, pp. 93–104.

[6] S. Nilsson and G. Karlsson, "Ip-address lookup using lc-tries," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 6, pp. 1083–1092, 1999.

[7] A. J. McAuley and P. Francis, "Fast routing table lookup using cams," in *Proc. of the 12th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '93)*, vol. 3, 1993, pp. 1382–1391 vol.3.

[8] F. Zane, G. Narlikar, and A. Basu, "Coolcams: Power-efficient tcams for forwarding engines," in *Proc. of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '03)*, vol. 1. IEEE, 2003, pp. 42–52.

[9] W. Li, X. Li, and H. Li, "Meet-ip: Memory and energy efficient tcam-based ip lookup," in *Proc. of the 26th International Conference on Computer Communication and Networks (ICCCN)*, 2017, pp. 1–8.

[10] E. Norige, A. X. Liu, and E. Torng, "A ternary unification framework for optimizing tcam-based packet classification systems," *IEEE/ACM Transactions on Networking*, vol. 26, no. 2, pp. 657–670, 2018.

[11] G. C. Sankaran and K. M. Sivalingam, "Design and analysis of fast ip address-lookup schemes based on cooperation among routers," in *Proc. of the 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)*, 2020, pp. 330–339.

[12] D. C. Feldmeier, "Improving gateway performance with a routing-table cache," in *Proc. of the 7th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '88)*, 1988, pp. 298–307.

[13] R. Jain, "Characteristics of destination address locality in computer networks: A comparison of caching schemes," *Computer networks and ISDN systems*, vol. 18, no. 4, pp. 243–254, 1990.

[14] X. Chen, "Effect of caching on routing-table lookup in multimedia environment," in *Proc. of the 10th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '91)*, vol. 3, 1991, pp. 1228–1236.

[15] S. Ravinder, M. A. Nascimento, and M. H. MacGregor, "Two-level cache architecture to reduce memory accesses for ip lookups," in *Proc. of the 23rd International Teletraffic Congress (ITC)*, 2011, pp. 278–285.

[16] Y. Tung and H. Che, "A flow caching mechanism for fast packet forwarding," *Computer Communications*, vol. 25, no. 14, pp. 1257–1262, 2002.

[17] S. Gamage and A. Pasqual, "High performance parallel packet classification architecture with popular rule caching," in *Proc. of the 18th IEEE International Conference on Networks (ICON)*, 2012, pp. 52–57.

[18] H. Yamaki, H. Nishi, S. Miwa, and H. Honda, "Data prediction for response flows in packet processing cache," in *Proc. of the 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, 2018, pp. 1–6.

[19] K. Tanaka, H. Yamaki, S. Miwa, and H. Honda, "Multi-level packet processing caches," in *Proc. of the 2019 IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS)*, 2019, pp. 1–3.

[20] M. Okuno, S. Nishimura, S. Ishida, and H. Nishi, "Cache-based network processor architecture: Evaluation with real network traffic," *IEICE trans. on electronics*, vol. 89, no. 11, pp. 1620–1628, 2006.

[21] M. Okuno and H. Nishi, "Network-processor acceleration-architecture using header-learning cache and cache-miss handler," in *Proc. of the 8th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2004)*, 2004, pp. 108–113.

[22] H. Yamaki, "Efficient cache architecture for packet processing in internet routers," in *Proc. of the 2020 Future of Information and Communication Conference (FICC 2020)*, vol. 1, 2020, pp. 338–352.

[23] H. Yamaki, "Flow characteristic-aware cache replacement policy for packet processing cache," in *Advances in Information and Communication Networks*, vol. 886, 2019, pp. 258–273.

[24] R. Balasubramonian, A. B. Kahng, N. Muralimanohar, A. Shafiee, and V. Srinivas, "Cacti 7: New tools for interconnect exploration in innovative off-chip memories," *ACM Trans. Archit. Code Optim.*, vol. 14, no. 2, pp. 1–25, 2017.

[25] T. Venkata Mahendra, S. Mishra, and A. Dandapat, “Self-controlled high-performance precharge-free content-addressable memory,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 8, pp. 2388–2392, 2017.

[26] S. Mishra, T. V. Mahendra, and A. Dandapat, “A 9-t 833-mhz 1.72-fj/bit/search quasi-static ternary fully associative cache tag with selective matchline evaluation for wire speed applications,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 11, pp. 1910–1920, 2016.

[27] WIDE MAWI WorkingGroup, “Mawi working group traffic archive,” http://mawi.wide.ad.jp/mawi/, (Accessed on 10/01/2019).

Artificial Intelligence: What it Was, and What it Should Be?

Hala Abdel Hameed
Information Systems Dep. Faculty of Computers and Information ,Fayoum University. Cairo, Egypt
Computer Science Dep. Khaybar Community College, Taibah University, KSA

***Abstract*—Artificial Intelligence was embraced as an idea of simulating unique abilities of humans, such as thinking, self-improvement, and expressing their feelings using different languages. The idea of "Programs with Common Sense" was the main and central goal of Classical AI; it was, mainly built around an internal, updatable cognitive model of the world. But, now almost all the proposed models and approaches lacked reasoning and cognitive models and have been transferred to be more data driven. In this paper, different approaches and techniques of AI are reviewed, specifying how these approaches strayed from the main goal of Classical AI, and emphasizing how to return to its main objective. Additionally, most of the terms and concepts used in this field such as Machine Learning, Neural Networks and Deep Learning are highlighted. Moreover, the relations among these terms are determined, trying to remove mysterious and ambiguities around them. The transition from the Classical AI to Neuro-Symbolic AI and the need for new Cognitive-based models are also explained and discussed.**

Keywords—Classical AI; machine learning; Neuro-Symbolic AI; Cognitive-based AI; deep learning

I. Introduction

Artificial intelligence (AI) is a wide-range models that empowers people to incorporate and analyze data to make insights and predictions that could be used in the decision making process, which is normally requires sufficient level of human expertise. In its early decades, the main challenge facing the artificial-intelligence researches was to learn the machines how to make relations between different states and a set of recognizable conditions, which have been maintained in its earlier models. During 1980s, AI models achieved a great value for probabilistic explanations over a set of discrete variables, e.g. machines can make interpretations and guess that a patient with specified symptoms may have a certain disease.

One of main objectives of AI models was to help people to anticipate problems or deal with issues as they come up, and operate in an intentional, and in an adaptive way. Despite the importance of the aforementioned objective, in 2015, Google apologized to a software engineer Jacky Alciné after he pointed out that the image recognition algorithms in Google Photos were classifying his black friends as "*gorillas*". Also, Google was algorithmically biased and showed an advertisement of a job to a male group rather than women [reported in Washington Post on July 6, 2015]. Another example that indicates a failure of AI systems, is when a street-sign recognition system used by self-driving cars mistaking the stop signs for speed limit with a little defacing. All these examples indicate the misbehaving of the current AI systems comparing to humans who can learn logical relations and make choices with little information. AI techniques, on the other hand, are more restricted in their abilities and require specific details to do their work.

The main and central objective of this paper is to illustrate how the AI field has been changed and deviated from its main goal, which causes that robust intelligence cannot be achieved. The paper also asserted that, without developing systems able to represent and reason the external world, and draw on substantial knowledge about its dynamics, this robustness will never be achieved. Recently, a lot of papers and researchers realized the importance of moving towards more adaptive, dynamic, and cognitive models. In addition, they provided comprehensive studies of the past, present and future of AI field, such as the work done in [1], [2].

The rest of this paper is organized as follows; In Section 2, an overview of different disciplines of AI is presented, while in Section 3, an overview of the history of the AI field was provided. Section 4 demonstrates how data-driven models overwrite the main goal of classical AI. Three different types of AI, Narrow, General, and Super AI are highlighted in Section 5, and the main challenges facing the Current AI are outlined in Section 6. The difference between Knowledge-Based, Cognitive-Based model and Consciousness is shown in Section 7. The importance of using a hybrid approach is discussed and explained in Section 8, while we conclude our study in Section 9.

II. The Disciplines and Terms of AI

In this section, different disciplines of AI that contribute to the emergence of the field are outlined. Also, the main terminologies and terms used are reviewed, keeping in mind, removing the ambiguities associated with them in several works of literature.

According to Russell & Norvig [3], different disciplines including Philosophy, Mathematics, Neuroscience, Economics, Computer engineering, Control theory, and Linguistics all together contribute to formalizing the principles of AI Philosophy that formulates a precise set of laws governing the rational part of the mind which allowed one to generate conclusions mechanically, given initial premises. Mathematics is the second foundation that formalizes the formal logic, computation, and probability. Economics which studies how to make decisions that maximize the profit is another foundation of AI, while Decision theory that consolidates probability theory with utility theory, to give a complete framework for decisions made under uncertainty. Also, Neuroscience, which studies the nervous system, is one of its main foundations, Camillo Golgi [4], was the first one who developed a staining technique allowing the observation of individual neurons in the brain. And Nicolas Rashevsky [5], was the first to apply mathematical models to the study of the nervous system. Another AI foundation is "Computer Engineering" which answers the question of how we can build an efficient computing machine. Economics which study how people make choices that lead to preferred outcomes is another foundation of AI. And finally, both Control Theory and Linguistics are the last two founders of AI, the first answers the question of how can machines operate under their own control, and linguistics main concern is how does language relate to thought.

AI, Machine Learning (ML), Deep Learning (DL), and Artificial Neural Networks (ANNs) are often used interchangeably, but this is not true; "Fig. 1" illustrates the relations between these different terms, it shows the relation between Symbolic Artificial Intelligence [it will be discussed in detail in section7], and the Current AI. Artificial Intelligence or sometimes called Narrow or Weak AI is s a broader concept, which is briefly, study how machines are used to simulate the way of thinking and perform the mental functions in an "intelligent" way.

Machine learning is a set of AI techniques that study how machines can learn from a dataset and perform new predictions based on that prior learning. Deep Learning, Artificial Neural Networks (ANN), or sometimes called Connectionist AI (duo to its structure as connections), includes algorithms that simulate the mental functions to detect patterns, and classify information. Current DL techniques include Supervised, Unsupervised, Semi-supervised, and Active Learning with different algorithms. On the other hand, Rule-Based AI is a synonym for Symbolic-AI which is a traditional way of representing the problem by applying specified rules to an input, and accordingly, the output is governed by those provided rules. In "Fig. 2", different algorithms for these categories are specified and listed.

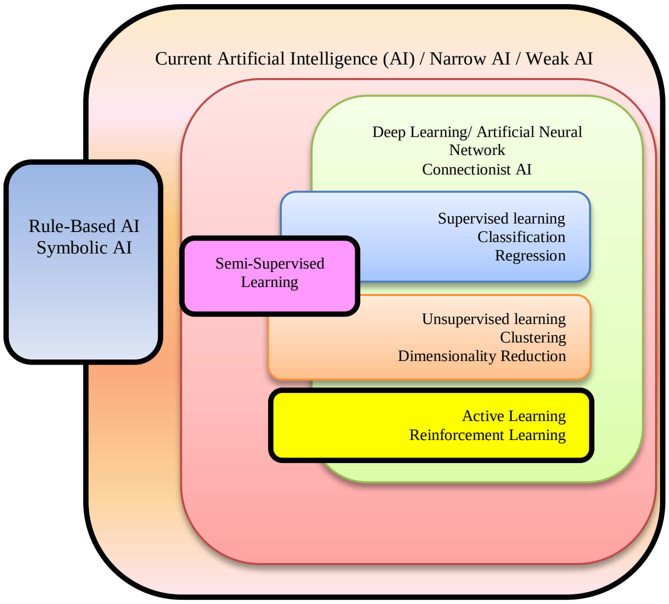


Fig. 1. The Big Picture of the Current AI.

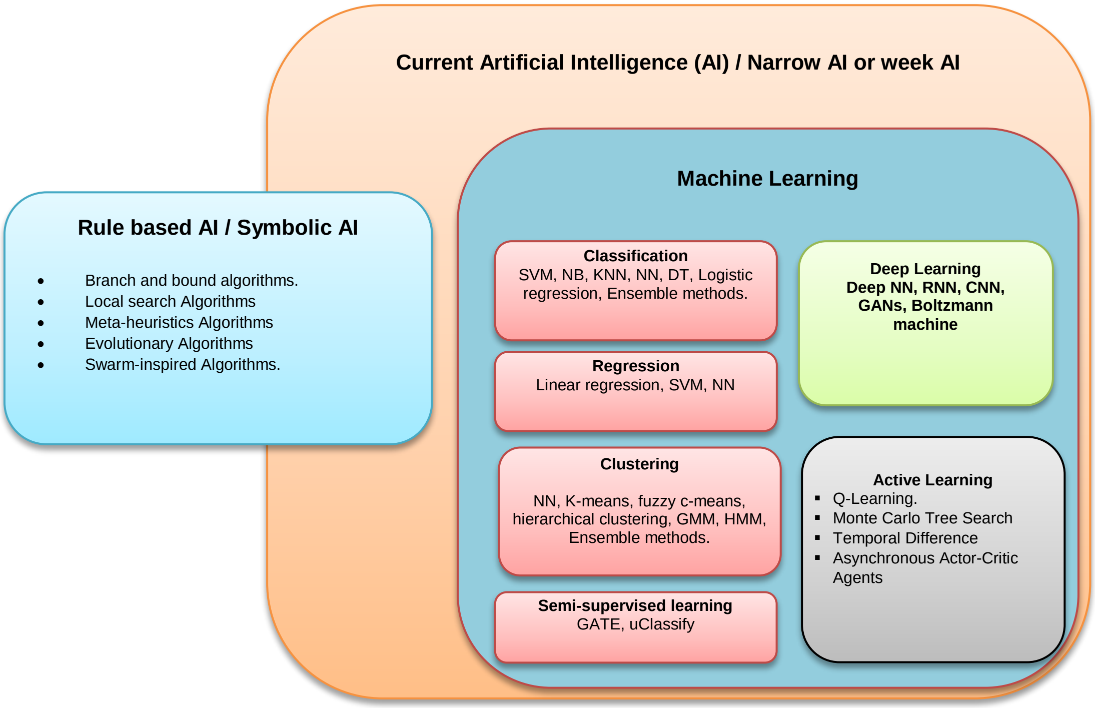

Current Artificial Intelligence (AI) / Narrow AI or week AI
Machine Learning
Rule based AI / Symbolic AI
Branch and bound algorithms.
Local search Algorithms
Meta-heuristics Algorithms
Evolutionary Algorithms
Swarm-inspired Algorithms.
Classification
SVM, NB, KNN, NN, DT, Logistic regression, Ensemble methods.
Regression
Linear regression, SVM, NN
Clustering
NN, K-means, fuzzy c-means, hierarchical clustering, GMM, HMM, Ensemble methods.
Semi-supervised learning
GATE, uClassify
Deep Learning
Deep NN, RNN, CNN, GANs, Boltzmann machine
Active Learning
Q-Learning.
Monte Carlo Tree Search
Temporal Difference
Asynchronous Actor-Critic Agents

Fig. 2. Algorithms used by Machine Learning and Rule-Based AI.

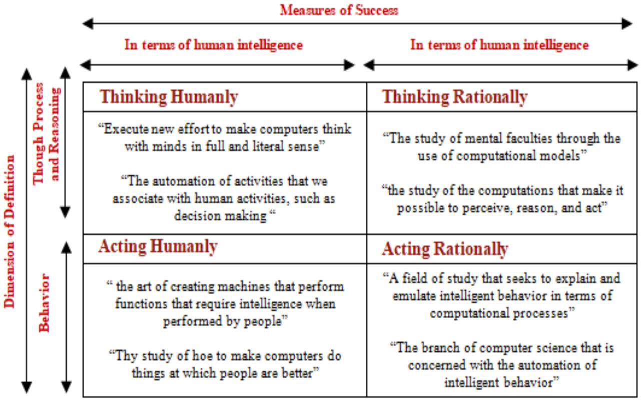
III. The Birth and Definition of AI

In this section, we will give a brief historical overview of the AI and its main idea. John McCarthy is an influential figure in AI, and Princeton is considered the true birthplace of AI, McCarthy. Minsky and others have organized a workshop for two months at Dartmouth in the summer of 1956 [6] and invited American researchers interested in automata theory, neural nets, and the study of intelligence. The proposal of the workshop, mainly stated that;" AI study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence a machine can be made to simulate it" [7]. Their main target was attempting to find how to make machines use language, form abstractions, solve problems reserved for humans, and improve themselves. Furthermore, they aimed at developing machines that will function autonomously in complex, changing environments.

Different definitions have been proposed to Artificial Intelligence from multiple dimensions. In the 1950s, Alan Turing provided an operational definition of intelligence that measured how the computer acting humanly through his proposed test [8], which is briefly, includes the following abilities for the computer:

- Ability to communicate successfully in English using natural language processing.
- Ability to store what it knows or hears using knowledge representation models.
- Automated reasoning ability which uses the stored information to answer questions and to draw new conclusions.
- Ability to adapt to new circumstances and detect and extrapolate patterns.

Wilson and Keil in 1999 [9] presented another definition known as The Cognitive Modeling definition, which is based on the dimensions of measuring how the computer can think and act humanly, in contrast with think and act rationally, their classification is shown in “Fig. 3”.


Measures of Success
In terms of human intelligence
In terms of human intelligence
Dimension of Definition
Though Process and Reasoning
Behavior

Thinking Humanly	Thinking Rationally
“Execute new effort to make computers think with minds in full and literal sense” “The automation of activities that we associate with human activities, such as decision making “	“The study of mental faculties through the use of computational models” “the study of the computations that make it possible to perceive, reason, and act”
Acting Humanly	**Acting Rationally**
“ the art of creating machines that perform functions that require intelligence when performed by people” “Thy study of hoe to make computers do things at which people are better”	“A field of study that seeks to explain and emulate intelligent behavior in terms of computational processes” “The branch of computer science that is concerned with the automation of intelligent behavior”

Fig. 3. Cognitive Modeling Definition of AI.

Moreover, their ability to do these things is going to increase rapidly until—in a visible future—the range of problems they can handle will be coextensive with the range to which the human mind has been applied” [10]. The overconfidence and promising performance of the early AI systems on simple examples were due to applying a simple syntactic manipulation of the problem, which is not suitable for large and difficult problems. Unfortunately, the early systems turned out to fail miserably when tried out on wider selections of problems and on more difficult problems. As an example, the translation project initiated in 1957 by the USA National Research Council is considered a typical example of early AI failure. The translation project was proposed to translate the Russian space scientific papers using simple words' replacement based on both the Russians and English grammar. The project was canceled, and it was stated that there was no machine could be used to translate human languages. The failure has been explained as; it was not sufficient to get the right meaning while the program of translation requires good background knowledge in order to resolve ambiguity and establish the content of the sentence.

IV. Data-Driven Models Built on Ruins of Classical AIx

“Programs with Common Sense” was the main and central concern of Classical AI. John McCarthy noted the value of commonsense knowledge in his pioneering paper [11], and Doug Lenat provided a representation of commonsense knowledge in a machine-interpretable form in his work [12][13][14]. The classical AI was, mainly built around an internal, updatable cognitive model of things, like individual people and objects, their properties, and their relationships with one another. But, almost all the recent models and approaches are lacking both reasoning, and rich cognitive models of the world [15], this may be due to the following reasons:

1) It was thought that using reasoning and data-cognitive models may be suitable for that small problem instance, while the scale of the problem has a proportional relation with sufficient hardware and larger memories.

2) To some extent, building human knowledge into machine learning systems has even been viewed within machine learning circles as cheating, and certainly not as desirable.

3) The complexity of the world is endless, and human minds are very complicated.

4) Lack of the essential methods used to capture the arbitrary complexity by finding good approximations of the world, there is a need to propose new AI systems that can discover like a human, not reinvent what he has already known.

Moreover, many saw this lack of encoded explicit knowledge or detailed cognitive models as an advantage rather than being anomalous; as they moved from classical AI and its core, towards different, more data-driven paradigms.

V. General, Narrow and Super AI

The paper written in 1958 by John McCarthy, introduced what is recently named General AI (AGI) concept, a hypothetical program (Advice Taker) was described, which considered the first complete AI system. In this system, axioms were defined to allow a model to generate a program to drive to the airport. The program was also developed to react autonomously to unexpected situations without being reprogrammed.

Narrow Intelligence, also known as "*Weak AI*" includes systems that perform a single narrow goal extremely well (e.g. chess playing). They are extremely centered around a single task and not robust and transferable to even modestly different circumstances. Such systems often work impressively well when applied to the exact environments on which they are trained, but in many cases, they are not reliable when the environment differs from that they are trained. Such systems have been shown to be powerful in the context of games, but have not yet proven adequate in the dynamic, open-ended flux of the real world.

When AI systems outperform the best human brains, Artificial Super Intelligence (ASI) will be achieved. [16]. In a public talk, Andrew NG, one of the key figures of AI, said that "the distance between AGI and ASI is very short; it may happen in mere months, weeks, or maybe the blink of an eye and will continue at the speed of light". Scientists have different views about that time.

VI. Problems of the Current AI

Despite the remarkable achievement accomplished by the AI in numerous applications, huge numbers of its initiators including McCarthy [17], Marvin Minsky, and Jouda Pearl accepted that it is strayed from its principle thought "machines that think, that learn and that make" as expressed in Simon's first workshop.

Classic Artificial Intelligence begins to break when it starts managing the untidiness of the world; for example, in image processing applications, in which computers used to gain high-level abstraction from digitized images. Consider the possibility that is needed to make a program recognize a cat; what number of rules is needed to make it. Another example is how it might be required to characterize the standards for a self-driving vehicle to identify all the various people on foot it may confront.

To illustrate the idea, consider Fig. 4 of a picture, which is known as a "Bongard Problem," named after its creator, Russian researcher Mikhail Moiseevich Bongard [18]. The problem is presented by two arrangements of pictures (six on the left and six on the right), the objective is to spot the key contrast between the two sets. As shown in the figure below, pictures in the left set contain one object, and pictures in the right set contain two objects. Although, it's simple for people to reach such inferences from such limited quantities of tests, yet there's still no neural network that can take care of the Bongard problem. In one investigation directed in 2016, computer-based intelligence scientists prepared a NN on 20,000 Bongard tests and tried it on 10,000 more; the NN's performance was much lower than that of average humans.

In the literature, there was a set of arguments and objections stated to answer the main question, "can a machine be intelligent?" Turing himself inspected a wide assortment of potential issues for building intelligent machines. In the following part, some of these objections including all intents that have been brought up in the last 50 years are highlighted.

A. Argument of Disability

This contention of inability implies that "machines can never, perform job Y", Y could be determined as a set of soft skills; as instances of X, to be benevolent, have initiative, have a sense of humor, commit errors, fall in love, enjoy strawberries, learn from experience, or accomplish something extremely new.

B. The Mathematical Objection

According to Gödel's incompleteness theorem which related to Halting Problem and Un-decidability, philosophers have asserted that machines are intellectually mediocre compared to people. Machines are formal frameworks that are constrained by the incompleteness theorem; they can't build up the reality of their own, while people have no such impediment [19], [20]. Briefly, for any mathematical system F contains a set of axioms which is assumed to be true without having any formal proof, Godel sentence or *F(X)* could be represented with these features:

- *F(X)* is a sentence of X, with no prof using F.
- If X is consistent, then *F(X)* is true.

Gödel's incompleteness theorem is applied only to formal systems, and this includes Turing machines, but Turing machines are infinite, whereas computers are finite. This implies that; any computer can, therefore, be described as a (very large) system in propositional logic, which is not subject to Gödel's incompleteness Theorem.

C. The informality of behavior Objection

AI is subjected to what is called "The qualification problem" As it was claimed by philosopher Hubert Dreyfus that computers are unable to interpret everything as a set of logical rules [21]. Theoretically, human behavior, such as human expertise and knowledge is very difficult to be represented by a set of rules, and because computers just follow these incomplete rules, consequently, they cannot generate behavior as intelligent as that of humans. In [24], [22], [23], similar criticisms, regarding this objection, were also produced.

VII. Knowledge-based Systems Versus Cognitive-Based Model and Consciousness Exploration

During the sixties to the early days of the eighties of the twentieth century, the field was ruled by what was named "Symbolic Artificial Intelligence" (Symbolic AI), or "Rule-Based AI," that includes transferring the human behavior and explicit knowledge into a set of codded rules. This approach is very efficient for systems where the rules are very obvious, and input can be represented by symbols. Symbolic AI used symbols to define things (chair, cat, trucks, etc.) and can represent conceptual objects (transfer statements) or things that are not tangible. Fig. 4 outlines some of the algorithms used by Symbolic AI compared to that of ML.

Despite all this success in AI models; according to Yoshua Bengio, the key weakness is lacking methods for defining objects in a conceptual way [25], [26]. This obviously occurred when it is required to generalize beyond the training distribution. As a principle, if a task can be broken down into objects, any AI model will be able to learn it, however, there is no way to give each conceivable labeled example of the problem to the model [27]. This leads to needs for using cognitive and consciousness exploration-based models.

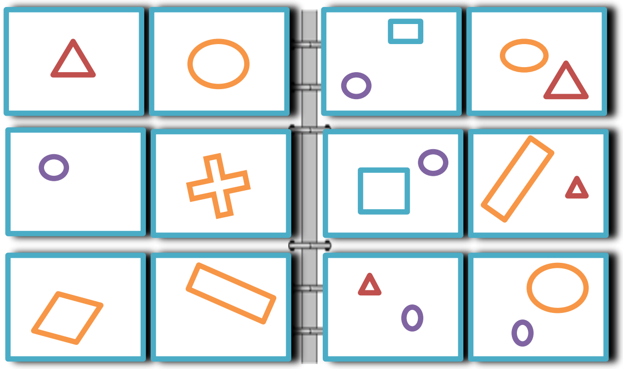


Fig. 4. Illustration of Bongard Problem.

1) Cognitive models: The Natural History Museum of Vienna has assaulted Facebook after the Facebook user was restricted from posting a photograph of a stripped ancient figurine of a lady which goes back to 29,500 years, Facebook replied, the ban was just an accident. Such failures demonstrate that there is no hope of accomplishing a complete intelligence system without first developing systems with what could be called deep understanding, which would involve an ability not only to correlate and recognize subtle patterns in complex data sets but also the capacity to look at any scenario and address unexpected situations. These limits become progressively clear in functional utilizations of the current AI. DL algorithms, for example, are data-driven, with no symbol or knowledge representation; consequently, it is difficult to be applied to systems that require reasoning and thinking [25]. Additionally, all DL models are prone to algorithmic bias because it gets its behavior from its training data. This implies that for any hidden or explicit biases embedded in the training examples will also find their way into the decisions the deep learning algorithm makes.

There is a need for the transferee to AI approaches that use cognitive models to overcome these limitations. Cognition is defined by psychological researchers as far as a sort of cycle; humans take in perceptual data from the surrounded environment, they assemble inner cognition models dependent on their view of that data and make their decisions accordingly. Psychological scientists perceive that such models might be imperfect, but they considered them to be the key to how humans see the world [28], [29]. However, what computational requirements needed to have systems that are capable of reasoning in a robust fashion must be studied.

2) Consciousness exploration: The Consciousness Prior Theory defined consciousness as "The perception of what passes in a man's own mind or awareness of an external object or something within oneself". It specifies that segments of our consciousness are chosen according to awareness methods and then communicate to the remainder of the brain, emphatically affecting downstream recognition [30]. After cognitive neuroscience, Yoshua Bengio turned his concentration to consciousness; he asserts that now is the ideal opportunity for ML to explore consciousness, which he says could bring "new priors to support abstraction and good speculation [31]. Yoshua aims that such research direction could permit AI systems to grow from representing what current systems are very good at, to represents more rational, sequential, logical, and intelligent models [32]. For his work, he only used those parts of consciences that include how humans express their felling in their own languages.

He used awareness as a mechanism of generating a set of related sequences for each event or thought; this sequence can be abstractly represented as an algorithm. In that way, consciousness can give motivation on how to build general models where agents are accomplishing something at a particular time at a specific place and have a specific impact [33]. That impact could have constant results all over the universe with the right abstractions.

VIII. Discussion (The Need for Hybrid Approach)

As illustrated in the former section, both cognitive and consciousness models are considered vital components for building a new robust AI system. Basically, "General knowledge" can be classified into two main categories; one includes all the ever known real-world factual knowledge that based on direct evidence, actual experience, or observation. The other reflects 'common sense', which is the sort of knowledge that humans assumed to be known intuitively without being told. For example, this simple fact "Once a baby born, he is alive" can't be inferred by any AI system. The main weaknesses in AI systems are that they don't get causation, they can see that a few occasions are related to different occasions, however, they don't find out which things legitimately cause different things to occur.

Fig. 5 illustrates the transition process of the AI, and its evolution in the last decades, features and challenges are maintained. The Rule-Based systems had deductive reasoning, logical inference, and a search algorithm that is used to finds a solution within the constraints of the specified model. It also used specified rules to deduce conclusions from the input data, to perform a certain goal. While in the Current AI, the rules of the model are not predefined, rather the data are provided and ML algorithms discover the rules from the training processes, and by applying statistical methods to adapt and tune different parameters till the optimal values are found.

Recently, influential steps towards building integral models that join features of the symbolic approaches with insights from ML, to obtain efficient techniques able to extract and generate abstract knowledge from stochastic data [34], [35]. For example, Geoffrey Hinton and others [36], use back-propagation algorithm to tackle the issue of enhancing the manner of adjusting synapses in order to enhance the performance. Backpropagation learns rapidly using synaptic updates and utilizes the connections of feedback to transfer error signals. So, a hybrid approach could be used to formalize the messiness of the problem in symbolic representation, then find all the correlations and induce some reasoning from it.

Central work of Neuro-Symbolic models is shown in [37] which analyzed the mappings between symbolic frameworks and neural systems, and indicated significant cutoff points on the sorts of information that could be represented in ANN, and showed the incentive in developing hybrid systems. Battaglia has produced a number of interesting papers on physical reasoning with systems that integrate symbolic graphs and deep learning [38]. A lot of similar work, such as [39] [40], [41] have been done to use ANN to give the answers from the messiness of the real world by learning. Then the symbolic part, forming internal symbolic representations, and create explainable rules to formalize the way that captures everyday knowledge, as shown in Fig. 5 [for clear resolution of the figure, refer to the last page].

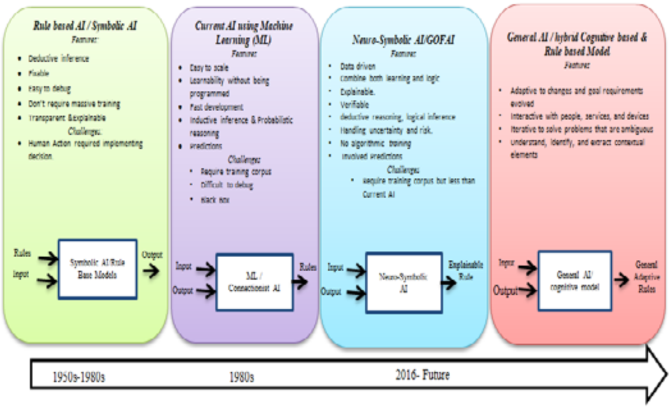

Rule based AI / Symbolic AI
Feature:
Deductive inference
Fixable
Easy to debug
Don't require massive training
Transparent &Explainable
Challenges:
Human Action required implementing decision
Rules
Input
Symbolic AI/Rule Base Models
Output
Current AI using Machine Learning (ML)
Feature:
Easy to scale
Learnability without being programmed
Fast development
Inductive inference & Probabilistic reasoning
Predictions
Challenges:
Require training corpus
Difficult to debug
Black box
Input
Output
ML / Connectionist AI
Rules
Neuro-Symbolic AI/GOFAI
Feature:
Data driven
combine both learning and logic
Explainable.
Verifiable
deductive reasoning, logical inference
Handling uncertainty and risk.
No algorithmic training
involved predictions
Challenges:
Require training corpus but less than Current AI
Input
Output
Neuro-Symbolic AI
Explainable Rule
General AI / hybrid Cognitive based & Rule based Model
Feature:
Adaptive to changes and goal requirements evolved
Interactive with people, services, and devices
Iterative to solve problems that are ambiguous
understand, identify, and extract contextual elements
Input
Output
General AI/ cognitive model
General Adaptive Rules
1950s-1980s
1980s
2016- Future

Fig. 5. The Transitions Affected the Evolution of AI.

In the history of AI, one of the largest efforts to create common-sense knowledge in a machine-interpretable form launched in 1984 by Doug Lenat, known as the CYC Project [42]. The main idea of the project was, to build a massive knowledge base containing static facts and heuristics, besides the cognitive and reasoning models needed to create what could be called common sense reasoning. According to Lenat; to simulate human thinking, CYC's team expected to code millions of facts crossing all different areas of human experience including science, society and culture, atmosphere and climate, cash and money, medicinal services, history, and other governmental issues. It was estimated that the CYC project requires a huge number (maybe thousands) of individuals to catch facts about brain science, governmental issues, financial aspects, science, and many, numerous different areas, all in logical structures. Simple declarative semantics models are used in knowledge representation, incorporating conjunctions, disjunctions, quantifiers, equality, and inequality operators. The CYC project has been depicted as "one of the most criticized projects of Artificial Intelligence". Machine learning researcher Pedro Domingos described the project as a "catastrophic failure" for several reasons, including the ceaseless amount of data required to produce any viable outcomes and the inability of evolving its own.

IX. Conclusion

A lot of the AI systems have become extremely powerful in many areas, such as medical diagnoses, translating languages, and image recognition, where they also can outperform humans at many complicated applications; however, they can be duped or confounded by situations they haven't seen before. Sometimes, the performance of AI systems, in their specialized domains, is very chaotic and weird, as none of them has a commonsense knowledge. This lack makes them brittle, its brittleness occurs when it is confronted by problems that were not foreseen by its designers. In this paper, we consider appealing to study how to integrate human experience and cognitive models with the current AI approaches in order to obtain more adaptive to the changes of the models. These models can interact with people, services, and devices and can understand, identify, and extract contextual elements.

As a future work, to enter the next decade of AI, more efforts must be done to build reliable AI systems that match basic reasoning of human, and can offer abstract solutions using insights, common sense and relatively little information. Apparently, in the next decade of AI, there is a need to redefine and refine the learning concepts, which are considered the main part of the AI models. Additionally, rich cognitive models must, intensively, be studied to represent models with rich- prior knowledge and sophisticated reasoning techniques.

References

[1] Gary F. Marcus, The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence, Computer Science, ArXiv, 2020.

[2] Tammet, Tanel. "Extending Automated Deduction for Commonsense Reasoning." ArXiv abs/2003.13159 (2020).

[3] Russell, Stuart J.; Norvig, Peter (2003), "Artificial intelligence: a modern approach" (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall, ISBN 0-13-790395-2.

[4] Golgi, Camillo. "On the Structure of the Grey Matter of the Brain." In Golgi Centennial Symposium: Perspectives in Neurobiology, ed. and transl. M. Santini, 647–50. New York: Raven, 1975.

[5] Abraham, T.H. Nicolas Rashevsky's Mathematical Biophysics. Journal of the History of Biology 37, 333–385 (2004).

[6] McCarthy, John, ``Programs with common sense", in Proceedings of the Teddington Conference on the Mechanization of Thought Processes, Her Majesty's Stationery Office, London, 1959.

[7] Moor, J., The Dartmouth College Artificial Intelligence Conference: The next fifty years, AI Magazine, Vol 27, No., 4, Pp. 87-9, 2006.

[8] Turing, Alan , "Computing machinery and intelligence", Mind, LIX (236): (October 1950), 433–460.

[9] Keil, F.C., "Nativism," in R.A. Wilson and F.C. Keil (eds.), "The MIT encyclopedia of the cognitive science" . Cambridge, MA: MIT Press, pp. 583–586. Keil , 1999.

[10] Simon, H. A.," Models of man", New York: Wiley, 1957.

[11] McCarthy, John ,``Programs with common sense", in Proceedings of the Teddington Conference on the Mechanization of Thought Processes, Her Majesty's Stationery Office, London. Reprinted in [McCarthy, 1990].

[12] Lenat, D. B., Prakash, M., & Shepherd, M. (1985). CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. AI magazine, 6(4), 65-65.

[13] Lenat, D. What AI can learn from Romeo & Juliet. Forbes, 2019.

[14] Lenat, D. B., Guha, R.V., "Building large knowledge based systems.", Addison Wesley, Reading, Massachusetts, 1990.

[15] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A. et al. (2017). Mastering the game of Go without human knowledge. Nature, 550(7676), 354-359.

[16] Superintelligence: Paths, Dangers, Strategies Reprint Edition, Kindle Edition.,2017.

[17] McCarthy, J., Minsky, M. L., Rochester, N., Shannon, C. E.: A Proposal, f or the Dartmouth Summer Research Project on Artificial Intelligence. The AI Magazine 27 (Winter 2006) 12-14.

[18] Bongard, Mikhail Moiseevitch. (1970). Pattern Recognition. Rochelle Park, N.J.: Hayden Book Co., Spartan Books.

[19] Kurt G¨odel. On Formally Undecidable Propositions of Principia Mathematica and Related Systems. Dover, 1962.

[20] Gödel, Nagel, Minds, and Machines, Solomon Feferman, The Journal of Philosophy, Vol. 106, No. 4, Special Issue: Our knowledge of nature and number: grounds and limits (Apr., 2009), pp. 201-219.

[21] Dreyfus, Stuart E.; Dreyfus, Hubert L. "A five-stage model of the mental activities involved in directed skill acquisitio, (February 1980).

[22] Stuart E. Dreyfus, "Coping with Change: Why people can and computers can't." logos (1986), 7:17-33.

[23] What Computers "Still" Can't Do: A Critique of artificial reason. Revised edition. Cambridge, Mass.: MIT Press, 1992.

[24] Harper & Row, " What computers can't do: a critique of artificial Reason", New York:, 1972.

[25] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives". IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), 35(8):1798–1828, 2013.

[26] Yoshua Bengio, Alexandra Luccioni, "On the morality of artificial intelligence" . IEEE Technol. Soc. Mag. 39(1): 16-25 (2020).

[27] Yoshua Bengio, Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, SarathChandar, "do neural dialog systems use the conversation history effectively? an empirical study. ACL (1) 2019: 32-37.

[28] Gallistel, C. R. "Learning, development, and conceptual change. The organization of learning". The MIT Press, 1990.

[29] C. R. Gallistel , Adam Philip King, Memory and the computational brain: why cognitive science will transform neuroscience, Wiley, ISBN:9781405122870, 2010.

[30] Baars, Bernard J. (2002) The conscious access hypothesis: Origins and recent evidence. Trends in Cognitive Sciences, 6 (1), 47-52.

[31] Yoshua Bengio, The Consciousness Prior, Université de Montréal, Mila, 2019.

[32] Yoshua Bengio. Learning deep architectures for AI. Now Publishers, 2009.

[33] Yoshua Bengio. Deep learning and cultural evolution. In Proceedings of the Companion Publication of the 2014 Annual Conference on Genetic and Evolutionary Computation, pages 1–2. ACM, 2014. URL http://dl.acm.org/citation.cfm?id=2598395.

[34] Dehaene and L. Naccache. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. Cognition, 79(1–2):1–37, 2001.

[35] Dehaene, H. Lau, and S. Kouider. What is consciousness, and could machines have it? Science, 358 (6362):486–492, 2017.

[36] Backpropagation and the brain, Timothy P. Lillicrap , Adam Santoro, Luke Marris, Colin J. Akerman and Geoffrey Hinton , 2020.

[37] The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision, Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, Jiajun Wu, arXiv.org , 2019.

[38] Cranmer, M. D., Xu, R., Battaglia, P., & Ho, S. (2019). Learning symbolic physics with graph networks. arXiv preprint arXiv:1909.05862.

[39] Raedt, L. D., Kersting, K., Natarajan, S., & Poole, D. "Statistical relational artificial intelligence: Logic, probability, and computation". Synthesis Lectures on Artificial Intelligence and Machine Learning, 10(2), 1-189 , 2016.

[40] Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T. et al. " Pyro: Deep universal probabilistic programming". The Journal of Machine Learning Research, 20(1), 973-978, 2019.

[41] Abraham, T.H. Nicolas Rashevsky's, "Mathematical Biophysics", Journal of the History of Biology 37, 333–385 (2004). https://doi.org/10.1023/B:HIST.0000038267.09413.0d.

[42] Lenat, D. B., Guha, R.V., "Building large knowledge based systems". Addison Wesley, Reading, Massachusetts, 1990.

Performance Assessment and Analysis of Blended Learning in IT Education: A Longitudinal Study in Saudi Electronic University

Mohamed Habib[1], Muhammad Ramzan[2]
Computer Science Department, College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia[1, 2]
Faculty of Engineering, Port Said University, Egypt[1]

Abstract—**Blended learning is a new educational model that binds traditional face-to-face learning with application of modern tools and technologies. This helps in retaining the positive features of the traditional learning while allowing students to realize potential of modern technologies. In blended learning, student perceptions and satisfaction plays a key role. Longitudinal studies can help identify patterns of these perceptions and expectations to evolve blended learning with changing times and technologies. In this paper, a longitudinal study has been carried out with the students and the faculty of Saudi Electronic University to identify major drivers and their role in shaping student perceptions and satisfaction. The results of this longitudinal study have been validated, and their subsequent comparisons are ascertained with the application of a decision tree based data mining technique. Based on the analysis and the findings of this study, the paper presents recommendations to improve blended learning experience and enhance the effectiveness of the teaching pedagogies developed consequently.**

Keywords—Blended learning; educational data; information technology; longitudinal study; data mining; decision tree

I. Introduction

Modern learning methods and various pedagogies to impart education have undergone a lot of enhancements during the last five decades [1]. Today, education is not confined to merely traditional classroom based learning. With the advancements in technology, new and versatile learning models have also evolved. Since the advent of the 21st century, integration of synchronous or asynchronous learning technologies have enabled educators to deliver education with innovative approaches using technology. Online learning found a great acceptance among academia for its flexibility and global reach to impart education where traditional infrastructure was hard to be provided. However, it is also a fact that there are certain unique aspects of traditional education, such as human connection, social interaction, spontaneity and personal attention; that cannot be substituted by any online learning approach. There has been always a need to find a learning environment that can combine positive aspects of both traditional and online learning approaches, while avoiding negative aspects associated with both. Blended learning was primarily introduced to achieve this goal.

Blended learning, as a concept, is a convergence of electronic-learning (e-learning) approach and face-to-face learning. It has been regarded as a new paradigm in modern education. The concept of blended learning emerged near the dawn of 21st century and soon found great acceptance amongst all levels of education. Today blended learning is being used at both the elementary and the higher education levels. The core feature of blending learning is its ability to incorporate technology while retaining features of face–to-face education. Today, blended learning is being adopted and studied as effective means of learning in all parts of the world.

For a successful blended learning model, achieving positive student perception plays a crucial role. The students are probably the most critical stakeholders in any academic system, and thus, their level of satisfaction directly reflects on quality of the education model. It is imperative in blended learning to know the level of adoption of technology among students and its measure of achieving students' learning goals. Being in a continuously changing environment of technology, practitioners of blended learning need to be constantly aware of student's perceptions, expectations and demands from their learning facilities, whether be traditional face-to-face or online. This interesting combination of traditional as well as technology driven education model therefore succeeds in captivating students' attention. Therefore, the current research in blended learning points towards the need for constant and up-to-date data collection regarding students' perception, with blended learning model under practice. However, there is still very little work done in the form of concrete studies to identify the role of various technological factors in students' satisfaction and meeting their goals. There is a need for concrete studies which observe student perceptions over a period of time to model their behavior and map it with significant features of blended learning.

Longitudinal studies are a very useful research method that involve continuous incremental data collection and its observation to find underlying models and patterns for any scientific phenomenon. These studies have proven to be effectively useful in many scientific domains such as environmental studies, biological studies and social sciences. Blended learning, as an area of research, can benefit greatly by application of longitudinal studies methods on student data to explore its effectiveness. Since such studies span over considerable period of time, their effectiveness in incorporating various relevant factors overpasses other similar research methods. Longitudinal studies are also a very good means of gauging the effectiveness of blended learning by successfully identifying factors which may be overlooked otherwise. This can help in identifying changes required to better cope with the student expectations and demands. This can also help in evolving learning model with changing technology and other platforms.

Saudi Electronic University (SEU) was established in 2011 as a premier institution of blended learning in Kingdom of Saudi Arabia with the aim to encourage provision of uniform quality education across all the regions of the kingdom. For this purpose, university was tasked to use blended learning as its medium of instruction. During all this time, a valuable amount of data about the student learning, performance, satisfaction, quality drivers and several other benchmarks has been collected and archived. In this paper, we have used a segment of that data to perform a longitudinal study determining student perception of blended learning model adopted in SEU. The data sample consists of information gleaned from more than 478 surveys of 243 students and 16 teachers over a period of four consecutive semesters from 2016 to 2018. The purpose of this study is to identify major quality drivers in positive student perception with blended learning. This study helps in understanding evolution of perception during time under study, which can further assist in shaping the learning model according to these perceptions and aspirations.

The remaining of this paper is organized as follows. After this brief introduction, literature review is presented in section 2. After literature review the data collection process is presented in section 3, followed by discussion and analysis of results in section 4, and decision tree based results validation in section 5. In the end, conclusion and future recommendations are presented in sections 6 and 7 respectively.

II. Literature Review

No academic model can be deemed successful if it fails to meet student perceptions of quality of learning and effectiveness of learning model. Students are the most critical stakeholders in academic systems. Success or failure of a model depends greatly on perception and acceptance of students. Quality of learning for students has been an area of interest for researchers since 1970s, which has helped in refining our concepts about changing dynamics of learning [2]. In [3] the factors that influence quality of learning are proposed, such as approach to learning, course material and student perceptions as shown in Fig. 1.

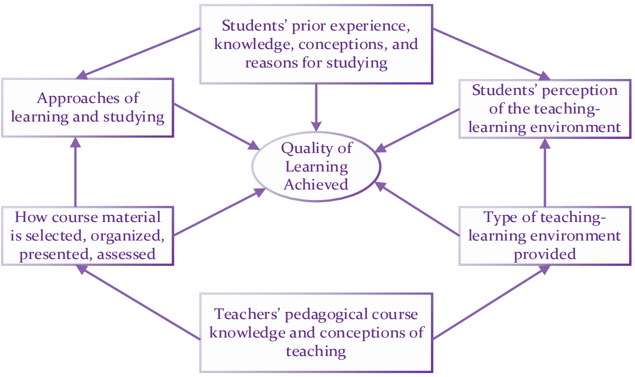

Students' prior experience, knowledge, conceptions, and reasons for studying
Approaches of learning and studying
Students' perception of the teaching-learning environment
Quality of Learning Achieved
How course material is selected, organized, presented, assessed
Type of teaching-learning environment provided
Teachers' pedagogical course knowledge and conceptions of teaching

Fig. 1. Concepts related to the quality of learning at university [3].

Blended learning as a medium of instruction has been around for more than two decades now. According to a study conducted in 2007, more than 45% undergraduate institutions in USA had adopted Blended learning as early as 2004. In 2001, assessment of effectiveness of blended learning had also begun. An evaluation on use of blended learning in a module at Master program at Cardiff University is presented by Banks [4]. A framework for blended learning is proposed in [5] which could be used to identify most suitable material for education at higher education level. In addition, the problem of creating effective number of assessments and its relation to blended learning environment has been introduced in [6] and [7], whereby, a data mining approach is implanted over educational data to predict the effect of the total number of assessment on student performance.

Pérez et al. [8] showed that applying blended learning can effectively reduce student attrition rate and increase their grasp of concepts resulting in better academic grades. Blended learning can be viewed as a combination of traditional face-to-face learning with e-learning [9]. Using distributed learning as a medium of education allows efficient interaction between faculty and students across different locations while retaining features of traditional face-to-face learning. In addition, it maintains a physical contact essential for effective and immediate guidance. There are definitions of blended learning that focus on percentage of time allocated to both face–to-face as well as distributed/online learning. For example, in [10] Bernard et al. proposed equal proportion (50%) contribution to both face to face and distributed learning. In [11], Yen and Lee emphasize that “blended learning, thoughtfully combining the best elements of online and face-to-face education, is likely to emerge as the predominant teaching model of the future”. Blended learning provides a personalized and adaptive learning approach to students that can be easily customized to suit the unique need of different students based on their unique characteristics and learning styles [12]. The blended course design involves thoughtful integration of various course delivery methods, learning principles, and instructional technologies which can provide the learners with a flexible, autonomic, and situated learning environment. Thus, blended learning is defined in [13] to be on way to becoming the new norm in higher education learning environments. A significant difference in success rate of students between blended learning environment and traditional environment, with higher success rate in blended learning, is presented in [14]. The authors attribute this higher success of blended learning to its ability in integrating face-to-face teaching that features the presence of an instructor, and e-learning with flexibility and accessibility in learning process. This course delivery method gives students opportunities to share and control learning, and to adapt to different learning context and situations.

Driscoll [15] proposed that blended learning could be explained as a combination of four approaches, namely, (i) Application of online and Information technology to achieve academic goals, (ii) Use of pedagogical approaches to improve learning outcomes, (iii) Amalgamation of information technology with traditional learning, and (iv) A mix of instructional methodology with actual job tasks. This study is very significant since it establishes the core competencies and scope of any successful blended learning model.

Student engagement is considered a fundamental and critical aspect of blended learning. It plays a crucial role in success of teaching model by contributing significantly in factors such as grades, persistence, and college completion [16]. By applying various modes of instructions (face to face as well as distributed), student's motivation to engage in self learning increases [17]. In [18], Dringus and Seagull consider student engagement to be the most critical factor for penetration of blended learning. Student engagement can involve many and often diverse factors, ranging from effort and persistence to learning motivation and involvement [19]. Self-report methods have been used effectively in domain of blended learning to measure student engagement. The data collected using this approach has been used to evaluate the blended learning model and study the relationship between student engagement and other important academic outcomes [20].

Longitudinal studies that intend to measure student engagement by capturing immediate student experience throughout blended learning, have been advocated as a means of effective quality enhancement mechanism in recent times [21], [22]. Longitudinal studies involve multiple measurements over time to model effectiveness of a process and to identify changes over the course of time [23]. These studies not only help in collecting specific data linked to activities that motivate student's performance in sphere of blended learning, but also provide researchers with ability to link these activities with driving factors and their relative influence. Another advantage of applying longitudinal studies is to have more transparency in data because of its multi-interval nature, since it is collected regularly over a period of time.

III. Data for Longitudinal Study

The student data collected for experimentation in this study is derived from Saudi Electronic University (SEU) systems. Various systems are in operation in SEU for academic purposes including Blackboard (Learning Management System (LMS)), Banner (registration system), and Attendance System etc. Details of data collected for analysis and experimentation can be summarized as following.

The dataset used consist of two parts. The first part is collected from surveys that were done on 243 different students having a total of 478 surveys and 16 faculty members over four consequent semesters of undergraduate students in IT program from 2016 to 2018. The second part of the dataset comprises of students' extracted data from the registration system (i.e. the Banner), along with students' data from Blackboard LMS.

In this study, three different experiments are conducted in order to measure multiple factors. The first part concentrates on students' perspective, measuring the following factors:

1) Effectiveness of Learning programming language in blended learning environment.

2) Effectiveness of applying practical labs for students versus having the same course without labs (Last two semester labs have been applied for programming courses).

3) Effectiveness of updating and modifying courses content.

4) Effectiveness of the total number of assignments per course.

5) Effectiveness of online quizzes

6) Effectiveness of participating on forums.

7) Effectiveness and easiness of using IT tools in Blended Learning System.

8) Rate of Satisfaction with IT Systems used for Blended Learning.

9) The second part of the study concentrates on faculty members and their satisfaction level for providing programming courses in blended learning environments

*10)*Faculty members' satisfactions with student performance in programming courses.

*11)*Faculty members' impression of applying practical labs.

*12)*Faculty members' satisfactions on updating and modifying courses contents.

*13)*Faculty members' satisfactions on predesigned assignments.

*14)*Faculty members' satisfactions on online quizzes

*15)*Faculty members' satisfactions on forums feedback to students.

*16)*Effectiveness and Easiness of using IT tools in blended learning system.

*17)*Rate of satisfaction with IT systems used for blended learning.

The third part of the study implements a decision tree on students' data to ensure the confidence of the analysis and results.

IV. Results and Analysis

In the following subsections, we present some of the salient findings of the study from all the three perspectives that we contemplated.

A. Student Feedback and Assessment

In this part of the study, we choose four courses in programming field, namely, (i) Computer Programming, (ii) Advanced Computer Programming, (iii) Web Technologies and (iv) Mobile Application Development. Major reason to select these course courses for experimental purposes is their applied nature that would require intensive collaboration between faculty and students on one hand, and offer an excellent opportunity to demonstrate effectiveness of IT systems in learning on the other hand. Subsequently, the survey was conducted after execution of each of these courses during the period of data collection. These four selected courses were then presented to the students in a blended learning environment through LMS comprising the academic and assessment resources.

Students answered an anonymous survey, which consisted of 32 questions in an online form after completion of each course. The total number of collected surveys were 478. Table I shows the total number of surveys collected every semester in different courses. Courses are represented as C1 => Computer Programming, C2 => Advanced Computer

Programming, C3 => Web Technologies and C4 => Mobile Application Development.

The students selected to participate in survey comprised of various cross-sections of participants. Student were from different age groups, as diverse as from 18 to 39 years. In addition, they belonged to different academic backgrounds such as high school, bachelors or diplomas in different areas of specializations. All these parameters were taken into consideration in building the students' performance prediction model. These surveys were conducted for five consecutive semesters in order to make sure that as many environmental factors as possible could be taken into consideration that could affect the outcome of the study. The resultant was a longitudinal study which test the following parameters:

1) P1: Effectiveness of learning programming language in blended learning environment.

2) P2: Effectiveness of applying practical labs for students versus having the same course without labs (Last two semester labs have been applied for programming courses)

3) P3: Effectiveness of updating and modifying courses content.

4) P4: Effectiveness of the total number of assignments per course.

5) P5: Effectiveness of online quizzes.

6) P6: Effectiveness of participating on forums.

7) P7: Effectiveness and easiness of using IT tools in blended learning system.

8) P8: Rate of satisfaction with IT systems used for blended learning.

These eight parameters are being used almost universally to measure the effectiveness of academic resources, assessment and application of knowledge. As seen in Fig. 2, for the first parameter we have more than 79% of students with great acceptance for programming courses in the blended learning environment. In addition, 84% of the students have approved the practical labs.

For the effectiveness of content update, we have around 73% of students. This shows that no real update on the content has been done and they feel no effectiveness of applied changes on the content. This depicts an effective need for constant review and upgrade of course contents in a blended learning environment. This is one of the significantly worrisome areas of blended learning, which shows that continuous administrative oversight is necessary to update the academic resources with changing environment.

TABLE I. NUMBER OF STUDENTS SURVEY OVER FIVE SEMESTERS

Semester	Number of Survey in each course				Total
	C1	C2	C3	C4	
First semester 2016/2017	29	26	20	15	90
Second semester 2016/2017	24	25	23	12	84
First semester 2017/2018	35	21	21	22	99
Second semester 2017/2018	27	32	19	19	97
First semester 2018/2019	37	25	30	16	108

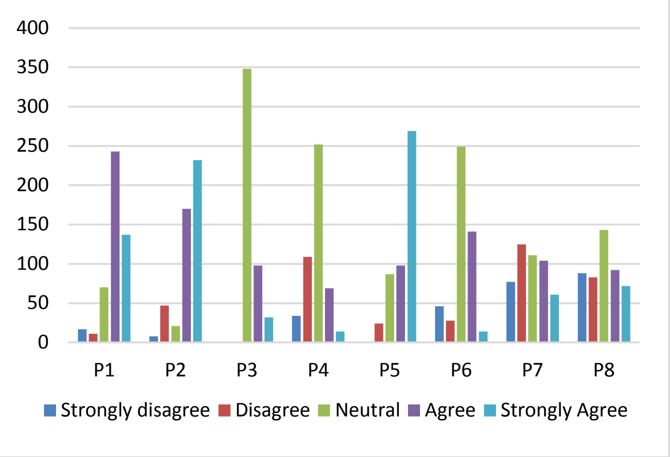


Fig. 2. Students Responses to different Parameters.

For the acceptance of a large number of applied assignments in blended learning environment, it has been shown from the results that around 83% of the students consider the number of assessments (assignments, quizzes, discussions and projects) to be excessive and are in favor of reducing these. Of all assessments results, it is evident that there is a good acceptance of online Quizzes, whereby, around 74% of students approve that. On the other hand, forums participation shows average acceptance from students for participation.

1) Analysis of student response: As is evident from Table II, more than 75% of the students have shown satisfaction with the first three parameters P1, P2 and P5. This indicates the higher acceptance from the students for programing courses, practical labs and having online quizzes in the blended learning environment.

On the other hand, for parameters P3, P4 and P6, most of students have natural response highlighting the need for updates and changes to course contents, number of assignments and forums, but not having high effectiveness on students.

Major areas of concern can be observed for parameters P7 and P8 where 30% have expressed facing difficulties in using IT tools as well as dissatisfaction with Blended Learning environment. Interestingly this is at par (30%) with students expressing greater acceptance and satisfaction of using IT systems in blended learning. One more observation during survey was about demographic variance and acceptance of blended learning. It is found that most of the students who faced difficulties are between 33 and 39 years old. Meanwhile, students who accepted blended learning tools are mostly under 30 years old. This shows that demographics can play a significant role in the success or failure of blended learning systems and should be considered as a significant parameter.

Fig. 3 show the results for the six parameters over the five semesters. The first figure (P1) indicates that student level of acceptance for blended learning environment is increasing over semesters, where the total number of students who agree and strongly agree is increasing in last three semesters.

TABLE II. PERCENTAGE OF STUDENTS RESPONSES ON DIFFERENT PARAMETERS

	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
P1	3.56%	2.3%	14.64%	50.84%	28.66%
P2	1.67%	9.83%	4.39%	35.57%	48.54%
P3	0%	0%	72.8%	20.5%	6.7%
P4	7.11%	22.8%	52.72%	14.44%	2.93%
P5	0%	5.02%	18.2%	20.5%	56.28%
P6	9.62%	5.86%	52.09%	29.5%	2.93%
P7	16.11	26.15%	23.22%	21.76%	12.76%
P8	18.41%	17.36%	29.92%	19.25%	15.06%

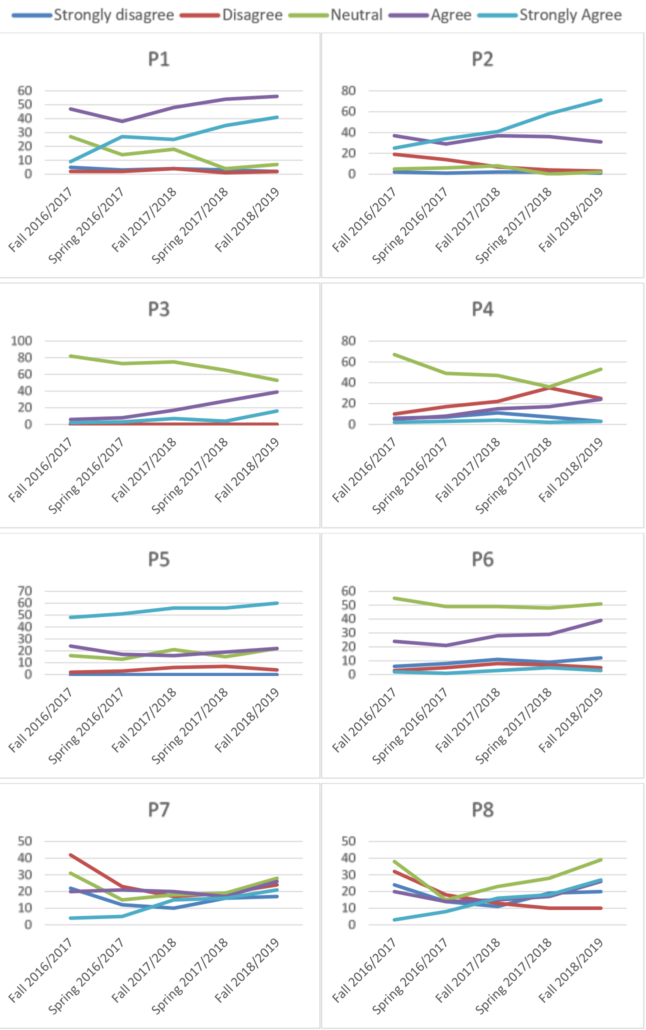


Fig. 3. Students Responses of Five Semesters.

For the second parameter, which measures the effectiveness of practical labs, at the first two semesters there appears to be a general resistance in applying the labs (P2). However, over a period of time, students are more eager to attend labs.

The third, fourth and sixth figures (P3, P4, and P5) show the effectiveness of course content updates, number of assignments and participation in forums, respectively. These patterns show that most of the students do not agree on these parameters, while the level of effectiveness is increasing over semesters with very low rate. In addition, the fifth figure, that measures the effectiveness of using online quizzes, shows a great impact and acceptance of students and it increases over semesters.

A very interesting observation is made about last two parameters. The seventh and eighth parameters start with high disagreement from students in initial levels of their education. However, it shows that over the semesters, students get more involved in blended learning environment and using different IT systems applied in this environment. In the last two semesters, the number of disagreed students is reduced, while more students found it easy to use IT tools and more satisfied with blended learning environment. This shows that student perception of blended learning changes as their expertise in working with systems involved improves. It also shows that in order to make blended learning universally effective, novice students need to be presented with more opportunities to interact with systems on experimental basis.

B. Faculty Feedback and Assessment

In this part, the survey were gathered over five semesters from teaching staff and faculty. The total number of staff members participating in the survey were 16, with 102 surveys. The faculty feedback about effectiveness and perception of blended learning systems can naturally vary to a great degree from students due to their expertise and knowledge. However, it can provide a unique opportunity to identify converging and conflicting factors between students and faculty.

1) P1: Faculty members satisfactions with student performance in programming courses.

2) P2: Faculty members impression of applying practical labs.

3) P3: Faculty members satisfactions on updating and modifying courses contents.

4) P4: Faculty members satisfactions on predesigned assignments.

5) P5: Faculty members satisfactions on online quizzes

6) P6: Faculty members satisfactions on forums feedback to students.

7) P7: Effectiveness and Easiness of using IT tools in Blended Learning System.

8) P8: Rate of Satisfaction with IT Systems used for Blended Learning. As is evident from

Faculty members' results shows a great acceptance of students' performance, applying practical labs, modifying courses contents, quizzes and forums. While about 84% of the faculty members are not satisfied with the currently designed assignments and advise to redesign new assignments for students. This result is very useful and reflect the same disagreement from the students against the current assignments. Table III shows the number of faculty and staff surveyed over a period of five semesters, whereas, Table IV and Fig. 4 presents the percentage of the staff responses on different parameters.

TABLE III. NUMBER OF STAFF SURVEY OVER FIVE SEMESTERS

Semester	Number of Staff Survey in each course				Total
	C1	C2	C3	C4	
First semester 2016/2017	5	4	3	3	15
Second semester 2016/2017	5	5	4	3	17
First semester 2017/2018	7	5	5	4	21
Second semester 2017/2018	6	7	5	5	23
First semester 2018/2019	9	6	6	5	26

TABLE IV. PERCENTAGE OF STAFF RESPONSES ON DIFFERENT PARAMETERS

	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
P1	0%	0.98%	16.67%	66.67%	15.68%
P2	0%	0%	0.98%	28.43%	70.59%
P3	0%	0%	15.69%	16.66%	67.65%
P4	3.92%	46.08%	34.31%	12.75%	2.94%
P5	0%	0%	0%	29.41%	70.59%
P6	4.9%	17.65%	42.16%	19.61%	15.68%
P7	3.93%	6.86%	21.57%	59.8%	7.84%
P8	4.9%	11.76%	50.98%	19.61%	12.74%

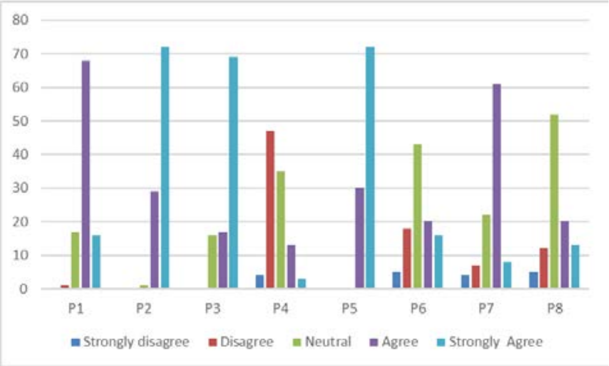


Fig. 4. Staff Responses to different Parameters.

1) Analysis for faculty response: Table IV shows feedback from faculty to various parameters being assessed during this study. It shows that most staff members have high confidence on parameters P1, P2, P3, P5 and P7. Faculty has shown a need to have more meaningful practical component in combination with conventional blended learning based teaching for programming courses. It was also observed that faculty showed greater confidence in change rate of course contents and online quizzes as compared to students. This can be attributed to active participation from faculty in actual revision and update process. In addition, the IT tools offered in blended learning were found to be very effective and easy to use by faculty members.

The biggest area in concern was discovered about predesigned assignments and assessment tools. Faculty showed a greater need to redesign and prepare custom assignments. For parameters P6 and P8, staff members have neutral response for the number of forums applied in each course. In addition, staff members show average satisfaction of current IT systems used, and provided suggestions for improving these systems.

C. Comparative Analysis of Student and Faculty Response

One important aspect of our study was to find areas of convergence and divergence between assessments made by faculty and students. The combined assessment is shown in Fig. 5.

Fig. 5 presents a comprehensive and interesting assessment where it can be observed that there is a great similarity between perceptions of both stakeholders for parameters P1, P2, P4, P5 and P6. The results show that high similarity is attained between the disagreement and agreement percentage for both the staff and the students for these parameters.

However, P3 is one parameter that indicates that staff members highly agree with the current changes and updates every semester on the course contents while students not agree on that. In addition, P7 shows the similar results which indicate that staff members found it easy and very effective to make use of different IT tools in blended learning environment. On the contrary, students responses do not depict that they agree on this. The last parameter shows the level of satisfaction to IT systems applied in blended learning environment. IT refers staff members have more neutral responses, while students show higher dissatisfaction, and suggest improvement to the currently applied IT systems.

These results show that even after the application of blended learning systems for a long time now, two major stakeholders in any academic system can differ greatly in their perception about effectiveness of the model. Our proposed decision tree based system, as described and explained in subsequent sections, is an attempt to present a uniform set of conditions that can help in ensuring uniformity of perception amongst both the students and the faculty alike.

V. DECISION TREE BASED RESULTS VALIDATION

In order to prove the results from the first two parts and ensure the confidence of the analysis, students data is extracted from the registration system (i.e. Banner), along with students data from the LMS. The extracted dataset represents all students participated in the survey enrolled in Information Technology program over five semesters. Data contains seven attributes as show in Table V.

In order to remove any errors and clean the data, a preprocessing phase is incorporated into the proposed scheme. This also removes any errors pertaining to the entry of the data in addition to the irrelevant attributes. The main objective of this approach is to find a relation between students' course GPA and the total number of assignments, quizzes and forums participation. Therefore, J48 classifier is used on the extracted and prepared dataset to provide a decision tree. The dataset, additionally, goes through a second step of preprocessing to convert the data into suitable format for the decision tree algorithm. Table VI shows the discretization rules applied on the dataset.

TABLE V. DATASET ATTRIBUTES

Attribute	Description
Age	Student age when he register the course
Gender	Male / Female
Previous Certificate	Student certificate before college (High school, diploma or other bachelor)
Assignments	Total number of assignments submitted by student during course
Quizzes	Total number of quizzes solved by student during course
Forums	Total number of participation in forums submitted by student during course
Course GPA	GPA for student in the course

TABLE VI. THE DISCRETIZATION RULES APPLIED ON THE DATASET

Attribute	Discretization Criteria
Age	1- Class 1: Age < 24 years 2- Class 2: Age between 24 years and 38 years. 3- Class 3: Age > 38 years
Gender	Male / Female
Assignments	1- Class 1: Less the 2 assignments 2- Class 2: between 2 and 5 assignments. 3- Class 3: Greater than 5 assignments.
Quizzes	1- Class 1: Less the 2 quizzes. 2- Class 2: between 2 and 4 quizzes. 3- Class 3: Greater than 4 quizzes.
Forums	1- Class 1: Less the 2 forums participations 2- Class 2: between 2 and 5 forums participations. 3- Class 3: Greater than 5 forums participations.
Course GPA	1- Class 1: Less than 2.0 (Reflect critical student) 2- Class 2: Greater than 2.0 (Reflect Stable student)

The data for the decision tress is divided into 66.6% for the training and 33.3% for the testing. The experimental results demonstrate that 78.6 % instances are correctly classified, while the incorrectly classified instances are 21.2%. Fig. 6 demonstrates the evolved decision tree structure. The results show that 92.2% of students, who solved more than five assignments during each course, got a higher GPA (over 2.0) and classified in stable stats during the IT program. While 88.8% of students who have less than 2 assignments, 2 quizzes and 2-forums participation are in critical status, with a GPA of less than 2.0.

This tree depicts very helpful results for the decision makers and for the academic advisors that could guide students during their enrollment in the IT program. In addition, we found that the age and gender are not critical attributes for students’ status during the program.

Table VII shows the following performance measures for each of the predicted classes:

- The True Positive (TP) rate: a measure for ratio where the model correctly predicts the positive class.
- The False Positive (FP) rate: measure for ratio where the model incorrectly predicts the positive class.
- Precision: a measure of how precise/accurate your model is out of those predicted positive, how many of them are actual positive.

$$Precision = \frac{TP}{TP + FP}$$

- Recall: calculates how many of the Actual Positives our model capture through labeling it as Positive.

$$Recall = \frac{TP}{TP + FN}$$

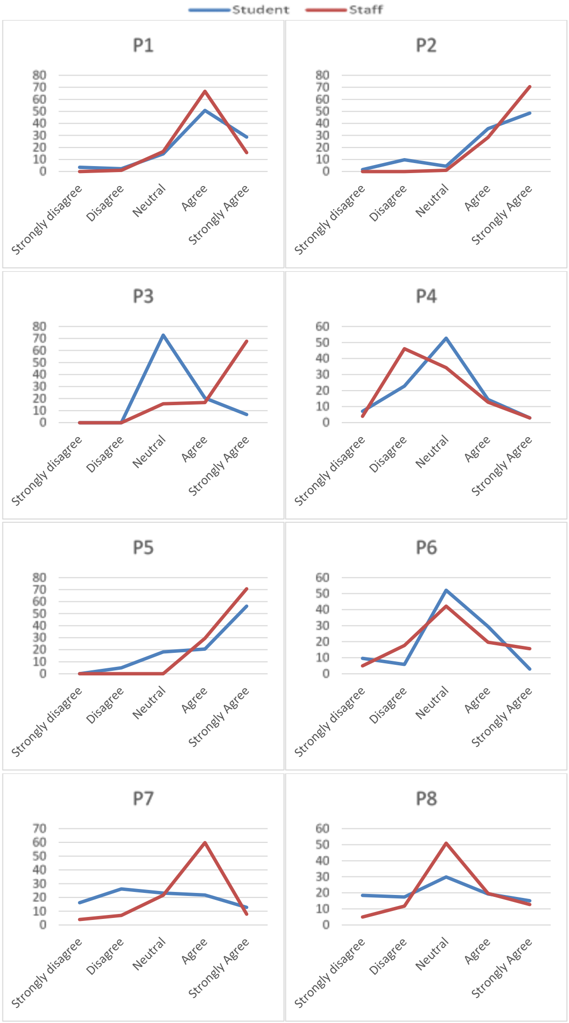


Fig. 5. Percentage of Students versus Staff Responses.

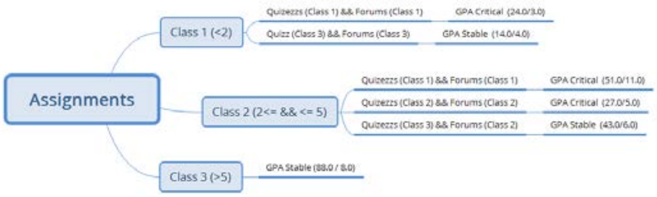


Fig. 6. Discovered Decision Tree for students’ performance.

TABLE VII. TP AND FP RATES FOR THE EXTRACTED CLASSES

Class	TP Rate	FP Rate	Precision	Recall
GPA>= 2.0 (Stable)	0.703	0.490	0.560	0.703
GPA < 2.0	0.510	0.297	0.658	0.510

In the second experiment, the results of applying practical labs and its influence on students' performance in programming courses is investigated. The same dataset, as was used in the previous experiment, is applied with some different attributes. These selected attributes are listed in Table VIII.

Initially, a preprocessing phase is applied on the dataset to remove the null values. J48 classifier is then used on the extracted and prepared dataset to provide a decision tree. The same discretization process, as presented in Table VI, is applied with the addition of a new rule for the new attribute Lab Attendance as presented in Table IX.

The resultant decision tree (Fig. 7) imparts a very important indication for the effectiveness of practical labs. As is evident that 91% of the students are less than 24 years old and attended less than two labs are in critical status. While, an average of 87% of male and female students are between 24 years and 38 years and attended more than 2 labs, have GPA over 2.0. In addition, for students over 38 years old, there is no direct impact from practical labs. Table X presents the discretization rule for the above mentioned investigation.

TABLE VIII. DATASET ATTRIBUTES

Attribute	Description
Age	Student age when he register the course
Gender	Male / Female
Previous Certificate	Student certificate before college (High school, diploma or other bachelor)
Lab Attendance	Total number of attended Labs
Course GPA	GPA for student in the course

TABLE IX. DISCRETIZATION RULE

Attribute	Discretization Criteria
Lab Attendance	1- Class 1: Less the 2 Labs attended 2- Class 2: between 2 and 5 Labs attended. 3- Class 3: Greater than 5 Labs attended.

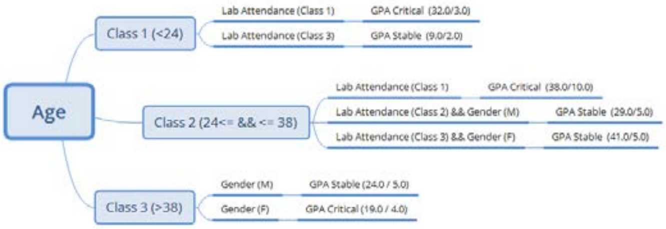


Fig. 7. Discovered Decision Tree for Students Performance Related to Practical Labs

TABLE X. DISCRETIZATION RULE

Class	TP Rate	FP Rate	Precision	Recall
GPA >= 2.0 (Stable)	0. 685	0. 479	0. 560	0. 685
GPA < 2.0	0. 521	0. 315	0. 651	0. 521

VI. Recommendations and Guidelines

This section introduce and summarize some suggestions for improving the effectiveness of blended learning environment based on students and staff members' responses which are summarized as follows:

Students have shown enhanced grasp of subject matter when subjected to practical work. Increasing practical labs for programming courses is preferable, as it would increase the success rate in these courses.

The need for formal course review was highlighted through both the student and the faculty responses. Course contents should be updated every semester and staff members should optimize application of various IT tools in blended learning for assessments and evaluation.

The number of home based assessments need to be revisited. The overall number of assessments in each course should be revisited and redesigned if necessary.

Online forums must be restructured in order to attract students and make it more effective during the learning process.

University should introduce training courses for both students and faculty on how to make best use of different IT systems applied in the Blended Learning environment.

The IT systems can be enhanced to be more user friendly and displaying course contents in attractive way to students, such as to increase the overall experience of blended learning.

VII. Conclusion

Blended learning combines good practices of traditional face to face learning and online learning. Blended learning relies heavily on identifying and meeting student perceptions for its success and evolution. It is very important to collect and analyze student data to meet this objective. In this paper a longitudinal study has been carried out on student data from Saudi Electronic University. This study has later been validated by comparing results with a decision tree based model of student as well as faculty responses. The results of this study show that both students and faculty have shown greater confidence in application of blended learning for education processes. However, the study also shows greater need for constant evolution and improvement in course content and assessment tools. In future, the authors intend to explore advanced heuristics based classifiers, such as Genetic Programming and Deep Learning, in order to exploit hidden dependencies in the solution domain which are often overlooked by traditional classifiers.

References

[1] J. Heywood, Assessment in higher education: Student learning, teaching, programmes and institutions. 2000.

[2] F. Marton And R. Säljö, "On Qualitative Differences In Learning: I-Outcome And Process*," Br. J. Educ. Psychol., 1976, Doi: 10.1111/J.2044-8279.1976.Tb02980.X.

[3] J. Entwistle, N., McCune, V., & Hounsell, "Approaches to study and perceptions of university teaching–learning environments: Concepts, measures and preliminary findings. Edinburgh, UK: Enhancing Teaching-Learning Environments in Undergraduate Courses Project," University of

Edinburgh, Coventry University, and Durham University, 2002. [Online]. Available: http://www.etl.tla.ed.ac.uk/docs/ETLreport1.pdf.

[4] J. Banks, "From boring to 'Blackboarding': Building participation through VLE group work," 2001. [Online]. Available: http://cebe.cf.ac.uk/learning/casestudies/case_pdf/jbanks.pdf.

[5] J. Wall and V. Ahmed, "Lessons learned from a case study in deploying blended learning continuing professional development," Eng. Constr. Archit. Manag., 2008, doi: 10.1108/09699980810852691.

[6] M. Alsuwaiket, A. Blasi, and R. Al-Msie'deen, "Formulating Module Assessment for Improved Academic Performance Predictability in Higher Education," Eng. Technol. Appl. Sci. Res., 2019, doi: 10.5281/zenodo.3249180.

[7] M. A. Alsuwaiket, A. H. Blasi, and K. Altarawneh, "Refining Student Marks based on Enrolled Modules' Assessment Methods using Data Mining Techniques," Eng. Technol. Appl. Sci. Res., 2020.

[8] M. V. López-Pérez, M. C. Pérez-López, and L. Rodríguez-Ariza, "Blended learning in higher education: Students' perceptions and their relation to outcomes," Comput. Educ., 2011, doi: 10.1016/j.compedu.2010.10.023.

[9] N. A. Williams, W. Bland, and G. Christie, "Improving student achievement and satisfaction by adopting a blended learning approach to inorganic chemistry," Chem. Educ. Res. Pract., 2008, doi: 10.1039/b801290n.

[10] R. M. Bernard, E. Borokhovski, R. F. Schmid, R. M. Tamim, and P. C. Abrami, "A meta-analysis of blended learning and technology use in higher education: From the general to the applied," J. Comput. High. Educ., 2014, doi: 10.1007/s12528-013-9077-3.

[11] J. C. Yen and C. Y. Lee, "Exploring problem solving patterns and their impact on learning achievement in a blended learning environment," Comput. Educ., 2011, doi: 10.1016/j.compedu.2010.08.012.

[12] Z. A. S. Al-Khanjari, "Applying online learning in software engineering education," in Overcoming Challenges in Software Engineering Education: Delivering Non-Technical Knowledge and Skills, IGI Global, 2014, pp. 460–473.

[13] A. Norberg, C. D. Dziuban, and P. D. Moskal, "A time-based blended learning model," Horiz., 2011, doi: 10.1108/10748121111163913.

[14] A. A. Y. Al-Qahtani and S. E. Higgins, "Effects of traditional, blended and e-learning on students' achievement in higher education," J. Comput. Assist. Learn., 2013, doi: 10.1111/j.1365-2729.2012.00490.x.

[15] M. Driscoll, "Blended learning: Let's get beyond the hype," 2002. [Online]. Available: https://www-07.ibm.com/services/pdf/blended_learning.pdf.

[16] G. M. Sinatra, B. C. Heddy, and D. Lombardi, The Challenges of Defining and Measuring Student Engagement in Science. 2015.

[17] K. A. Meyer, "Student Engagement in Online Learning: What Works and Why," ASHE High. Educ. Rep., 2014, doi: 10.1002/aehe.20018.

[18] L. P. Dringus and A. B. Seagull, "A five-year study of sustaining blended learning initiatives to enhance academic engagement in computer and information sciences campus courses," in Blended Learning: Research Perspectives, Volume 2, 2013.

[19] S. L. Reschly, A. L., & Christenson, "Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. Handbook of research on student engagement," 2012. [Online]. Available: http://www.springerlink.com/index/X0P2580W60017J14.pdf.

[20] R. Junco, G. Heiberger, and E. Loken, "The effect of Twitter on college student engagement and grades," J. Comput. Assist. Learn., 2011, doi: 10.1111/j.1365-2729.2010.00387.x.

[21] M.-T. Eccles, J., & Wang, "Part 1 Commentary: So what is student engagement anyway? In Editor 1 & Editor 2 (Eds.)," in Handbook of research on student engagement, New York, NY: Publisher., 2012, pp. 133–145.

[22] M. A. Lawson and H. A. Lawson, "New Conceptual Frameworks for Student Engagement Research, Policy, and Practice," Rev. Educ. Res., 2013, doi: 10.3102/0034654313480891.

[23] M. Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, Experience sampling method: Measuring the quality of everyday life. Thousand Oaks, CA: Sage, 2007.

Usability and Design Issues of Mobile Assisted Language Learning Application

Kashif Ishaq[1], Fadhilah Rosdi[2], Nor Azan Mat Zin[3]*, Adnan Abid[4]
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia[1, 2, 3]
School of Systems and Technology (SST), University of Management and Technology, Lahore, Pakistan[4]

Abstract—**This paper aims to look at teachers, government officials, and students for Literacy & Numeracy Drive (LND), a smartphone app for students in Punjab province, Pakistan, to teach languages and math. Furthermore, to recognize LND usability and design problems while its use for grade three in schools. As the usability and design issues of LND were not discussed since the launch of this application. The methodology for this study is the questionnaire for teachers and semi-structured interviews for government officials of District Sheikhupura and students. The result shows that LND has various usability and design problems in its current form, i.e., buttons, icons, color schemes, sluggish performance, and fonts. Besides, teachers, government officials, and students suggested that game-based learning consists of an interactive interface, phonics, key animations to be created and adopted. Highly engaging and appealing delivery of the curriculum and improvements in the appraisal will improve the participation of students and deliver better outcomes.**

Keywords—Educational technology; language learning; literacy and numeracy drive; mobile application (App); m-learning; usability; user interface design

I. Introduction

Mobile learning (m-learning) technology involves the usage of mobile devices for instructional applications. Digital technology enhances the learning and students' success by providing the privilege of being able to study anywhere at all times, depending on the comfort of students [1]. The instructions should not be done at a fixed place or within the specified period such as a classroom [2]. M-learning seeks to bring colleges and organizations at the center of educational innovation and to fulfill user expectations for versatility and ubiquity [3]. However, one of the challenges confronted by mobile application (app) developers is that mobile devices present new barriers to usability that are difficult to model using conventional usability approaches. Usability standards and guidelines for desktop software used for designing mobile apps are not applicable and do not deal with problems associated with existing mobile devices constraints [4].

Two critical factors for the actual implementation of any information system are usability and user experience. Usability is the degree to which a single person may allow the usage of a program, invention, or service for the achievement of the objectives with quality, reliability and productivity in the form of a particular application [5]. The user experience is the observations and reactions of someone resulting from the use of software, device, or service or its expected use [6]. Usability is a critical issue in mobile apps, which can be avoided from being challenging to use by an adequately designed software, which is one of the main determinants of numerous apps' performance.

Usability testing of apps for portable devices is a new field of research that is confronted with several challenges due to the distinctive feature of small screen devices, restricted input capacity, and the evolving user perspective [7]. The usability testing of mobile learning systems is a critical mechanism for the functionality of mobile apps to ensure that mobile learning is easy, practical, and secure [8]. The technical deficiency and high rates of consumer satisfaction of mobile devices are important. Nevertheless, the use of virtual learning and evaluation in schooling advances slowly, owing to difficulties, such as digital technology, curricula, job growth, organization and management [9].

In comparison, issues including low efficiency, screen-size, reduced bandwidth, poor display quality, storage space, availability of different file formats, lack of input features, usage of multiple modules, and operating systems for mobile devices are all at risk for the accessibility of M-learning applications. The indicators used for M-learning applications include learning power, effectiveness, modification, efficiency, coherence performance, quality, inaccuracy and satisfaction [10]. In compliance with the technology acceptance model (TAM) [11], if the approach is deemed user-friendly and practical, the consumer will adopt the knowledge structure as it is compatible with the intent to be used.

The techniques for developing functional mobile apps are not considered to be successful usability test methods. For this reason, the usability of mobile apps must be assessed through the development and implementation of appropriate usability testing methodologies. Nevertheless, to m-learning researchers and designers, this is a significant challenge along with usability and design issues. Therefore, this study aims to examine the usability and design issues of a mobile-assisted language learning application for public sector schools in Pakistan, named Literacy and Numeracy Drive (LND).

A. Research Background

1) Literacy and numeracy drive: Given the strengths of m-learning, the local government of Punjab has launched the LND, a mobile learning program for teaching and evaluation in public schools for grade three students (8-10 years old). The program was introduced in 2015 in 36 districts throughout the province of Punjab, where the conventional students'

*Corresponding Author

assessment procedures such as the Punjab Examination Commission (PEC) are pricey, infrequent, and complex. The initiative comprises 52,394 schools; 403,172 teachers and 12,268,981 students [12] [13] [14] [15] [16].

The Department of School Education and the Punjab Information & Technology Board, therefore, implemented a low-cost tablet-pc examination program for students to assess them during their monthly school visits by the Monitoring and Evaluation Assistant (MEA) (Fig. 1). The assessment software is linked to a massive question bank, with each question identifying the corresponding learning outcomes for students [14]. Currently, the Learning Outcomes (LOs) for English (Fig. 2), Urdu (Fig. 3) and Math’s (Fig. 4) are measured for grade three students. The areas assessed are comprehension (Fig. 5), sentence completion, two and three-digit addition and subtraction, multiplication (Fig. 6) and division.

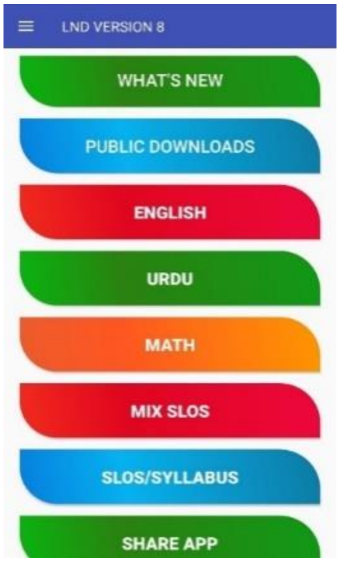


Fig. 1. LND Interface

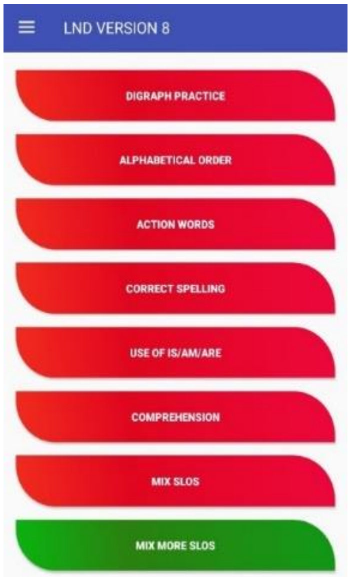


Fig. 2. Interface for English.

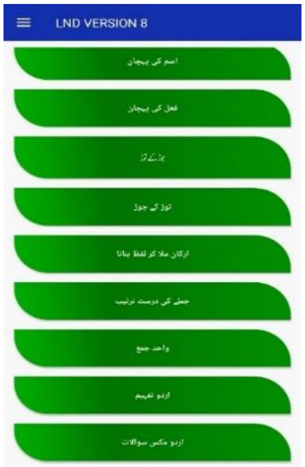


Fig. 3. Interface for Urdu.

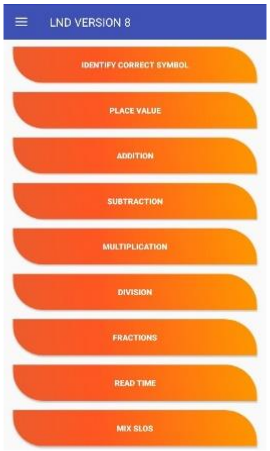
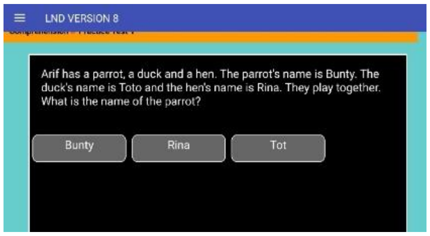
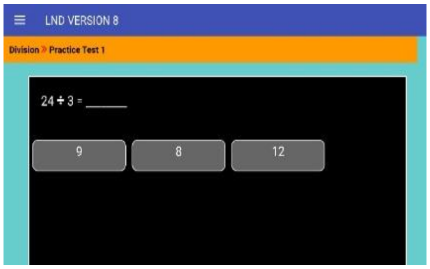


Fig. 4. Interface for Math.



Arif has a parrot, a duck and a hen. The parrot's name is Bunty. The duck's name is Toto and the hen's name is Rina. They play together. What is the name of the parrot?

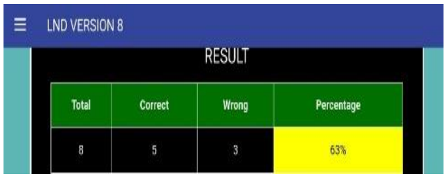
Fig. 5. LND English Test.



24 ÷ 3 = ______

Fig. 6. LND Math Test.

The MEA, who also has an LND mobile app on his tablet, evaluates the students. An evaluation tablet allows questions to be periodically rendered from a central question-bank for each student, subject, and key LO, and multiple-choice questions are presented to the students (Fig. 5). A total of seven questions were tested with a student sample in a school. There are more than 47,000 public schools in Punjab, and the assessment cycle is accomplished in less than five minutes per student. The monthly evaluation is carried out with almost 329,000 participants and to date, nearly 6.7 million assessments have been completed by the MEAs (Fig. 7). The statistics of each evaluation is shared with training managers using the internet portal and full-term here SMS notifications [13] [14].



RESULT

Total	Correct	Wrong	Percentage
8	5	3	63%

Fig. 7. LND Test Result.

B. *Motivation of this Research*

Although smartphone utilization is constantly popularized and increasing, users prefer to utilize mobile apps with advanced features for their study activities [17]. The LND implementation in government schools in Punjab Province is the focus of this research. This software has been downloaded and implemented in 52,394 [12] schools and used for teaching and evaluation, but no study has been done until now on its usability. Therefore, this study aims to determine the usability of the LND app. Evaluation of products and technologies is necessary [18]; thus, the usability evaluation for LND is carried out. Furthermore, since its implementation, there has been no usability evaluation carried out for the LND app. This evaluation is carried out among the stakeholders - teachers and administrators and students to gather feedback on the LND mobile app.

This paper is structured accordingly in which the related literature is summarized in Section 2. Section 3 presents the research method used. Section 4 summarized the results and discussion of the research, and finally, Section 5 concludes the paper with suggestions for improvements.

II. Literature Review

A. *Mobile Learning*

Mobile learning or m-learning is an extension of e-learning that enables consumers to learn with small and mobile wireless devices. Mobile apps are being created for online education in the educational world, which offers opportunities for students to learn whenever feasible, based on their convenience. Learning and teaching do not require a particular location or schedule but are versatile and can be performed anywhere at any time. M-learning has several distinct literature meanings, but all are restricted to learning from mobile devices and other hand-held technologies that are not time and location-based [19]. Mobile technology has continued to be used by academic organizations around the world because the relevance of using mobile technology to support teaching and learning is apparent and inevitable [20]. Smartphones have expanded instructional resources by reducing costs and increasing flexibility. To date, many efforts to introduce digital learning in institutions identify possible advantages and disadvantages of using these devices to allow access to learning such as (a) the potential of students to build self-centric pedagogy, (b) the ability to establish automated learning pedagogy, (c) the ability to promote useful communication tools for learning and support, and track the learners' knowledge [21] (d) the ability to provide flexibility in learning without the restriction of time and place, (e) to provide the ability to define the content according to the need either audio, video or presentation/images, and (f) to enhance the motivation for learning interactively [22].

2) Challenges for mobile learning: However, there are challenges for mobile learning students and teachers; m-learning (a) can separate technologically sound students from non-technically successful students, (b) can develop a sense of detachment among students and teachers, and (c) is hugely dependent on a networked resource [21]. Designers of mobile apps experience several challenges due to two leading causes, the usage context and device capacity:

a) Usage Context is the traditional usability methods apply to a standardized and well-established framework. In contrast, mobile learning environments are often volatile and hard to detect, forecast and simulate meaning. Users often employ different tools and use various application assessment methods.

b) Device capacity is a physical restriction of portable devices, particularly in narrow-screen size and resolution, which significantly impacts mobile apps' usability as it is visually uncomfortable to read human behavior and so from a small screen. Additionally, small icons and touchscreens minimize feedback and increase human mistakes because the usability performance depends mainly on the use of multiple-input and output processes. I think device capacity should include processing power too.

B. *Design and Issues of Mobile Educational Applications*

The user interface (UI) design concept for m-learning apps is one of the most significant mobile architecture dimensions comprising various functions such as ease of usage, customer retention, appeal, and learning abilities [23]. The design objective of a product is to increase consumer retention and engagement by the utility, ease of usage and pleasure in the experience. As it relies on enhancing consumers' awareness about what they are searching for, what they need and what they hear, effective design is thus critical for users to be significantly influenced by a product [24]. When designing the user interface, designers will take into account the form of users communicating with the program, the particular user requirements and device functionalities coupled with robustness, reduced error levels and excellent efficiency for a greater framework adoption. It will provide a well-built user experience to make sure the code is appropriately suited. Accuracy should be preserved across growing channels and apps. The Graphical User Interface (GUI), where user inputs are recognized by specific mobile device buttons or pointing methods that correspond to the screen, is the first form for the mobile user interface. The GUI presents important details on handheld app screens [10] [23].

Three user interface considerations need to be addressed when designing a mobile content; 1) Provide just a primary function. The mobile devices are tiny so, the specification should contain only the elements necessary for the task; the GUI may otherwise be frustrating for the users. 2) Have continuous association. Assignments should be built sequentially such that the consumer can only access a specific interface at a time. There is a significant distinction between a portable device and a computer user device. A function may be broken into a screening series if it needs many moves. 3) Limited on-screen assistance. Mobile users prefer to invest less time on the app than computer users. Therefore, GUI should be as simple as possible to complete the function while pursuing a transparent and rational flow. This element strongly refers to user experience, accuracy, minimum surprise concepts [25], and consistency [26]. Consistency is the key principle in design because usability and learnability enhance when similar components have consistent looks and functions in a similar way.

Older people usually face issues utilizing cell phones, including tiny keys and rubbery edges, limited text size displays, small keys, and letters. They are likely to hit the keyboard numbers inappropriately because they are not acquainted with slide or touch screen interface style plus dynamic menu configuration [27]. A major drawback is the screen size. Furthermore, small screen size causes visual problems, eye strain, or difficulty for visually impaired people.

Moreover, websites are not often optimized for smaller screens [28]. As a result, there are significant limitations on the size of the objects which can be viewed as well as the size of the font. A font size 12pt is probably too small and hard to read, even on a larger screen [29]. Besides, restricted space and memory and the ability for editing documents may also restrict mobile instructional practices [28]. The keys which are used must be wide enough and distinct from the rest of the interface component so that users can operate effectively. The common idea is that only graphical pictures can be used in the keys, and the text on them should also not be ambiguous and frustrating for children [29].

C. Usability and Issues of Mobile Educational Applications

Usability was introduced at the end of the 1980s [30] and is used regularly for the consistency and recognition of goods and services [31]. Nine standardized usability attributes have been established in current mobile usability studies, which are Learnability, Efficiency, Simplicity, Errors, Memorability, comprehensibility, user satisfaction, and learning output. Such usability attributes seek to measure the quality and usefulness of certain products [7]. Using these attributes, the present research aims to measure usability and identify design problems with the LND app currently used in schools [14]. Usability testing for mobile apps is an emerging field of research, as it has unique attributes, including touchscreen width, limited input area, and increasing user experiences, which is difficult to achieve [1] [7].

The small size of mobile devices can cause issues such as poor graphical appearance [32] and inadequate multimedia output [33]; poor display resolution [34], restricted input functions [35], keyboard limitations [36], restricted storage capacity [37], and low processing speed [38]. Other challenges that need to be addressed include difficulty reading on a small-screen [39], failure in assignment completion [7], speed of Internet connectivity, evaluations on the smartphone, and students handling gadgets as toys from outside school [40].

A study by [8] found the dissatisfaction shared by students at the National University of Fiji for mobile learning apps in which thirty students participated. The results indicated significant usability issues and further improvement recommendations. Similarly, Primo, a discovery tools' usability study, was also administered on a medium-sized research tool for the library that detects user search behavioral patterns. The researchers investigated essential design concepts and functionality to grasp Primo's accessibility for users and carried out predictive usability tests which revealed that users encounter several technology challenges [41].

The effectiveness of mobile game-based learning needs positive smartphone learner behaviors, user experience, usability, design, and useful system [42]. Several mobile apps have been designed for use in technologically improved smartphones. Still, usability and design in almost all of the applications are not the main focus, whereas usability and design are the primary cause for those apps not being usable [43]. The smartphone LND app has been used in 52,000 schools since 2015 [14] but the English assessment [13] indicates its ineffective use because of poor design, low performance, font size problem, not suitable content and not being used by teachers [14]. Nonetheless, to date, no research has been conducted on the usability of this mobile app. Identifying usability and design issues is important to enhance the user experience of the LND mobile application, to achieve learning outcomes.

III. Methodology

A. Setting and Participants

The population of this study are primary school teachers and students of the Punjab province in Pakistan. There are 403,172 teachers (male and female) and 12,268,981 pupils (male and female).

The samples were randomly selected from 21 schools, out of the 1,247 in the Sheikhupura district. Fifty-seven (57) teachers who use LND applications in the classroom for teaching languages and Math subjects participated in the study. Besides, two government officials, the District Monitoring Officer (DMO) and the Assistant Commissioner (AC) and a total of 300 students also participated in the study. The teacher samples were self-administered questionnaire survey; thus, the returned rate obtained was 100 percent. The DMO and AC together with student participants, were interviewed using a semi-structured interview schedule. Students were interviewed instead of the given questionnaire because of the difficulty of reading and understanding the questionnaire items.

B. Instrument

1) Questionnaire: The questionnaire consists of 25 items (5 demographic and 20 usability items) adapted from [14]. The measured alpha Cronbach is 0.845 which indicates a high degree of internal uniformity. Items in the questionnaire cover demographic information, user experience, usability items (ease of use (5 items), accessibility (2 items), User Experience (10 items)). Demographic information includes age, class, education and school location. All items besides demographic use a 5-point Likert scale (1-Strongly disagree to 5-Strongly agree).

2) Interview schedule: A semi-structured interview schedule in Urdu which is the national language of Pakistan, was used. The interviews were conducted in Urdu because the participants are public office holders and also they felt comfortable using their mother tongue. Government officials were selected because they have been actively involved in public sector school educational activities in the district. They concurred with the need for the study and acknowledged that no one had investigated the smartphone LND app to date.

The interview schedule consists of demographic information for officers (age, employment, job experience) and (age only) for students in the first section and LND user experience, ease of use, and design issues are in the second section. LND specific issues are the context of the color scheme, font size, layout and design. Open-ended questions on the LND features that need enhancement to make it more effective, engaging, and beneficial for students were also included along with recommendations. The interview was carried for one hour with the DMO and AC in their respective offices.

The same instrument was used for students in the semi-structured interview session, scheduled during the free class time and break time, where the researcher interviewed the students himself. This interview was also in Urdu because students are not able to speak English at the Primary level. The interview started with the demographic information of the participant, and then questions related to the LND application were asked. The questions about the usability and design issues were inquired along with suggestions for improvement.

C. Data Analysis

The Statistical Package for Social Science (SPSS 25.0) is used to evaluate teacher's questionnaire results, frequencies for demographic, Mean and Standard Deviation for the Usability and Design problem data. A qualitative method is used to analyze the interview results from students and the DMO and AC. Every interview was transcribed and then translated. The results were analyzed by identifying the answers from the interview, then coding and classifying them based on the theme. The results and analyses are translated into English, from the Urdu transcripts. Two language experts and corrections verified the translation of transcripts were made where required.

IV. Results and Discussions

There are 12 (21.1%) males and 45 (78.9%) females among the 57 teacher respondents. The age range of 18 (31.6%) respondents are between 26-30 years old, whereas 9 (15.8%) respondents are in the 31-35 age group while 13 (22.8%) and 17 (29.8%) respondents are in the 36-40 and above 40 age groups, respectively. As for students' participants,138 (46.0%) are males, and 162 (54.0%) are females. The age range of 6 participants (2.0 %) was 5-7 years old, 217 (72.3 percent) participants are from 8-10 years old, 76 (25.3 percent) are 11-13 years old, whereas only one (0.3 percent) were older than 13 years. There were only two government officials, one male over 40 years of age and one female over 35 years.

A. Survey Result

The results from the teachers' survey are presented in Table I.

1) Usability issues of LND: Table I indicates that all items except items 8 and 9 have low mean scores (between 1.02 to 2.44), which translates to low usability. For any products such as smartphone apps, usability measure is important because a better usability score means an application is being used smoothly, efficiently and with satisfaction. Items 1 to 7 represent the usability issues encountered with the LND app. Respondents reported that it is difficult to find the icons and that the interface does not make the app easy to use, as indicated by items 1 (̄mean=2.44) and 2 (̄mean=2.25). Also, the respondents disagreed that the touch screen provides an easy input method as indicated by a very low item 3 score (̄mean=1.56). The difficulty of using the app is confirmed by respondents (item 8 mean=3.78) Assistance is important to guide a user in an application but items 4 (̄mean=2.14) and 5 (̄mean =1.45), indicate that there is little or assistance provided by the app. Additionally, there is no instruction to fix errors (item 6 mean=1.15) and there is no effort to improve user experience by the app provider (item 7 mean=1.02).

TABLE I. Results of LND App Usability Test

Item No.	Items	N	M	SD
1	The application icon is easy to find.	57	2.44	1.09
2	The application interface is easy to use.	57	2.25	.61
3	The application provides easy to use touch screen input.	57	1.56	.50
4	The application provides step by step assistance to use it.	57	2.14	.72
5	The application provides assistance in difficulty.	57	1.45	.50
6	The application instructs to fix the problem automatically.	57	1.15	.37
7	The application provider is taking steps to improve the user experience of the application.	57	1.02	.13
8	The application is difficult to use.	57	3.78	.54
9	The performance of the application is slow.	57	4.75	.43
10	The application makes me skillful in learning English.	57	1.84	.65
11	The use of application makes me confident.	57	1.21	.41
12	It helps me to enhance my vocabulary.	57	1.54	.57

Note: Scale ranging from 1-Strongly Disagree to 5-Strongly Agree

Furthermore, the performance of the LND app is slow during class usage (Item 9, mean=4.75). Moreover, items 10 ($\overline{}$mean=1.84), 11($\overline{}$mean=1.21) and 12 ($\bar{X}$=1.54) show that respondents are not satisfied with the learning performance using the LND app.

Table II shows the results related to the design of the LND app.

2) Design issues of LND: In items 13 to 20 of Table II, the issues in the design of the LND are illustrated. The size of the font is a crucial element to consider for text reading on a mobile phone app, and it is challenging to use such an application if the font is not readable or difficult to read. Item 13 ($\overline{}$mean=1.94) indicates that the majority of respondents face difficulty in reading text from the app screen. Navigation keys help users use the app easily and efficiently but the LND app lack this feature (Item 14 mean=1.14). Respondents very much disagreed that the icons are attractive and recognizable (Item 15 mean=1.70), or that the color scheme of buttons and application screen is attractive (Item 16 mean=1.95, and item 17 mean=1.56). The application's efficiency also has an impact on its usability because the user will not be able to learn quickly when the application operates poorly while carrying out the task. Voice instruction, simulations and instructional videos motivate people to learn quickly and effectively besides giving a clear demonstration of concepts realistically. However, this app has none of these multimedia elements as indicated by the lowest scores for items 18 to 20 ($mean = 1.00$).

Table III presents the results for LND app accessibility.

3) Accessibility of LND: According to Table III for LND accessibility, item 21 ($\overline{mean} = 4.91$) shows that the app has a significant number of advertisements that create access and usage problems along with distraction from the learning process. Items 22 (mean=1.19) and 23 (mean=1.02) have the lowest mean scores, which indicate that the app does not provide self-recommendations for questions, and developers are not taking steps to improve the user experience of the application. Item 24 ($\overline{Xmean} = 1.63$) indicates that student does not get equal access to the app or equal time for practice during class (Item 25 mean=1.63) because every school is supplied with only one tablet for 30 students per class. Thus learning using the app hardly occurs since it requires proper attention and time to do practices as exercises.

In sum, usability testing of the LND app shows that it has low usability and accessibility besides many design problems, which also relate to app usability, as shown in Table IV.

Technology is important to enhance student success, dedication, and overall involvement in language learning. It provides students with unrestricted access to different services and methods to promote language acquisition in schools utilizing mobile apps [19]. Primary education is where the pupil needs extra support to learn languages in the classroom using a mobile application. If the mobile language learning application is convenient and simple to use, then the success of the students would also be good. However, results from this study demonstrate that there are usability, design, and accessibility issues of the LND app, which may not have contributed to students learning, as shown by their poor examination performance [36].

TABLE II. RESULTS FOR DESIGN ISSUES OF LND

Item No.	Items	N	M	SD
13	The font size is easy to read.	57	1.94	.44
14	The application provides navigation keys.	57	1.14	.35
15	The icons and buttons are attractive and recognizable.	57	1.70	.57
16	The color scheme of buttons is attractive.	57	1.95	.23
17	The color scheme of the application screen is attractive.	57	1.56	.50
18	The application provides useful voice instructions.	57	1.00	.00
19	The application provides animations for learning.	57	1.00	.00
20	The application provides videos for learning.	57	1.00	.00

Note: Scale ranging from 1-Strongly Disagree to 5-Strongly Agree

TABLE III. RESULTS FOR ACCESSIBILITY OF LND APP

Item No.	Items	N	M	SD
21	The application shows too many advertisements.	57	4.91	.29
22	The application provides a variety of questions in its question bank.	57	1.19	.40
23	The application provides self-recommendations for questions.	57	1.02	.13
24	Each student gets equal access to the application in class.	57	1.61	.82
25	Each student gets equal time for the practice of the application in class.	57	1.63	.70

Note: Scale ranging from 1-Strongly Disagree to 5-Strongly Agree

TABLE IV. OVERALL MEANS AND STANDARD DEVIATION FOR SUBSCALES

Item No.	Subscales	N	Mean	Std. Deviation
1	Usability	57	2. 095	.218
2	Design Issues	57	1.412	.097
3	Accessibility	57	2.073	.317

Note: Scale ranging from 1-Strongly Disagree to 5-Strongly Agree

B. DMO, AC, and Students' Interview Results

The analysis of interview from DMO, AC, and students is presented in Table V:

The interview results from DMO, AC, and students in Table V concur with teachers' survey findings. Usability test findings from the interview confirmed that the LND app has a complex structure, is not easy to use, has a problem in recognizing the icons and cannot build user confidence after using it. For the design issues, the font size used in the application creates a readability problem. Other issues like navigation keys, icons and buttons, buttons color scheme and screens demotivate users from using this application efficiently. Lastly, the content issue raised by DMO, AC, and students that content used in the application is different from the textbook, thus create issues for the students who still need to finish up the text-book based syllabus for the examination.

TABLE V. INTERVIEW RESULT

Item no.	**Category**	**Sub Category**	**DMO and AC's Response**	**Students' Response**
1	**Usability**	Icons	The icons in the application is difficult to find.	The icons and menu in the application is difficult to recognize.
		Interface	The interface of the application is complex and not easy to scroll.	The interface is not easy to understand.
		Ease of use	The application is not easy to use for users.	The application is not easy to use.
		Assistance to use	It does not provide any assistance for using it.	The application does not help or provide assistance in complex tasks.
		Consistency	There is too many inconsistencies in the application.	There is too much inconsistency in the application.
		Confidence	It cannot build confidence for the user after using the application.	The application is not building confidence after using it.
		Complexity	The application is unnecessarily complex.	The structure of the application is complex.
	Design (Colour scheme, Icons, and Interface)	Font Size	The font size of the text is not suitable to read.	It is not easy to read the text in the application
		Navigation Keys	The application does not provide navigation keys.	The application does not provide navigation keys.
		Icons and Buttons	The icons and buttons are not attractive or recognizable.	The icons and buttons are not attractive to recognize.
		Colour Scheme (Buttons and Screens)	The color scheme of icons and buttons is not attractive.	The color scheme of icons and buttons is not suitable.
		Useful Features (Voice, Video, Animation, and Translation)	The application does not provide voice, video, animations, and pronunciation for learning.	The application does not provide voice, animations, and translation for learning complex words.
3	**Other Issues**	Content	The content in the application is different from the content of the textbook which creates ambiguity for the students to cover up the content in application and textbook as well.	The content in the application is different from the textbooks which is the issue of learning both syllabus at same time.
4	**Overall Comments**		The outdated teaching and assessment methods are not efficient, along with LND mobile application. The application is not rich enough in its current form and also not useful in interactively helping students. It is needed to develop a game-based learning method to teach students effectively.	The current form of LND is not efficient, which could help in learning efficiently and smoothly. Fun based learning should be adopted to overcome all the issues.

V. Conclusion

The purpose of the study was to investigate the usability and design issues from teachers, students, and government officers regarding the LND mobile app used in public schools of Punjab. The results of this study were based on the researchers' and practitioners' understanding of usability and design issues of the LND application. Based on the results, it was found that the interface was not easy to use; even the icon of the application was not easy to find. The majority of respondents were not interested to use the application due to poor interface design, small font size, unattractive color schemes, no assistance in difficulty and the app is non-interactive. Additionally, the content of the application was not consistent with the school syllabus, whereas students are supposed to learn using LND app alongside the teacher's classroom teaching using standard syllabus for their assessment promotion to the next level. Furthermore, feedback from the stakeholders has never been gathered to bring useful improvement to the app. The findings indicate that the application need to be redesigned by addressing all the identified issues and the content should be based on the syllabus of the target class level to achieve learning outcomes. The app should also be highly interactive, attractive by leveraging on visuals such as graphics, appropriate colour scheme and animation, that can be achieved by developing a digital game, as proposed by the respondents. Therefore, future research will involve design of a mobile game-based app following the current mobile design principles and guidelines so that the output will be usable and effective for users [44].

Acknowledgment

The authors would like to acknowledge UKM for the publication sponsorship under grant PP-FTSM-2020. They also acknowledged the efforts of School Education Department, Sheikhupura and Senior Headmistress of Government Girls High School, Kot Sondha, Sheikhupura, Punjab for helping to conduct this research.

References

[1] Kumar and P. Mohite, "Usability of mobile learning applications: a systematic literature review," Journal of Computers in Education, vol. 5, no. 1, pp. 1–17, Dec. 2017.

[2] M. A. Chilton, "Technology in the Classroom: Using Video Links to Enable Long Distance Experiential Learning," Journal of Information Systems Education, vol. 23, no. 1, pp. 51–62, 2019.

[3] W.-L. Siu, T.-S. Lim, Y.-R. Chen, Y.-L. Chen, Y.-A. Jou, and Y.-C. Chen, "Using an English language education APP to understand the English level of students," 2018 27th Wireless and Optical Communication Conference (WOCC), pp. 1–3, Apr. 2018.

[4] J. Pettit and A. Kukulska-Hulme, "Going with the grain: Mobile devices in practice," Australasian Journal of Educational Technology, vol. 23, no. 1, 2007.

[5] ISO 9241-11:2018, "Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts". Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en.

[6] ISO 9241-210:2010, "Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems". Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:en.

[7] 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, no. 1, p. 1, 2013.

[8] B. A. Kumar and P. Mohite, "Usability guideline for mobile learning apps: an empirical study," International Journal of Mobile Learning and Organisation, vol. 10, no. 4, p. 223, 2016.

[9] S. A. Nikou and A. A. Economides, "A comparative study between a computer-based and a mobile-based assessment," Interactive Technology and Smart Education, vol. 16, no. 4, pp. 381–391, 2019.

[10] M. Sarrab, M. Elbasir, and S. Alnaeli, "Towards a quality model of technical aspects for mobile learning services: An empirical investigation," Computers in Human Behavior, vol. 55, pp. 100–112, 2016.

[11] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," MIS Quarterly, vol. 13, no. 3, p. 319, 1989.

[12] "PMIU," Real-time Monitoring and Student Learning Data Across Punjab. [Online]. Available: https://open.punjab.gov.pk/schools/home/landing. [Accessed: 05-Jan-2020].

[13] K. Ishaq, N. A. M. Zin, F. Rosdi, A. Abid, and U. Farooq, "Effectiveness of Literacy & Numeracy Drive (LND): A Students Perspective," 2019 International Conference on Innovative Computing (ICIC), 2019.

[14] K. Ishaq, N. Azan, F. Rosdi, A. Abid, and Q. Ali, "Usability of Mobile Assisted Language Learning App," International Journal of Advanced Computer Science and Applications, vol. 11, no. 1, 2020.

[15] K. Ishaq, N. Azan, F. Rosdi, A. Abid, and Q. Ali, "Usefulness of Mobile Assisted Language Learning in Primary Education," International Journal of Advanced Computer Science and Applications, vol. 11, no. 1, 2020.

[16] K. Ishaq, N. Azan, F. Rosdi, A. Abid, and Q. Ali, "Usefulness of Mobile Assisted Language Learning Application," International Journal of Engineering and Advanced Technology Regular Issue, vol. 9, no. 3, pp. 518–525, 2020.

[17] I. Arif, W. Aslam, and M. Ali, "Students' dependence on smartphones and its effect on purchasing behavior," South Asian Journal of Global Business Research, vol. 5, no. 2, pp. 285–302, 2016.

[18] K. Griggs, L. M. Bridges, and H. G. Rempel, "Library/mobile: tips on designing and developing mobile web sites," Code4lib journal, vol. 8.

[19] M. J. W. Lee and A. Chan, "Pervasive, lifestyle-integrated mobile learning for distance learners: an analysis and unexpected results from a podcasting study," Open Learning: The Journal of Open, Distance and e-Learning, vol. 22, no. 3, pp. 201–218, Apr. 2007.

[20] I. M. I. Ezzelden, "MOBILE LEARNING FOR ENGLISH LANGUAGE LEARNING: BENEFITS AND CHALLENGES," European Journal of Open Education and E-learning Studies, vol. 4, no. 1, pp. 44–59, 2019.

[21] N. Yawasabere, "Benefits and Challenges of Mobile Learning Implementation: Story of Developing Nations," International Journal of Computer Applications, vol. 73, no. 1, pp. 23–27, 2013.

[22] A. Pandey, "Top 5 Benefits of Mobile Learning," EIDesign, 21-Mar-2018. [Online]. Available: https://www.eidesign.net/top-5-benefits-mobile-learning/. [Accessed: 05-Jan-2020].

[23] A. Ali, M. Alrasheedi, A. Ouda, and L. F. Capretz, "A STUDY OF THE INTERFACE USABILITY ISSUES OF MOBILE LEARNING APPLICATIONS FOR SMART PHONES FROM THE USER'S PERSPECTIVE," International Journal on Integrating Technology in Education, vol. 3, no. 4, pp. 1–16, Dec. 2014.

[24] N. Ismail, F. Ahmad, N. Kamaruddin, and R. Ibrahim, "A review on usability issues in mobile applications," IOSR Journal of Mobile Computing & Application, vol. 03, no. 03, pp. 47–52, 2016.

[25] R. P. Cortez and D. Roy, "Screen Interface Design for Mobile-assisted Language Learning in EFL Context: A Case Study in Japan," Journal of Language Teaching and Research, vol. 3, no. 3, Jan. 2012.

[26] A. Nikolov, "Design principle: Consistency," Medium, 03-May-2017. [Online]. Available: https://uxdesign.cc/design-principle-consistency-6b0cf7e7339f. [Accessed: 25-Mar-2020].

[27] C. Y. Wong, R. Ibrahim, T. A. Hamid, and E. I. Mansor, "Usability and Design Issues of Smartphone User Interface and Mobile Apps for Older Adults," Communications in Computer and Information Science User Science and Engineering, pp. 93–104, 2018.

[28] H. Hashim, M. M. Yunus, M. A. Embi, and N. A. M. Ozir, "Mobile-assisted Language Learning (MALL) for ESL Learners: A Review of Affordances and Constraints," Sains Humanika, vol. 9, no. 1-5, 2017.

[29] R. Kraleva, V. Kralev, and D. Kostadinova, "A Conceptual Design of Mobile Learning Applications for Preschool Children," International Journal of Computer Science and Information Security, vol. 14, no. 5, pp. 259–264, May 2016.

[30] K. A. Butler, "Usability engineering turns 10," interactions, vol. 3, no. 1, pp. 58–75, Feb. 1996.

[31] J. Dumas and J. Fox, "Usability Testing," Human Factors and Ergonomics Human-Computer Interaction, pp. 231–251, Feb. 2009.

[32] H.-J. Jung, "Fostering an English Teaching Environment: Factors Influencing English as a Foreign Language Teachers Adoption of Mobile Learning," Informatics In Education, vol. 14, no. 2, pp. 219–241, Jan. 2015.

[33] M. Park and T. Slater, "A Typology of Tasks for Mobile-Assisted Language Learning: Recommendations from a Small-Scale Needs Analysis," TESL Canada Journal, vol. 31, p. 93, 2015.

[34] T.-T. Wu, Y.-M. Huang, H.-C. Chao, and J. H. Park, "Personlized English reading sequencing based on learning portfolio analysis," Information Sciences, vol. 257, pp. 248–263, 2014.

[35] T. Kurisu, S. Matsumoto, T. Kashima, and M. Akiyoshi, "A study on constructing user adaptive learning environment to realize sustainable self-study," 2014 IEEE 7th International Workshop on Computational Intelligence and Applications (IWCIA), 2014.

[36] D. Bozdoğan, "MALL Revisited: Current Trends and Pedagogical Implications," Procedia - Social and Behavioral Sciences, vol. 195, pp. 932–939, 2015.

[37] J. Burston, "MALL: the pedagogical challenges," Computer Assisted Language Learning, vol. 27, no. 4, pp. 344–357, 2014.

[38] Y. Liu, H. Li, and C. Carlsson, "Factors driving the adoption of m-learning: An empirical study," Computers & Education, vol. 55, no. 3, pp. 1211–1219, 2010.

[39] S.-C. Cheng, W.-Y. Hwang, D.-W. Wen, S.-Y. Wu, C.-H. Hsiehe, and C.-Y. Chen, "A Mobile and Web System with Contextual Familiarity and its Effect on Campus English Learning," 2010 Third IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning, pp. 222–224, Apr. 2010.

[40] L.-H. Wong and C.-K. Looi, "Mobile-Assisted Vocabulary Learning in Real-Life Setting for Primary School Students: Two Case Studies," 2010 6th IEEE International Conference on Wireless, Mobile, and Ubiquitous Technologies in Education, pp. 88–95, Apr. 2010.

[41] A. Nichols, A. Billey, P. Spitzform, A. Stokes, and C. Tran, "Kicking the Tires: A Usability Study of the Primo Discovery Tool," Journal of Web Librarianship, vol. 8, no. 2, pp. 172–195, Mar. 2014.

[42] R. Tahir and F. Arif, "Framework for Evaluating the Usability of Mobile Educational Applications for Children," Society of Digital Information and Wireless Communications, 2014.

[43] S. Shafiq, and A. T. Khan, "Role & value of usability in educational learning via game based apps," International Journal of Scientific and Technology Research, vol. 7, no. 11, pp. 70-77, 2018.

[44] S. A. Alserri, N. A. M. Zin, and T. S. M. T. Wook, "Gender-based engagement model for designing serious games," 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI), 2017.

RESEARCH PROFILE

Kashif Ishaq is a Ph.D. scholar in Faculty of Information Science and Technology, Univers iti Kebangsaan Malaysia. He is working on his Ph.D. research title "Serious Game Design Model for Language Learning in Cultural Context". He received his Masters in Information Technology from Department of Computer Science, University of Management and Technology, Pakistan. His area of expertise are Serious Games, MALL, E-Learning, and Usability.

Nor Azan Mat Zin received the Ph.D. in 2005 and working as Professor in Research Centre for Software Technology and Management (SOFTAM), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia. She is t he Head of Games lab, Multimedia Software and Usability Research Group and area of specialization is Serious games, HCI (Accessibility), E-learning Technology.

Fadhilah Rosdi received the Ph.D. from University of Malaya and working as Senior Lecturer in Research Centre for Software Technology and Management (SOFTAM), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia. Her area of specialization is Speech Processing Knowledge Based System.

Adnan Abid was born in Gujranwala, Pakistan, in 1979. He received the B.S. degree from the National University of Computer and Emerging Science, Pakistan, in 2001, th e M.S. degree in information technology from the National University of Science and Technology, Pakistan, in 2007, and the Ph.D. degree in computer science from Politecnico Di Milano, Italy, in 2012. He is currently an Associate Professor with the Department of Computer Science, University of Management and Technology, Pakistan. His research interests include computer science education, MALL, information retrieval, and data management. He is a member of the IEEE Computer Society.

Intelligent Risk Alarm for Asthma Patients using Artificial Neural Networks

Rawabi A. Aroud[1], Anas H. Blasi[2]
Computer Science Department[1]
Computer Information Systems Department[2]
Mutah University Al Karak, Jordan[1, 2]

Mohammed A. Alsuwaiket[3]
Computer Science and Engineering Technology Department
Hafar Batin University, Hafar Batin
Saudi Arabia

Abstract—**Asthma is a chronic disease of the airways of the lungs. It results in inflammation and narrowing of the respiratory passages; which prevents air flow into the airways and leads to frequent bouts of shortness of breath with wheezing accompanied by coughing and phlegm after exposure to inhalation of substances that provoke allergic reactions or irritation of the respiratory system. Data mining in healthcare system is very important in diagnosing and understanding data, so data mining aims to solve basic problems in diagnosing diseases due to the complexity of diagnosing asthma. Predicting chemicals in the atmosphere is very important and one of the most difficult problems since the last century. In this paper, the impact of chemicals on asthma patient will be presented and discussed. Sensor system called MQ5 will be used to examine the smoke and nitrogen content in the atmosphere. MQ5 will be inserted in a wristwatch that checks the smoke and nitrogen content in the patient's place, the system shall issue a warning alarm if this gas affects the person with asthma. It will be based on the Artificial Neural Networks (ANN) algorithm that has been built using data that containing a set of chemicals such as carbon monoxide, NMHC (GT) acid gas, C6H6 (GT) Gasoline, NOx (GT) Nitrogen Oxide, and NO2 (GT) Nitrogen Dioxide. The temperature and humidity will be also used as they can negatively affect asthma patient. Finally, the rating model was evaluated and achieved 99.58% classification accuracy.**

Keywords—Asthma; ANN; data mining; intelligent systems; machine learning; traffic-related pollution

I. Introduction

The development of human beings today has led to a heavy price. It is pollution that exists in our time, which increases continuously with every drop of fuel burned by human, and with declining air quality in urban areas, the risk of stroke, heart disease, lung cancer and acute and chronic respiratory diseases, including asthma is increasing [1]. Moreover, Air Pollution and Children's Health report, according to the World Health Organization in 2018, shows that 93% of children worldwide those under the age of fifteen breathe polluted air that puts their health and development at great risk. Estimates that around 600,000 children died in 2016 from acute respiratory infections caused by polluted air [2].

Asthma is a condition in the airways that occurs in the lungs. The muscles tighten around the airways, and excess swelling and irritation of the airways is called inflammation. In fact, it causes narrowing of the airways, coughing, wheezing, chest tightness, or trouble breathing. So, if asthma is left untreated, it may cause long-term lung function loss. Furthermore, when you are exposed to an asthma trigger, the air passages become more inflammatory or swollen than usual, making breathing more difficult or making illnesses worse [3].

The chemicals that have been studied in this paper are affecting the asthma patients negatively. For example, carbon monoxide is produced from partial oxidation (incomplete combustion of carbon) and organic compounds such as coal, this occurs when oxygen is scarce, or when the heat is very high. NMHC (GT) acid gas, especially in the natural gas field, is any gas mixture containing significant amounts of Hydrogen Sulfide (H2S) or carbon dioxide (CO2) or similar gases with an acidic character. C6H6 (GT) Benzene (or benzol) is liquid volatile color and one gasoline vehicles (fuel) and its highly flammable fumes are carcinogens and have a strong smell and jet. NOx (GT) Nitrogen Oxide is also known as dioxide nitrogen oxide or nitrogen monoxide, it is famous in the name of laughing gas for its stimulant effects when inhaled, it is a chemical compound with the chemical formula N2O, in the natural state it is a colorless gas, non-flammable, has a pleasant breath. NO2 (GT) Nitrogen Dioxide. Nitrogen dioxide is one of many nitrogen oxides, having the formula NO2 is a natural gas, brownish-red in color with a sharp pungent odor.

Other important factors have been considered in this paper are the temperature which can cause irritation or inflammation in the airways of an asthma patient, and Relative Humidity (RH) which is the biggest problem for asthma patients, it increases airway resistance to air flow, in addition to narrowing the bronchi, provoking coughing and an increase in mucus secretion, the secret of the bed bug recovering with high humidity is hidden to everyone. In addition to reducing vitamin.

Finally, the Absolute Humidity (AH) which may cause shortness of breath and respiratory diseases in healthy children. In addition, it could be the reason behind the emergence of asthma in the most sensitive and exposed individuals and also this may be due to mold, fungi, bacteria, dust insects, and even cockroaches.

This paper aims to predict the environmental chemicals such as chlorine gas, sulfur dioxide and smoke, which are the most gases affecting the asthma patient using Artificial Neural

Networks (ANN). The focus of this paper is on asthma patient. In general, asthma occurs when inhaling chemical vapors or gases. In 2015, the number of asthma patients exceeded 358 million compared to 183 million in 1990, causing 397,100 deaths in 2015 [4]. There is no effective treatment for asthma, but symptoms can be alleviated. It is necessary to follow a specific plan for the proactive management and control of symptoms. This plan includes reducing exposure to allergens and assessing the severity of symptoms and the use of medications [5].

Researchers recently created an artificial intelligence-based diagnostic algorithm by programming a GPU to act as a neural network, and by applying deep learning using a GPU. However, the team trained the neural network to identify and differentiate diseases. Finally, they showed a reliable result with high accuracy [6]. In this paper, a sensor system called MQ5 is used to examine the smoke and nitrogen content in the atmosphere. It is inserted in a wristwatch that checks the smoke and nitrogen content in the person's place, the system shall issue a warning if this gas affects the person with asthma. It will be based on the ANN algorithm.

The paper is organized as follows: Section 2 reviews the related work about predicting the risk of asthma symptoms using different data mining techniques, Section 3 describes the process followed to prepare the data including data understanding, selecting, transforming, and model building. Section 4 describes the interpretation and evaluation of the results. Finally, Section 5 discusses the conclusions and draws the future work.

II. Related Work

There are a large number of research papers were used data mining in asthma, some of them concentrate on predicting the risk of asthma symptoms' and others on understanding the relationship between allergens and asthma [7], including those who focused on smart diagnosis of asthma [8] and so on. However, the difference between this paper and other studies is that this paper is predicting the chemicals examined by a sensor called MQ5 to relieve the symptoms of asthma, so that a person with asthma should take precaution from where they are located. While other studies are focusing on the causes and symptoms of asthma and very limited who used the data mining for this matter.

In [9], a skilled diagnosis system for asthma and pulmonary embolism was developed and an algorithm to correctly distinguish between asthma and pulmonary embolism was developed as well. The researcher collected the data where he obtained 3657 records of the disease and the data were processed. Artificial neural networks algorithm was used to determine the need to operate the EDS system and perform confirmatory. Artificial Neural Networks algorithm (ANN) has a high 95% accuracy for asthma and pulmonary embolism samples, among the 1492 patients with respiratory disease, 1442 were classified correctly.

The authors [10] talked about the initial prediction of asthma away from the traditional method and with the help of Deep Neural Network (DNN) and Support Vector Machine (SVM). The aim of this study is to develop an algorithm and determine its effectiveness in the diagnosis of asthma. Data were obtained from Kendai University hospital. 566 patient records were collected. However, the search network was conducted on the basis of medium, after using both algorithms, the deep DNN obtained a high accuracy rate with 98% of the total prediction.

Another paper [11], the researcher sought to predict the risk of asthma attacks using machine learning approaches such as naïve Bayes (NB), Support Vector Machines (SVM), and Random Forests (RF). However, the study was conducted on 5 million records of infected patients and the goal was to reduce mortality so that it works to predict early in the risk that causes death. Logistic regression was used which was optimal in predicting the event. The data was validated using the Asthma Learning Health System (ALHS) and was developed to validate the asthma health system educational model. This work was carried out and with the support of asthma in the UK also an (ALHS) data set was established with funding from the National Council for Environmental Research (NCER).

Another paper focused on skilled diagnosis of asthma through machine learning algorithms [12], the k-NN and SVM algorithms were used, in addition to 169 people with asthma were tested and set of processes were used to implement the algorithm such as input organization, preprocessing, data tuning, and output evaluation. Tehran hospital was used to obtain the data, 250 records were taken, and data processing was done in two steps the first was the removal of incomplete data and the second step is to select the most important features that can be utilized in the algorithm. In the results, the researcher obtained the data through Canvas Orange and the implementation was in Python. Finally, SVM algorithm achieved the best results.

Another paper where authors introduced the development of the Lasso logistic regression model in 2015 [13] to predict asthma. In this research, the focus was on pediatric patients receiving medical care. The goal was to use administrative claims data for pediatric residents enrolled in Medicaid to train and test already deep in practice by comparing their predictive power. The Lasso logistic regression model served as a benchmark comparing the results of the deep learning model.

According to [14], the authors talked about machine learning was applied to the continuous biomarker so that the data provides an automatic respiratory novel for asthma in children using the Pediatric Asthma Guide (PAS) as a standard for clinical care. The ANN algorithm was applied to create an automatic respiratory score and validated by two approaches. However, ANN was compared with normal regression models and Poisson. Finally, results obtained an initial group of 186 patients and 128 patients met the inclusion criteria.

According to [15], the authors focused on predicting the disease of asthma using machine learning classification algorithms. Authors in this paper were used some machine learning algorithms such as SVM, ANN, k-NN and random forest algorithms. SVM algorithm have achieved 98% compared to other algorithms. MLP achieved 100% specificity compared to other algorithms. ANN achieved 100% sensitivity compared to other algorithms.

Due to the large proportion of people who affected by asthma in this world. Researchers have prepared studies to reduce its risks. Most papers have been working on developing a smart system that can distinguish between asthma and another disease, and others have predicted the initial diagnosis of asthma, also some of them who trained more than one algorithm to know which algorithms are the most accurate in predicting asthma and its symptoms.

This paper differs from other research in that it focuses on chemicals that have a significant impact on asthma patient in order to design a wrist watch can predict chemicals to issue a warning at the appropriate time that the patient's location is a danger to him/her to take an appropriate action to prevent the danger.

III. Research Methodology

In order to get a better insight into the best predictability in chemicals a patient with asthma greatly helps to know the percentage of chemicals present everywhere. The dataset which collected for this study of the most common chemicals that appeared in one of the cities of Italy that contain gas Carbon monoxide, NMHC(GT), C6H6(GT), NOx (GT), and NO2(GT). These gases and other factors such as temperature and humidity would help to test and extract the model and know the accuracy that we can get from the ANN algorithm.

The critical step here as shown in Fig. 1 is a Knowledge Discovery in Databases (KDD) methodology which will be used as a methodology to manage all the processes that include data selection, preprocessing, data cleansing, building a data mining model and evaluating the results.

A. Selection

Data were obtained from the UC Irvine Machine Learning Repository website [17] containing 9358 rows of decimal numbers for chemicals within an Italian city. Data recorded from March 2004 to February 2005. The data set properties are multivariate; time series and the attribute characteristics is real.

B. Preprocessing

The data contained a set of empty and missing rows data was reorganized by disposing of empty and incomplete rows using Python programing language to convert numbers to be between 0 and 1. Data transformation through the alternative standardization is scaling features to lie between a given minimum and maximum value, often between zero and one, or so that the maximum absolute value of each feature is scaled to unit size. This can be achieved using MinMaxScaler or MaxAbsScaler in Python.

Table I shows a sample preprocessed data, where the target (output) is representing the scale of risk between 0 (lowest risk) and 1 (highest risk) for asthma patients, while the other eight attributes are the chemicals (inputs) of the proposed model.

C. Data Mining

In the past ten years, Artificial Intelligence (AI) systems have been the best performing. Deep learning is actually a new name given to the AI approach and it has been called Artificial Neural Networks (NN), which started a long time ago more than 70 years ago. ANN were first proposed in 1944 by Warren McCullough and Walter Bates, researchers at the University of Chicago who moved to the Massachusetts Institute of Technology in 1952.

ANNs were a major area of research in both neuroscience and computer science until 1969. To do machine learning, where the computer learns to perform some tasks by analyzing training examples [18]. The structure and operation of the ANN can be described by the abstract model of the neural network of neurons, also called units or nodes. They can capture information from outside or from other neurons, transfer them to other neurons, or output them as a final result. There are positive and negative weights that represent an exciting or inhibiting effect. If the weight is zero, one neuron does not affect communication on the other hand. Neural networks can have a variety of different structures. These networks are also referred to as feedback networks or feedback networks [19][20]. In addition to a simple visualization mentioned in Fig. 2 to show how the inputs compared with each other.

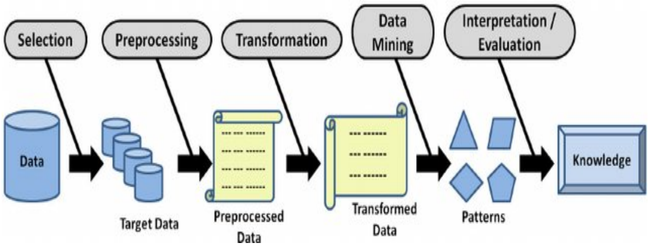


Fig. 1. Knowledge Database Discovery (KDD) Processes [16].

TABLE I. Sample of Preprocessed Data

Carbon monoxide	NMHC (GT)	C6H6 (GT)	NOx (GT)	NO2 (GT)	T	RH	AH	Target
0.66	0.79	0.79	0.60	0.58	0.6	0.6	0.8	0.66
0.84	0.47	0.47	0.93	0.76	0.4	0.7	0.5	0.65
0.78	0.72	0.72	0.80	0.51	0.4	0.8	0.5	0.65
0.9	0.35	0.35	0.93	0.65	0.3	0.8	0.4	0.65
0.85	0.35	0.35	0.92	0.67	0.3	0.8	0.5	0.64
0.64	0.79	0.79	0.55	0.51	0.6	0.7	0.8	0.64
0.96	0.35	0.35	0.91	0.56	0.3	0.8	0.4	0.64
0.66	0.79	0.79	0.60	0.58	0.6	0.6	0.8	0.66

ANNs represent a series of algorithms that seek to identify basic relationships in a set of data through a process that mimics the way the human brain works. It can adapt to changing inputs; therefore, the network generates the best possible result without having to redesign the output standards. In fact, the algorithm of the neural network in this paper was the best in predicting the chemical gases in the weather. ANN algorithm will be explained using Python programming language.

A graphical representation of the proposed ANN as mentioned in Fig. 2. However, the first set of 8 nodes is the inputs. The second set of 5 nodes is the hidden layer. The last set of three nodes is the output layer.

Fig. 3 shows how the inputs have compared to each other and how to represent the ratios of the chemicals that were recorded, in addition to represent the outputs in terms of risk with representations of the impact on the asthma patient. However, the high risk is in blue color, the medium risk is in orange color, and the low risk is in green color.

Python programming language is considered the most suitable and has been used in this paper because it is designed to be extendable with compiler code for proficiency, also many tools are available to facilitate Python integration and software code.

In this paper, a deep learning model has been applied using ANN algorithm to build the classification model. As mentioned previously in this paper, the ANN algorithm is a neural network of nutrition consisting of more than one hidden nonlinear layer. It is characterized by a combination of weight matrices, bias vectors, and a non-linear activation function [21]. Then, the ANN algorithm was built to construct the required model with combinations of testing and training the input layer.

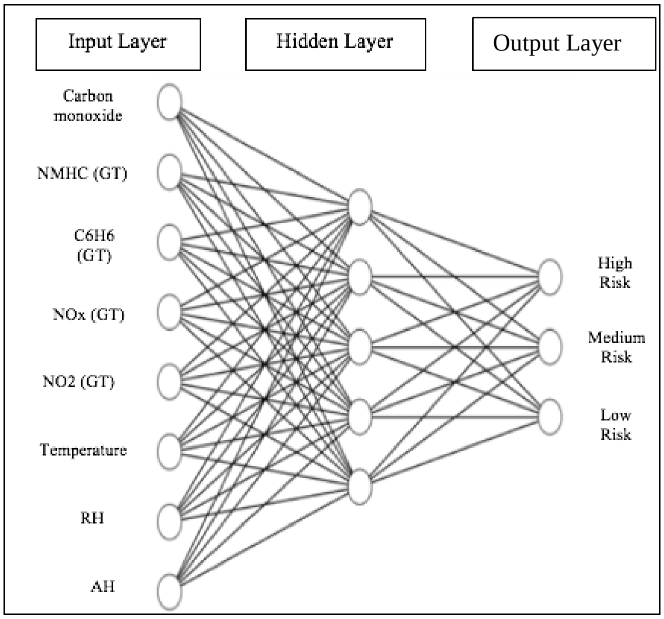


Fig. 2. Graphical Representation of the Proposed ANN.

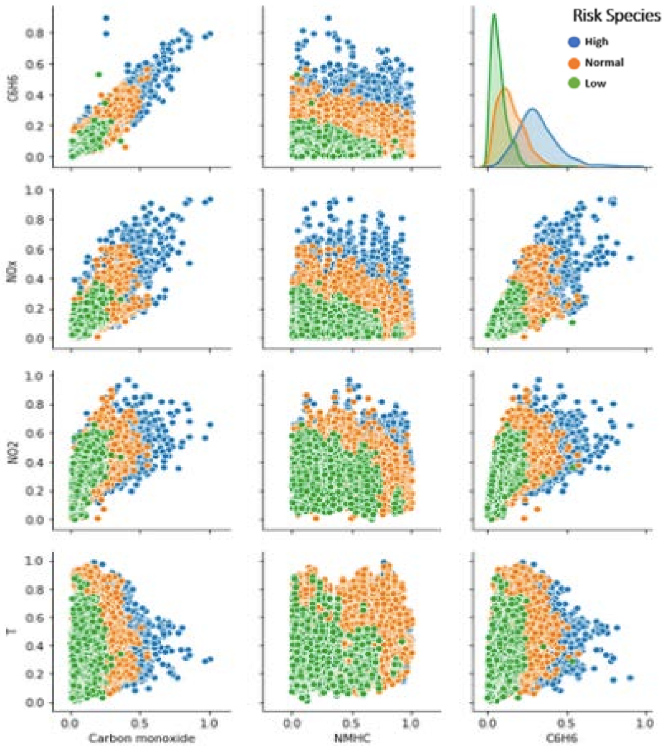


Fig. 3. Sample of Inputs Visualization with Comparison.

The proposed model has been used to predict the risk of these chemicals present in the region if they are high, medium or low risk for the asthma patients. The data have been divided into two sets; one is the training set with 70% of the data used to train the classifier for the prediction result. The other set is the test set with 30% of the data used to test the classifier. The proposed ANN has three layers, input, hidden and output layer. However, the Input layer has all the chemical gases on which the study was conducted with temperature and humidity. On other side, the output layer was made up of three values, each one indicating the types of risk, meaning that the values are focused on the type of risk, if it is high, medium or low risk. The weight matrix has been used to connect the inputs to the hidden layer of the proposed ANN. Each node in the input layer is connected to each node in the hidden layer. Weight values were randomly selected between -1 and 1. The Classification Accuracy (CA) that achieved from the proposed model is 99.58%.

IV. Results Evaluation

In recent years, education has revolutionized science and knowledge in machine learning, especially seeing the computer. In this approach, the Artificial Neural Network (ANN) is trained, often in a supervised manner, using backpropagation due to an ANN and an error function, the method calculates the error function of the neural network weights. Huge quantities of specific training examples are needed, but the resulting classification accuracy is impressive, and sometimes it beats humans. The application is implemented in Python programming language.

The results obtained are in terms of percentage of accuracy. ANN gives high accuracy which is 99.58%. It was rated that the proportion of chemical gases if exceeded 0.6 out of 1, it will be considered high risk and can affect the asthma patient and in this case the patient will be warned to move from the area of risk. If the chemical gases are in between 0.3 and 0.6, it will be considered medium risk of the asthma patient and also in the case the patient will be notified of the level of risk, but if it was less than 0.3 out of 1, the proportion of chemical gases do not affect the asthma patient and will not warn the patient for any risk.

Table II shows the stratified cross-validation that seeks to ensure that each fold is representative of all strata of the data. Generally, this is done in a supervised way for classification and aims to ensure each class is approximately equally represented across each test fold (which are of course combined in a complementary way to form training folds), knowing that the weight values were randomly selected between -1 and 1 using Python.

Stratified cross validation in Table II consists of some important evaluation methods that assess the proposed model of ANN. Here the value of correctly classified instances (Classification Accuracy) is very high with 99.58, and the value of Mean Square Error (MSE) is very low with 0.0028. However, these results show that the proposed model has achieved very good results.

The intuition behind this relates to the bias of most classification algorithms in Table III detailed accuracy by class obtained after feature selection. They tend to weight each instance equally which means overrepresented classes get too much weight. Table III shows some evaluation methods for each class (High, Normal, Low).

However, these methods are True Positive rate (TP Rate) which represents the predicted instances as positive and are actually positive (higher better), False Positive (FP) which represents the predicted instances as positive and are actually negative (lower better), Precision which is the percentage of positive instances out of the total predicted positive instances (higher better), Recall which is the percentage of positive instances out of the total actual positive instances (higher better), F-measure is the harmonic mean of precision and recall. This takes the contribution of both, so higher the F1 score, the better. Finally, ROC Area stands for receiver operating characteristic and the graph is plotted against TPR and FPR for various threshold values. As TPR increases FPR also increases.

In general, all results are closely similar to each other and they have very competitive results.

TABLE II. STRATIFIED CROSS-VALIDATION

Correctly classified instances	99.5822
Incorrectly classified instances	0.4178
Kappa statistic	0.9925
Mean absolute error	0.0028
Root mean squared error	0.0528
Relative absolute error	0.7407
Root number of instances	12.1721

TABLE III. DETAILED EVALUATION METHODS BY CLASS

TP Rate	**FP Rate**	**Precision**	**Recall**	**F-measure**	**ROC Area**	**Class**
0.996	0.002	0.991	0.996	0.993	0.987	High
0.996	0.005	0.997	0.996	0.996	0.995	Normal
0.995	0.001	0.998	0.995	0.997	0.994	Low
0.996	0.003	0.996	0.996	0.996	0.993	Weighted avg.

TABLE IV. CONFUSION MATRIX

a	**b**	**c**	**Classified as**
1358	6	0	**a= High**
13	4113	3	**b= Normal**
0	7	1441	**c= Low**

A confusion matrix as shown in Table IV is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand. Here, the row represents the predicted instances and the columns represents the actual instances for the dataset.

V. CONCLUSION AND FUTURE WORK

This study demonstrates that machine learning techniques such as ANNs were utilized to analyze parameters of simple vital signs and finite data. The potential impact of such an outcome is to improve and standardize data management to see the aggravation of acute asthma. This paper revealed several barriers to the integration of disparate data sources, it also processed and disposed of incomplete data. Further validation of the algorithm is imperative to improve data integrity, and to improve and expand the contribution features. Because asthma in children is the most prevalent chronic childhood disease in the future, this study has endeavor to design a wristwatch that contains a sensor MQ5 that examines chemicals in the weather and when the danger increases, it sends an alert message to the patient concerned alert him/her has been exposed to high pollution.

In future work, a different data set will be applied from different regions of the world and different settings of hidden layers will be tested in ANN as well. In addition to use other machine learning algorithms such as decision tree DT [22] or/and fuzzy logic [23], then compare the results with ANNs results.

REFERENCES

[1] "Asthma", www.nhs.uk,19-2-2018, Retrieved 21-3-2020. Edited.

[2] World Health Organization, https://www.who.int/ar/news-room/detail/29-10-2018-more-than-90-of-the-world%E2%80%99s-children-breathe-toxic-air-every-day, Retrieved 25-3-2020.

[3] Webmd website: https://www.webmd.com/asthma/eosinophilic-asthma-symptoms#1, Retrieved 10-2-2020. Edited.

[4] National Heart, Lung, and Blood Institute. Expert panel report 3: guidelines for the diagnosis and management of asthma. NHLBI website. https://www.nhlbi.nih.gov/files/docs/guidelines/asthgdln.pdf. Published August 28, 2007. Retrieved 1-4-2020.

[5] Create an asthma management plan. American Lung Association. https://www.lung.org/lung-health-and-diseases/lung-

disease-lookup/asthma/living-with-asthma/managing-asthma/create-an-asthma-action-plan.html. Retrieved 2-2-2020.

[6] Raja Koduri, “Why the healthcare industry is hacking graphics technology to power machine intelligence”. https://www.cbronline.com/ehealth/healthcare-industry-hacking-graphics-technology-power-machine-intelligence/. Retrieved 2-4-2020.

[7] Platts-Mills TAE[1], Perzanowski M[2]."The use of machine learning to understand the relationship between IgE to specific allergens and asthma" PLOS Medicine Published: November 20, 2018.

[8] Mostafa Langarizadeh Taha Samad-Soltani Maryam Zolnoori zahra mahmoodvand "Intelligent Diagnosis of Asthma Using Machine Learning Algorithms "International Research Journal of Applied and Basic Sciences Available online at www.irjabs.com ISSN 2251-838X / Vol, 5 (1): 140-145 ,2016.

[9] Eddie Custovic , Lejla Gurbeta, Almir Badnjevic,"An Expert Diagnostic System to Automatically Identify Asthma and Chronic Obstructive Pulmonary Disease in Clinical Settings" Scientific Reports volume Published, 2018.

[10] Katsuyuki Tomita ,RyotaNagao[a], HirokazuTouge[a], TomoyukiIkeuchi[a], HiroyukiSano[b], AkiraYamasaki[c], YujiTohda[b]. "Deep learning facilitates the diagnosis of adult asthma "Allergology International. Volume 68, Issue 4, Pages [456-461], October 2019.

[11] Holly Tibble, Athanasios Tsanas, Aziz Sheikh "Predicting asthma attacks in primary care " BMJ Open 9 July 2019.

[12] Mostafa Langarizadeh Taha Samad-Soltani Maryam Zolnoori zahra mahmoodvand. "Intelligent Diagnosis of Asthma Using Machine Learning Algorithms "International Research Journal of Applied and Basic Sciences Available online at www.irjabs.com ISSN 2251-838X / Vol, 5 (1): 140-145 ,2016.

[13] Xiao Wang, Zhijie Wang, Yolande M. Pengetnze, Barry S. Lachman, Vikas Chowdhry. "Deep Learning Models to Predict Pediatric Asthma Emergency Department Visits" Deep Learning MonitorarXiv:1907.11195v1, 2019.

[14] AI[1], Bui N[2], Wagner BD[3], Szefler SJ[1], Vu T[2], Deterding RR[1]. "Novel pediatric -automated respiratory score using physiologic data and machine learning in asthma ". Original Article, 9 December 2018.

[15] Mrs. J. Cathrin Princy 1 , Mrs. K. Sivaranjani 2 "International Journal of Computer Science and Mobile Computing "Survey on Asthma Prediction Using Classification Technique "IJCSMC, Vol. 5, Issue. 7, pg.515 – 518, July 2016.

[16] J. Han, M. Kamber, and J. Pei, “Classification,” Data Mining, pp. 327–391, 2012. doi:10.1016/b978-0-12-381479-1.00008-3.

[17] UC Irvine Machine Learning Repository web site: https://archive.ics.uci.edu/ml/datasets/Air+quality.

[18] MIT News http://news.mit.edu/2017/explained-neural-networks-deep-learning-0414.

[19] Yiming Jiang,[1] Chenguang Yang,[1] Jing Na,[2] Guang Li,[3] Yanan Li,[4] and Junpei Zhong[5]. ”A Brief Review of Neural Networks Based Learning and Control and Their Applications for Robots”. Hindawi Complexity.Volume 2017, Article ID 1895897, 14 pages https://doi.org/10.1155/2017/1895897.

[20] Anas H. Blasi, “Performance increment of high school students using ANN model and SA algorithm”. Journal of Theoretical and Applied Information Technology 95(11):2417-2425. 2017.

[21] KatsuyukiTomita,RyotaNagao[a]HirokazuTouge[a]TomoyukiIkeuchi[a]HiroyukiSano[b]AkiraYamasaki[c]YujiTohda[b]"Deep learning facilitates the diagnosis of adult asthma "Allergology International Volume 68, Issue 4, Pages [456-461], October 2019.

[22] Mohammad Al Lababede, Anas H. Blasi, Mohammed A. Alsuwaiket. “Mosques Smart Domes System using Machine Learning Algorithms”. International Journal of Advanced Computer Science and Applications, Vol. 11, No. 3, 2020.

[23] Anas H. Blasi, “Scheduling food industry system using fuzzy logic”. Journal of Theoretical and Applied Information Technology 96(19):6463-6473. 2018.

Cultural Algorithm Initializes Weights of Neural Network Model for Annual Electricity Consumption Prediction

Gawalee Phatai[1], Sirapat Chiewchanwattana[2*], Khamron Sunat[3]
Department of Computer Science
Faculty of Science, Khon Kaen University
Khon Kaen, Thailand

***Abstract*—The accurate prediction of annual electricity consumption is crucial in managing energy operations. The neural network (NN) has achieved a lot of achievements in yearly electricity consumption prediction due to its universal approximation property. However, the well-known back-propagation (BP) algorithms for training NN has easily got stuck in local optima. In this paper, we study the weights initialization of NN for the prediction of annual electricity consumption using the Cultural algorithm (CA), and the proposed algorithm is named as NN-CA. The NN-CA was compared to the weights initialization using the other six metaheuristic algorithms as well as the BP. The experiments were conducted on the annual electricity consumption datasets taken from 21 countries. The experimental results showed that the proposed NN-CA achieved more productive and better prediction accuracy than other competitors. This result indicates the possible consequences of the proposed NN-CA in the application of annual electricity consumption prediction.**

***Keywords*—Neural network; weights initialization; metaheuristic algorithm; cultural algorithm; annual electricity consumption prediction**

I. Introduction

Electricity is a major driving force for economic development in many countries. The overall demand for power increases continuously, even more, prominent in the future.

APEC is the acronym of the Asia-Pacific Economic Cooperation that is a cooperative economic group in the Asia-Pacific region. The high growth rates in recent decades of APEC results in a significant increase in electricity consumption. APEC energy data has proved essential in tracking energy consumption, reduction, and in determining the group’s renewable energy goals. APEC is committed to improving efficient energy technologies; by setting targets and action plans, thereby creating the necessity to predict future electricity consumption usage accurately.

The artificial neural network (ANN) computation is based on the learning process of human perception and the function of the brain’s nervous system, which has been widely applied to various problems in classification, pattern recognition, regression, and prediction. In general, humans have learning processes in which processes are characterized by pattern recognition. The pattern-based learning method is described as follows: People observe unknown objects and perceive their identities as distinct from others, especially when viewed more often and in different ways, which results in learning and memory. The human brain contains numerous processing units linked by several nervous systems that perform rapid analysis and decision making. The artificial neural network represents a simulation of the human brain [1][2]. Many studies regarding ANNs have been conducted for solutions in various disciplines.

A. Background

ANN is a distributed data processing system consisting of several simple calculation elements working together through a weighted connection. This calculation architecture was inspired by the human brain, which can learn intricate data patterns and classify them into general data. ANN can be categorized into several types according to not only instructional and unattended learning methods but also feedback-recall architectures.

ANN's most commonly used architecture is the multilayer neural perceptron (MLP). The weights of MLP can be adjusted using both the gradient-based process and the stochastic-based process. The original gradient-based supervised training algorithm of MLP is the error back-propagation (BP) algorithm [3]. BP and its variants are the most frequently used neural network techniques for classification and prediction [4][5].

However, the gradient-based method has two significant disadvantages: slow convergence speed and being trapped at a local minimum easily because of having a high dependency on the initial parameters (weights) [6][7]. Metaheuristic algorithms can overcome those disadvantages of the gradient-based algorithms. Algorithms of this kind use randomization-based techniques to perform the exploration and exploitation searches [8], which are capable of generating solutions to complex real-life problems that gradient-based methods are unable to solve [9]. The population-based structure is the most efficient and commonly used architecture in metaheuristic algorithms. The two often used categories of metaheuristic algorithms are evolutionary and swarm intelligence algorithms [10][11].

Metaheuristic algorithms were applied as supervised training algorithms of MLPs. For a given problem (input and target values), both the structure and weights of an MLP can be optimized. In this paper, we focus on selecting proper initial values of the connecting weights in an MLP network. A

*Corresponding Author

metaheuristic algorithm will perform the initial weights selection. The existing metaheuristics that used to train MLPs for the annual electricity consumption prediction included the Artificial Bee Colony (ABC) [12][13], Teaching-Learning-Based Optimization (TLBO) [13], Harmony Search (HS) [14] and Jaya Algorithm (JA) [15]. Techniques from prior studies found that the applied ANN model (ANN-TLBO), optimized by the TLBO algorithm to predict electric energy demand outperformed the ANN-BP and ANN-ABC models [13]; In other studies conducted to predict the electricity consumption of the ANN-TLBO in comparison with the ANN-BP, ANN-ABC, ANN-HS, ANN-TLBO, and ANN-JA models; the ANN-TLBO yielded better efficiency than that of the other models [15]. TLBO algorithm itself is two phases algorithm; a teacher phase and a learner phase [16].

Not Free Lunch Theorem (NFL) said that there is no superior optimization algorithm for all optimization problems [17]. Although a variety of evolution-based algorithms have been implemented and examined in the literature for MLP training, recognizing that the question of reaching local minima still exists. The Cultural Algorithm (CA) is very similar to the TLBO because it is also a two-stage algorithm; the population level and the belief space level [18]. This characteristic might lead to a more efficient in the initial weights selection. Therefore, we propose, herein, a new MLP training method based on the CA, in which to develop a single hidden layer neural network for annual electricity consumption prediction.

II. Methodology

A. Multilayer Perceptron for Neural Model Training

MLP is a widely used type of feedforward neural network having a multi-layered structure for complex tasks. There are several layers, namely the input layer, hidden layers, and the output layer. Each layer of MLP comprises of numerous neurons and the connecting weights between the two consecutive layers. The connecting weights are represented by real numbers in [−1, 1]. The input layer is responsible for receiving information for the neural network and sending it to the first hidden layer through the connecting weights. Each hidden layer will contain a layer that is responsible for receiving information for the neural network and sending it to the hidden layer. In an MLP fully interconnected by weights, each neuron of the hidden layer contains summation and activation functions. The weighted summation of input is described in Eq. (1), where I_i is the input variable i, and ω_{ij} is the connection weight between I_i and the hidden neuron j. An activation function is used to trig the output of neurons based on the value of the summation function. The Sigmoid function is most often applied. However, different types of activation functions may be utilized in the MLP.

Each node of the hidden layer calculates its output by Eq. (2). The production of the node j in the hidden layer is described in Eq. (2.) [19].

$$S_j = \sum_{i=1}^{n} \omega_{ij} I_i + \beta_j \tag{1}$$

$$f_j(x) = \frac{1}{1+e^{-S_j}} \tag{2}$$

The outcomes of the lower hidden layer are fed to the adjacent layer. Once all neurons in the last hidden layer produce results, the production of the network will be obtained by Eq. (3).

$$\hat{y}_k = \sum_{i=1}^{m} W_{kj} f_i + \beta_k \tag{3}$$

The initialization of the weights of a neural network is one of the essential problems, as network initialization can speed up the learning process. Zero initialization [20] and Random initialization [21] are generally practiced techniques used to initialize the parameters. Traditionally, the weights of a neural network are set to small random numbers.

B. Cultural Algorithm

Cultural algorithms (CA) is a kind of evolutionary algorithms; it is first presented by R. G. Reynolds [18]. Their computational models are based on principles of human social Cultural evolution that make practical use of the learning process through various agent-based techniques based on experience and knowledge gained over time. The cultural process allows for improved efficiency in finding the optimal solution within a search space and making it easier to find the optimal global solution. The cultural changes within an optimization problem model represent information transmitted within and between populations. The main principle of the CA is to preserve socially accepted beliefs and discard unacceptable beliefs.

The CA can be divided into two main components as a population space and a belief space. Each member of the former part is evaluated through a performance function and may be carried out by an Evolutionary Algorithm (EA). An acceptance function then determines which individuals are to impact the belief space. At each generation, the knowledge acquired in the population search (e.g., the population's best solution) will be memorized in the belief space [22]. The interaction and help between the two spaces are similar to the evolution of human culture [23]. The significant components of CA are shown in Fig. 1.

The CA uses a dual evolutionary mechanism, while lower-level populations help periodically enter the top level of beliefs. On the other hand, a high level of belief will evolve these elite people to influence the lower communities [25]. This mechanism results in the improvement of the population diversity and the convergence characteristics, accordingly. The interested reader can see [18] for more details of CA.

C. Cultural Algorithm for Training Neural Network Model

We propose CA as a training algorithm of the Neural Network model. CA will find a proper set of the initial weights for an MLP, and from now on, we call the proposed algorithm as NN-CA. It can be applied not only for a single hidden but also several hidden layers. Two main aspects must be considered when the approach is used: the representation of the weights as the search agent of the CA; and the selection of the fitness function.

The representation is straightforward, as all the weights of an MLP are organized and indexed to be a row vector. This vector is a search agent of CA. The fitness function will be explained after the presentation of the workflow.

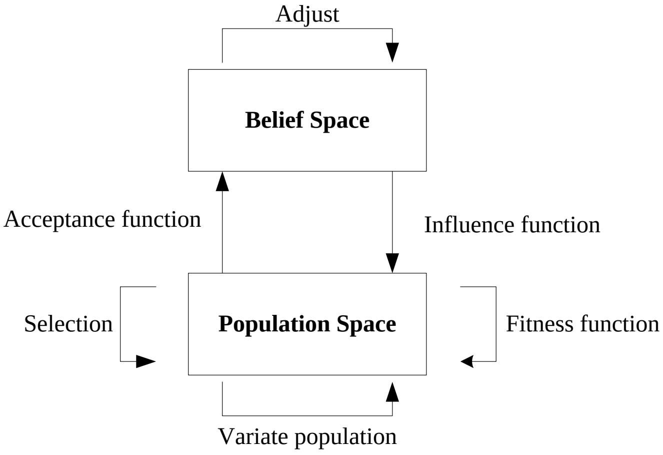


Fig. 1. Flowchart Diagram of Cultural Algorithm [24].

The general steps of the proposed NN-CA are depicted in Fig. 2.

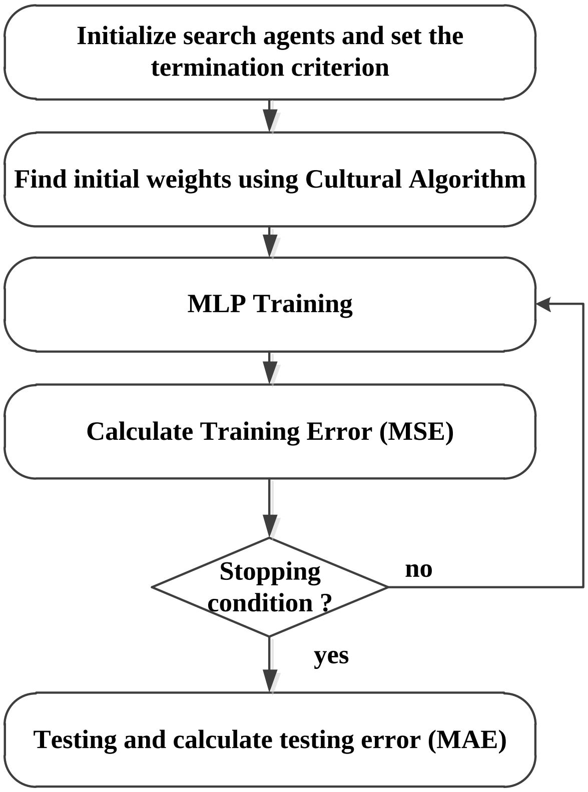

Initialize search agents and set the termination criterion
Find initial weights using Cultural Algorithm
MLP Training
Calculate Training Error (MSE)
Stopping condition ?
no
yes
Testing and calculate testing error (MAE)

Fig. 2. General Steps of the NN-CA.

The workflow of the CA approach applied to train the neural network model may be described in the following steps:

1) Initialization: the search agents in the population and belief spaces are randomly generated for training. Each search agent in a belief space represents a possible MLP. Each dataset is separated as the training part and the testing part.

2) Fitness evaluation: Each possible MLP is evaluated its quality through a fitness function. All the weights of a search agent of belief space are mapped to an MLP, and then each MLP is assessed by the selected fitness function. Typically, the Mean Squared Error (MSE), which is dependent on the neural network training model and the problem of interest, is selected to perform.

3) Update the accepted population in the belief space.

4) Steps 2 to 3 are repeated until the terminated condition is found.

5) The reliability evaluation of the neural network model that has the lowest MSE value will be conducted on the testing part of the dataset to determine the Mean Absolute Error (MAE).

The MSE, which is the average of the error-squared for all training samples, as shown Eq. (4), acts as the fitness function. It depends on the difference between each actual (or the target) its associated output values of the MLP.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(EC_o - EC_p\right)^2 \qquad (4)$$

The Mean Absolute Error (MAE) that evaluates the reliability of each model is shown in Eq. (5).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\left(EC_o - EC_p\right)^2\right|, \qquad (5)$$

where EC_P is annual electricity consumption value produced from MLP and EC_O is the actual annual electricity consumption value.

III. Experimental Results

The experiments aimed to examine the effectiveness of the proposed method for the annual electricity consumption prediction. The neuron network model used was a single hidden layer MLP. All the experiments are programmed in MATLAB, and ran on Intel 2.9 GHz, 8 GB memory. The operating system is Windows 10.

Our study utilized the Asia-Pacific Cooperation (APEC) energy database, which contained the annual electricity statuses of 21 countries in the Asia-Pacific region. There are four input variables: Population (million person), GDP (billion US$), imports (billion US$), and exports (billion US$) were independent variables in model annual electricity consumption (TWh).

Data were divided into two parts: training data (1990 to 2008) and testing data (2009 to 2017); which consisted of population, GDP, imports, and exports data from the World Bank [26]; and annual electricity consumption data, from the Expert Group on Energy Data and Analysis (EGEDA) [27]. Annual Electricity consumption target data are shown in Fig. 3.

The Pearson correlation coefficient (*R*) was applied to examine the dependency between each input variable and annual electricity consumption. All related *R*-values are shown in Table I.

From Table I, the GDP of Russia has a relatively low *R*-value ($R < 0.5$). That means the annual electricity consumption in Russia does not maintain a linear relationship with its GDP parameters.

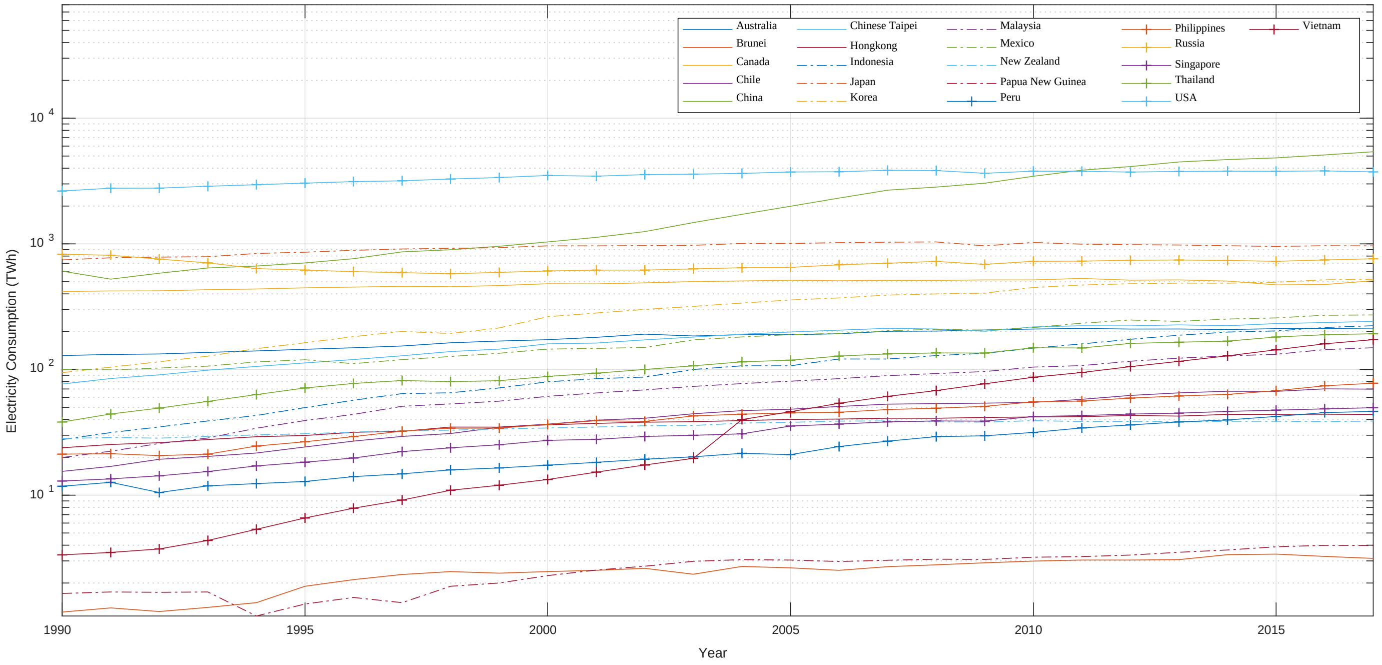


Fig. 3. Variation of annual Electricity Consumption in various Datasets.

TABLE I. EXAMINATION OF THE RELATIONSHIPS (*R*) BETWEEN INPUT PARAMETERS AND OUTPUT OF NEURAL NETWORK MODEL

Country	**Population**	**GDP**	**Import**	**Export**
Australia	0.9379	0.8367	0.8591	0.8451
Brunei	0.9587	0.7768	0.7094	0.7292
Canada	0.7980	0.7971	0.8242	0.8836
Chile	0.9962	0.9196	0.9033	0.9092
China	0.8984	0.9853	0.9893	0.9919
Chinese Taipei	0.9936	0.9435	0.9702	0.9724
Hongkong	0.9886	0.8971	0.8615	0.8751
Indonesia	0.9892	0.9391	0.9174	0.9310
Japan	0.9839	0.5534	0.6855	0.7611
Korea	0.9880	0.9618	0.9385	0.9512
Malaysia	0.9922	0.9443	0.9423	0.9346
Mexico	0.9890	0.9601	0.9889	0.9878
New Zealand	0.9392	0.8402	0.8756	0.8671
Papua New Guinea	0.9462	0.7558	0.6475	0.7930
Peru	0.9288	0.9722	0.9610	0.9439
Philippines	0.9902	0.9414	0.9788	0.9759
Russia	-0.2056	0.5540	0.5528	0.5117
Singapore	0.9867	0.9327	0.9437	0.9487
Thailand	0.9507	0.9287	0.9208	0.9585
USA	0.9458	0.9179	0.9205	0.8443
Vietnam	0.9088	0.9932	0.9939	0.9881

Based on the electric consumption data we studied, the size of MLP is 4:*h*:1, where *h* is the number of neurons in the hidden layer. Because the prediction accuracy depends on the MLP size or *h*, we compare two strategies to study the effect *h*. The first strategy *h* was assigned as 2×N +1, where N is the dimension of dataset features or the dataset features [19]. The second strategy appointed the number of hidden neurons to be 5, 10, 15, and 20 [13]. There are some predefined settings: all BP experiments were executed with 5,000 iterations, each metaheuristic algorithm evolved 5,000 iterations, the population size of CA is 50, the MLPs weights must be in the interval of [-1, 1].

A. *Comparing the Results of Neural Network Models*

The proposed NN-CA was compared with MLP trained by the error back-propagation algorithm (which is label as BP), as well as other metaheuristic algorithm trainers, based on the MSE evaluation measures. Input, hidden layer neurons, and output variables were assigned before starting the experiment. From Table I, there are four input variables: population, GDP, imports, and exports. To determine the suitable network architecture, the BPs were trained with a single hidden layer, incorporating nine hidden nodes that specified by the first strategy, and 5, 10, 15, 20 hidden nodes as determined by the second strategy.

Table II presents the average ranks, Friedman test [28], produced by each competitor, where a lower score is better. The significant differences do exist between the six algorithms. As seen from Table II, NN-CA with a 4-20-1 architecture produced the best overall ranking in comparison with other algorithms, which shows the merits of the proposed NN-CA.

The annual electricity consumption variable was provided in the output data. The overall results that the 4-20-1 architecture of the neural network model was the most superior.

We can see that NN-CA outperforms the other algorithms in all the results of the Friedman rank test with an even lower number of hidden neurons, such as the 4-5-1 neural network model architecture.

The overall results, the 4-20-1 architecture of the neural network model and the 4-5-1 neural network model architecture presented in Tables III and IV.

As demonstrated in each table above, the proposed NN-CA outperformed all other training optimizers and BP. CA can select a proper search agent to be the initial weights of an MLP. Each algorithm within each dataset was statistically compared via the Friedman test. This comparison confirmed the significance, and contrast, of the NN-CA's ability with that of the other trainers.

Fig. 4 shows the MSE convergence curves of the 4-20-1 neural network model architecture utilized in the prediction datasets and trained by ABC, CA, HS, JA, and TLBO.

TABLE II. RESULTS OF THE FRIEDMAN RANK TEST

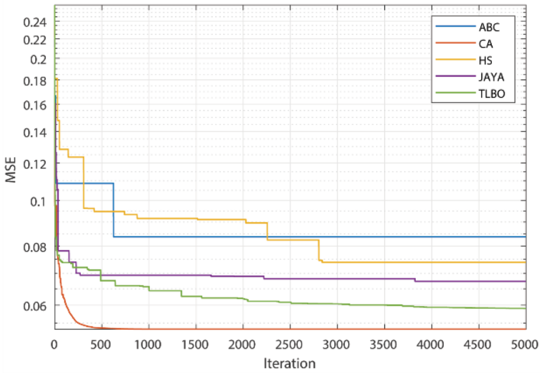
Algorithm	Ranking				
	4-5-1	4-10-1	4-15-1	4-20-1	4-9-1
BP	6.00	6.00	6.00	5.95	6.00
ABC	4.24	4.24	4.24	4.71	4.19
CA	1.31	1.31	1.31	**1.29**	1.43
HS	4.48	4.48	4.48	4.24	4.48
JA	3.24	3.24	3.24	3.05	3.24
TLBO	1.74	1.74	1.74	1.76	1.67

TABLE III. MSE AND MAE VALUES OF DIFFERENT METAHEURISTIC ALGORITHMS WITHIN THE 4-20-1 NEURAL NETWORK MODEL ARCHITECTURE

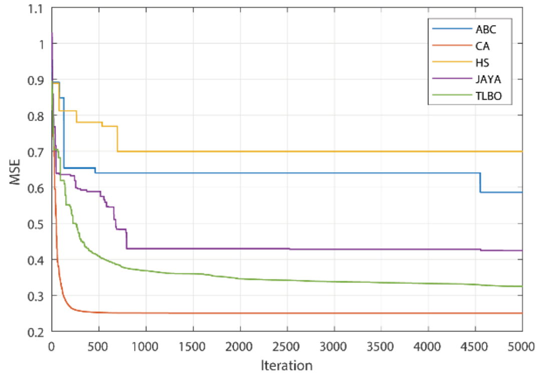
Country	**MSE from Training**						**MAE from Testing**					
	BP	**ABC**	**CA**	**HS**	**JAYA**	**TLBO**	**BP**	**ABC**	**CA**	**HS**	**JAYA**	**TLBO**
Australia	1.8558	0.0838	**0.0533**	0.0739	0.0674	0.0590	10.5534	7.1501	5.1081	6.3893	5.8884	5.2598
Brunei	1.0547	0.5863	**0.2504**	0.6996	0.4240	0.3247	0.4232	0.3971	0.2488	0.4317	0.3233	0.2947
Canada	1.4451	0.3669	**0.0847**	0.4876	0.4062	0.2450	23.3678	16.4532	7.0591	20.9438	18.6190	14.0389
Chile	0.6715	0.1843	**0.0508**	0.1698	0.1135	0.0825	5.5987	5.6646	2.8245	5.4038	4.4637	3.6630
China	4.8680	0.1330	0.0113	0.0398	0.0380	**0.0068**	28.8633	48.9373	10.6728	27.9172	21.3205	6.6458
Chinese Taipei	1.1856	0.0676	0.0185	0.0354	0.0272	**0.0142**	9.4288	11.2984	6.0540	8.0929	7.4067	5.2358
Hongkong	0.7192	0.2669	**0.1195**	0.2321	0.2053	0.1332	2.7506	2.4894	1.4485	2.3302	2.2148	1.4924
Indonesia	1.5123	0.3734	0.0372	0.1236	0.0871	**0.0260**	30.3337	24.8620	6.8910	14.2380	13.2884	6.8579
Japan	2.5163	0.2062	**0.0900**	0.2037	0.1085	0.0941	37.5703	29.8761	17.5061	30.1001	21.8478	19.2708
Korea	2.3068	0.0762	**0.0232**	0.0808	0.0564	0.0358	52.9777	28.9537	16.4857	28.4239	25.5435	18.3189
Malaysia	1.0742	0.3935	0.1245	0.4744	0.2740	**0.1068**	19.8689	18.1442	8.5110	18.9106	14.0960	8.3304
Mexico	2.1067	0.1150	**0.0240**	0.0446	0.0271	0.0249	26.3023	13.8243	5.2876	8.9366	6.4515	5.5591
New Zealand	0.5732	0.2191	**0.1490**	0.2559	0.2225	0.1897	1.7637	1.5507	0.9873	1.4720	1.2715	1.1473
Papua New Guinea	0.4207	0.3025	0.0661	0.2260	0.1488	**0.0640**	0.4476	0.4624	0.1658	0.3498	0.2745	0.1551
Peru	0.7525	0.2076	**0.0241**	0.1279	0.0577	0.0356	7.2074	3.3716	1.3156	3.1291	1.8440	1.1969
Philippines	1.0347	0.2377	**0.0513**	0.1272	0.0981	0.0541	9.4401	5.7589	2.4693	4.5334	3.8628	2.5846
Russia	2.8502	0.4656	**0.1939**	0.4462	0.3205	0.2265	34.3029	33.5951	17.9250	34.1849	27.1204	20.6213
Singapore	0.7203	0.2771	**0.0572**	0.2237	0.1497	0.1131	5.8382	4.9481	2.0600	4.3414	3.4922	2.9320
Thailand	1.3158	0.2126	**0.0756**	0.1322	0.0815	0.0834	19.6150	15.2629	8.3153	11.7838	9.3112	8.0198
USA	5.2565	0.0643	**0.0137**	0.0394	0.0281	0.0161	20.6635	14.4222	6.9762	10.8600	10.8860	7.4154
Vietnam	1.3056	0.1194	0.0062	0.0431	0.0243	**0.0034**	33.5102	13.7530	2.6370	8.0160	5.9483	2.0628

TABLE IV. MSE AND MAE VALUES OF DIFFERENT METAHEURISTIC ALGORITHMS WITHIN THE 4-5-1 NEURAL NETWORK MODEL ARCHITECTURE

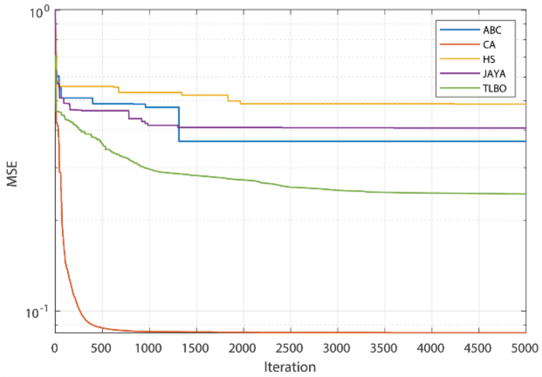
Country	**MSE from Training**						**MAE from Testing**					
	BP	**ABC**	**CA**	**HS**	**JAYA**	**TLBO**	**BP**	**ABC**	**CA**	**HS**	**JAYA**	**TLBO**
Australia	1.5729	0.0658	**0.0461**	0.0672	0.0669	0.0582	8.5458	7.0993	4.9810	6.2592	5.9593	5.3299
Brunei	0.9042	0.5360	**0.3201**	0.6360	0.5469	0.5052	0.4258	0.4001	0.3009	0.4179	0.3795	0.3654
Canada	1.3952	0.3125	**0.0959**	0.4858	0.4712	0.4367	20.4912	13.9783	8.2325	20.4358	20.0953	19.7515
Chile	0.6546	0.1033	**0.0539**	0.1341	0.1226	0.0933	4.9831	4.5533	2.9904	4.7539	4.4267	4.0753
China	4.6984	0.0359	0.0135	0.0170	0.0212	**0.0061**	26.1438	22.6626	9.9075	12.0142	16.7640	6.7927
Chinese Taipei	1.1471	0.0446	0.0211	0.0349	0.0294	**0.0175**	9.7913	7.8809	6.3223	8.3315	7.3382	5.8382
Hongkong	0.6731	0.1734	**0.1276**	0.2009	0.1944	0.1583	2.7399	2.1919	1.6193	2.0849	1.9471	1.4592
Indonesia	1.3943	0.1177	0.0393	0.0575	0.0612	**0.0351**	36.7804	16.6837	8.8297	9.4206	10.1301	7.4799
Japan	2.5308	0.1097	**0.0951**	0.1701	0.1334	0.1087	37.7423	24.2206	20.1802	26.4450	24.2393	21.5002
Korea	1.9595	0.0784	**0.0398**	0.0706	0.0674	0.0483	58.3515	29.5178	20.0582	25.8154	26.5983	22.0616
Malaysia	0.9710	0.2567	**0.1556**	0.3388	0.2905	0.2466	19.5628	14.4917	9.6247	15.1668	15.3101	12.4294
Mexico	1.8606	0.0411	0.0252	0.0280	0.0273	**0.0224**	26.0858	8.2147	5.8130	6.3860	6.3732	5.0313
New Zealand	0.5570	0.1702	**0.1597**	0.2420	0.2261	0.2147	1.5785	1.5426	1.1897	1.4519	1.3562	1.2779
Papua New Guinea	0.2874	0.1446	**0.0891**	0.1318	0.1358	0.0911	0.3320	0.2438	0.2016	0.2531	0.2533	0.1908
Peru	0.7471	0.0977	**0.0392**	0.0851	0.0585	0.0435	6.6411	2.5816	1.3081	2.4125	1.7299	1.4332
Philippines	0.9596	0.1122	0.0627	0.0957	0.0750	**0.0572**	10.0227	4.2566	2.5691	3.7190	3.1442	2.5663
Russia	2.7264	0.2626	**0.1577**	0.3593	0.3216	0.2544	30.1140	26.9641	17.0105	26.5047	24.4000	21.4613
Singapore	0.6790	0.1926	**0.0818**	0.1602	0.1598	0.1235	6.3694	4.0392	2.6527	3.6022	3.6631	3.1494
Thailand	1.2956	0.1195	**0.0841**	0.1008	0.0975	0.0866	19.6608	10.7466	8.3404	9.4941	9.2509	8.2743
USA	5.0420	0.0343	**0.0198**	0.0261	0.0306	0.0201	19.6188	11.7549	8.5884	10.1170	10.1933	8.0254
Vietnam	1.3203	0.0279	0.0054	0.0146	0.0132	**0.0035**	47.5357	6.5294	2.3461	4.5924	4.5872	2.1222



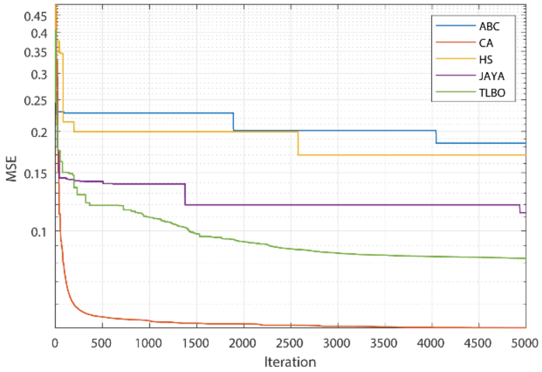
Australia



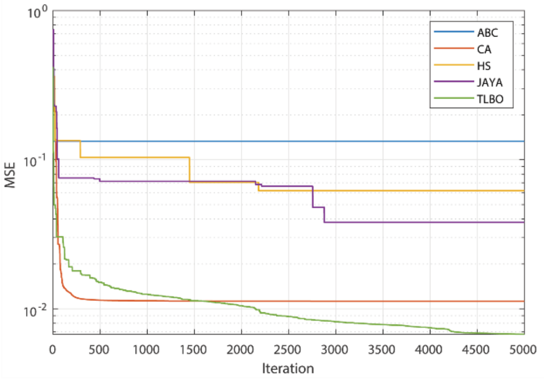
Brunei



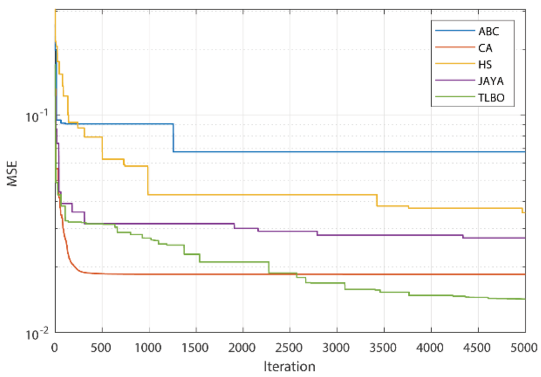
Canada



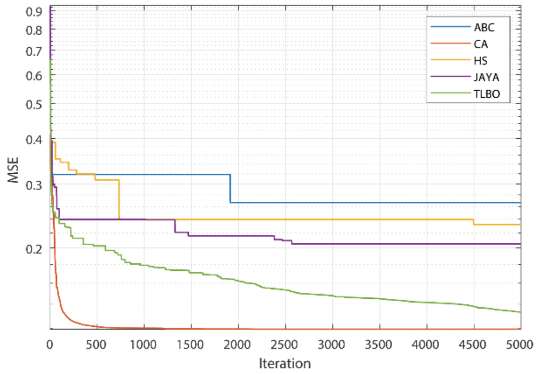
Chile



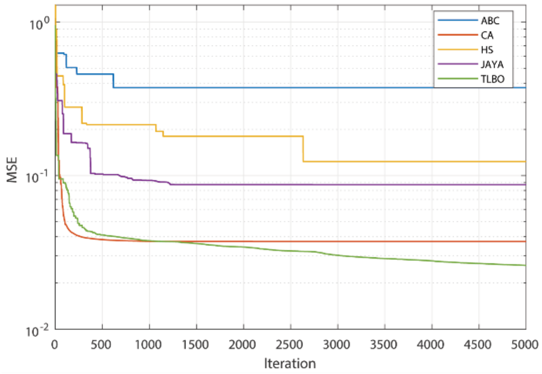
China



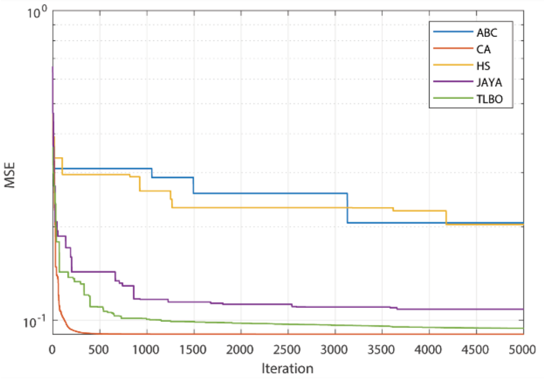
Chinese Taipei



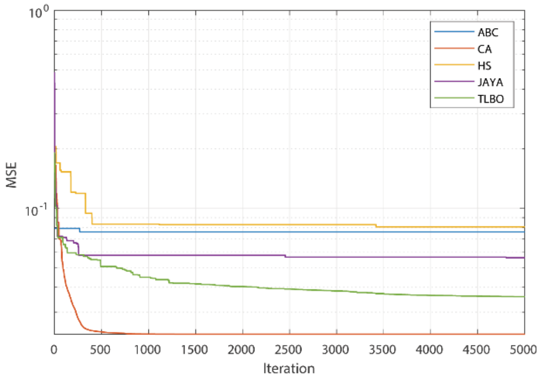
Hongkong



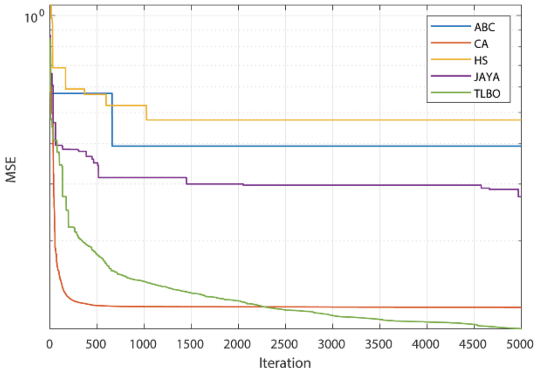
Indonesia



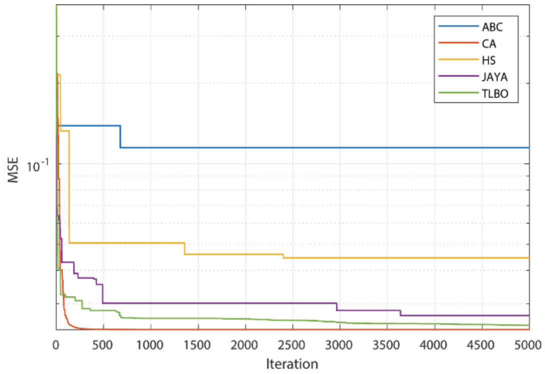
Japan



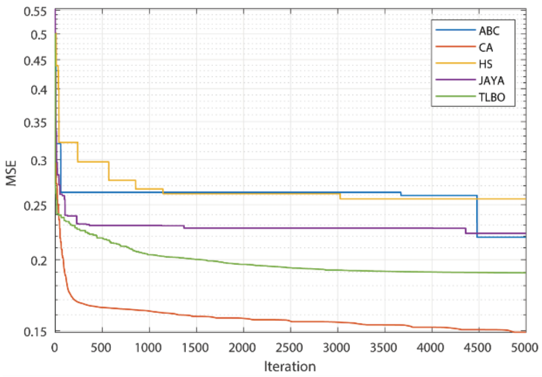
Korea



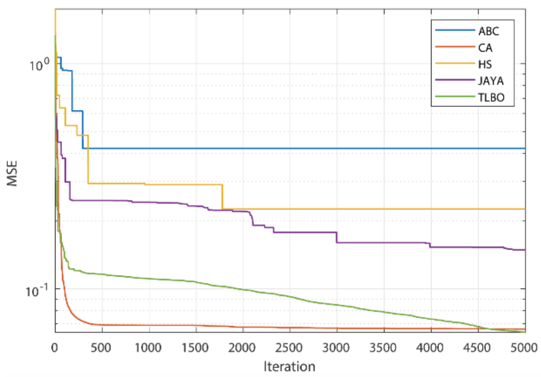
Malaysia



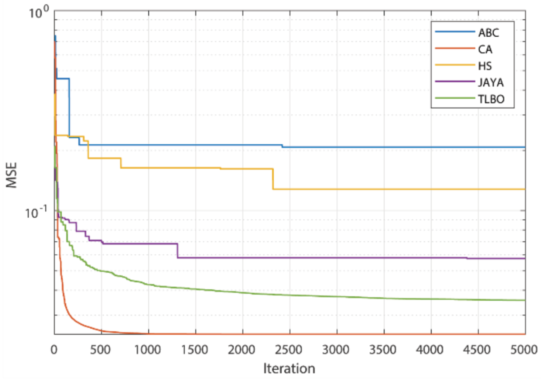
Mexico



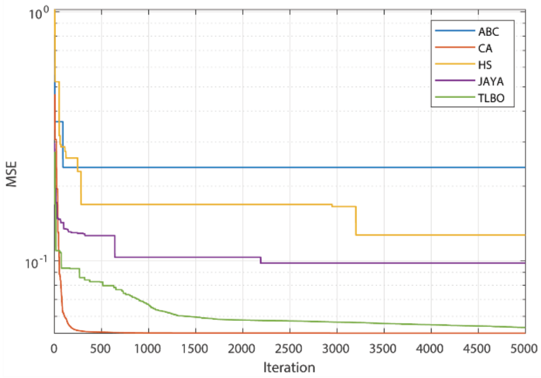
New Zealand



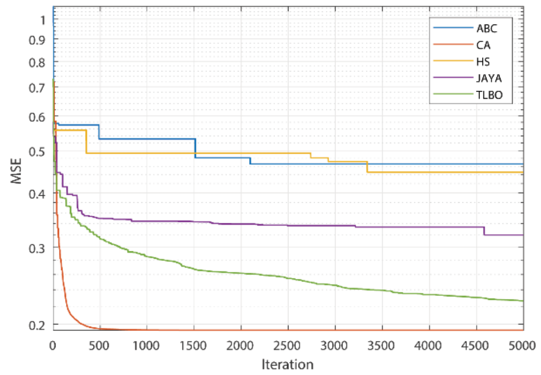
Papua New Guinea



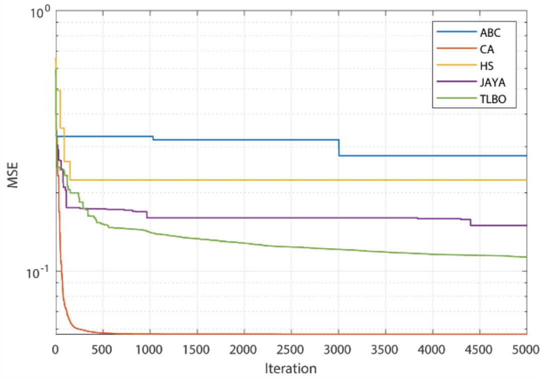
Peru



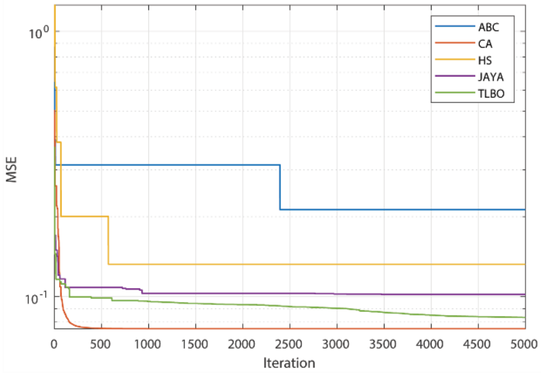
Philippines



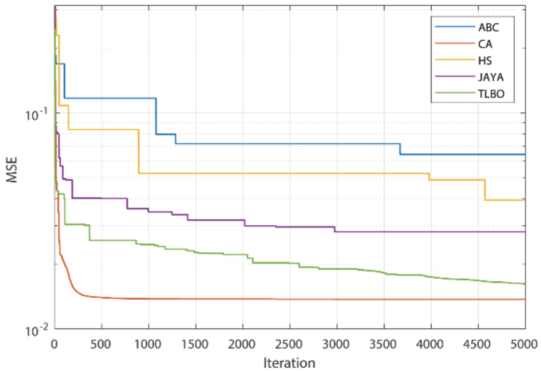
Russia



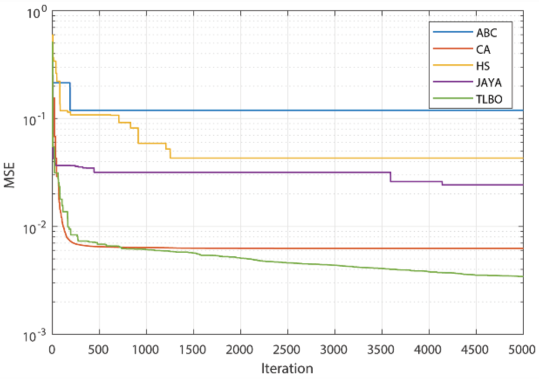
Singapore



Thailand



USA



Vietnam

Fig. 4. MSE Convergence Curves of different Metaheuristics Algorithm with 4-20-1 Neural Network Model Architecture.

The results indicated that the NN-CA algorithm was the fastest convergence speeds in Australia, Brunei, Canada, Chile, Hong Kong, Japan, Korea, Mexico, New Zealand, Peru, Philippines, Russia, Singapore, Thailand, and USA datasets. Within other datasets, the NN-CA was not deemed best; its results remained very competitive in each case.

IV. Conclusion

In this paper, MLP predicted consumer electricity usage, a method based on metaheuristic algorithms for the weights initialization of an MLP was implemented, as well as to analyze annual electricity consumption. The goals of the training problem were to avoid high local optima with convergence to the best solution in the predefined time. The result of the proposed technique was an MLP that has the lowest MSE. The proposed NN-CA outperformed all competitive algorithms, found the best-initialized weight values that the error back-propagation algorithm did not stick at a local minimum and can reduce the MSE effectively. The proposed method was proved to be suitable for the annual electricity consumption prediction, which will accurately support the power network infrastructure plan.

Because the MLP in this paper is a fully connected network, there are a lot of unnecessary weights or links. An MLP having only the necessary weights is not only more compact but also more accurate than the MLP with whole weights. Therefore, determining those weights and removing them from the final model is a necessity. However, this problem is very time-consuming. That is our future work.

Acknowledgment

This research was partially supported by the Advance Smart Computing Lab of the Department of Computer Science, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand.

References

[1] I. A. Basheer and M. Hajmeer, “Artificial neural networks: fundamentals, computing, design, and application,” J. Microbiol. Methods, vol. 43, no. 1, pp. 3–31, 2000.

[2] G. Panchal, A. Ganatra, Y. P. Kosta, and D. Panchal, “Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers,” Int. J. Comput. Theory Eng., vol. 3, no. 2, pp. 332–337, 2011.

[3] D. E. Rumelhart, R. Durbin, R. Golden, and Y. Chauvin, “Back-propagation: The basic theory,” Backpropagation Theory, Archit. Appl., pp. 1–34, 1995.

[4] M. S. Kıran, E. Özceylan, M. Gündüz, and T. Paksoy, “Swarm intelligence approaches to estimate electricity energy demand in Turkey,” Knowledge-Based Syst., vol. 36, pp. 93–103, 2012.

[5] M. E. Günay, “Forecasting annual gross electricity demand by artificial neural networks using predicted values of socio-economic indicators and climatic conditions: Case of Turkey,” Energy Policy, vol. 90, pp. 92–101, 2016.

[6] M. Črepinšek, S.-H. Liu, and M. Mernik, “Exploration and exploitation in evolutionary algorithms: A survey,” ACM Comput. Surv., vol. 45, no. 3, pp. 1–33, 2013.

[7] S. Mirjalili, S. Z. M. Hashim, and H. M. Sardroudi, “Training feedforward neural networks using hybrid particle swarm optimization and gravitational search algorithm,” Appl. Math. Comput., vol. 218, no. 22, pp. 11125–11137, 2012.

[8] X.-S. Yang, Engineering optimization: an introduction with metaheuristic applications. John Wiley & Sons, 2010.

[9] F. W. Glover and G. A. Kochenberger, Handbook of metaheuristics, vol. 57. Springer Science & Business Media, 2006.

[10] T. Bäck, D. B. Fogel, and Z. Michalewicz, Handbook of evolutionary computation. CRC Press, 1997.

[11] E. Bonabeau, M. Dorigo, D. de R. D. F. Marco, G. Theraulaz, and G. Théraulaz, Swarm intelligence: from natural to artificial systems, no. 1. Oxford university press, 1999.

[12] F. Gürbüz, C. Öztürk, and P. Pardalos, “Prediction of electricity energy consumption of Turkey via artificial bee colony: a case study,” Energy Syst., vol. 4, no. 3, pp. 289–300, 2013.

[13] M. Kankal and E. Uzlu, “Neural network approach with teaching–learning-based optimization for modeling and forecasting long-term electric energy demand in Turkey,” Neural Comput. Appl., vol. 28, no. 1, pp. 737–747, 2017.

[14] H. Ceylan, H. Ceylan, S. Haldenbilen, and O. Baskan, “Transport energy modeling with meta-heuristic harmony search algorithm, an application to Turkey,” Energy Policy, vol. 36, no. 7, pp. 2527–2535, Jul. 2008.

[15] G. Phatai, S. Chiewchanwattana, and K. Sunat, “A Comparative of Neural Network with Metaheuristics for Electricity Consumption Forecast Modelling,” in 2018 22nd International Computer Science and Engineering Conference (ICSEC), 2018, pp. 1–4.

[16] R. V Rao, V. J. Savsani, and D. P. Vakharia, “Teaching–learning-based optimization: A novel method for constrained mechanical design optimization problems,” Comput. Des., vol. 43, no. 3, pp. 303–315, 2011.

[17] D. H. Wolpert and W. G. Macready, “No free lunch theorems for optimization,” IEEE Trans. Evol. Comput., vol. 1, no. 1, pp. 67–82, 1997.

[18] R. Reynolds, An Introduction to Cultural Algorithms. 1994.

[19] I. Aljarah, H. Faris, and S. Mirjalili, “Optimizing connection weights in neural networks using the whale optimization algorithm,” Soft Comput., vol. 22, no. 1, pp. 1–15, 2018.

[20] S. Masood and P. Chandra, Training neural network with zero weight initialization. 2012.

[21] W. Cao, X. Wang, Z. Ming, and J. Gao, “A review on neural networks with random weights,” Neurocomputing, vol. 275, pp. 278–287, 2018.

[22] E. Talbi, “Population-based metaheuristics,” Metaheuristics from Des. to implementation, John Wiley Sons, Inc., Hoboken, New Jersey, pp. 190–200, 2009.

[23] R. G. Reynolds and B. Peng, “Cultural algorithms: modeling of how cultures learn to solve problems,” in 16th IEEE International Conference on Tools with Artificial Intelligence, 2004, pp. 166–172.

[24] S. khan, I. Qureshi, F. Zaman, B. Shoaib, A. Naveed, and A. Basit, “Correction of Faulty Sensors in Phased Array Radars Using Symmetrical Sensor Failure Technique and Cultural Algorithm with Differential Evolution,” ScientificWorldJournal., vol. 2014, p. 852539, Jan. 2014.

[25] H. Ma and Y. Wang, “Cultural Algorithm Based on Particle Swarm Optimization for Function Optimization,” in 2009 Fifth International Conference on Natural Computation, 2009, vol. 3, pp. 224–228.

[26] The World Bank, “World Development Indicators.” [Online]. Available: https://databank.worldbank.org/source/world-development-indicators.

[27] Expert Group on Energy Data and Analysis (EGEDA), “the APEC Energy Database.” [Online]. Available: https://www.egeda.ewg.apec.org/egeda/database_info/index.html.

[28] D. W. Zimmerman and B. D. Zumbo, “Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks,” J. Exp. Educ., vol. 62, no. 1, pp. 75–86, 1993.

Learning based Coding for Medical Image Compression

Abdul Khader Jilani Saudagar
Information Systems Department, College of Computer and Information Sciences
Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Saudi Arabia

***Abstract*—The area of Image processing has emerged with different coding approaches, and applications which are ranging from fundamental image compression model to high quality applications. The advancement of image processing, has given the advantage of automation in various image coding applications, among which medical image processing is one of the prime area. Medical diagnosis has always remained a time taking and sensitive approach for accurate medical treatment. Towards improving these issues, automation systems have been developed. In the process of automation, the images are processed and passed to a remote processing unit for processing and decision making. It is observed that, images are coded for compression to minimize the processing and computational overhead. However, the issue of compressing data over accuracy always remains a challenge. Thus, for an optimization in image compression, there is a need for compression through the reduction of non-relevant coefficients in medical images. The proposed image compression model helped in developing a coding technique to attain accurate compression by retaining image precision with lower computational overhead in clinical image coding. Towards making the image compression more efficient, this research work introduces an approach of image compression based on learning coding. This research achieves superior results in terms of Compression rate, Encoding time, Decoding time, Total processing time and Peak signal-to-noise ratio (PSNR).**

Keywords—Image compression; medical image processing; neural network; learning based coding; peak signal-to-noise ratio

I. Introduction

Image processing and its related applications has ascended in different levels of coding approach which are stretching from rudimentary image compression model to astronomical data processing and clinical image processing, considered as high end applications. For example in telemedicine, sending and receiving images by overcoming the bandwidth limitation is a major problem faced by the hospitals nowadays. In this situation, an engineer’s main aim is to develop new methods using which the transmission of multiple images with lower bitrate can be made easy. It should also need to achieve a good image quality at receiver side in image processing applications like progressive image encoding, multimedia transmitting, image browsing etc. Since, the images are of huge features, coding without the loss of information into lower bit rates may intern results to the degradation in the quality of image under retrieval. Along with this, encoding in the noise environment becomes too much complex and results in the heavier degradation in the quality of image. Several approaches were proposed in the past for encoding and compression of images, but none of them found to be efficient under specific environments.

All the earlier proposed approaches give efficient results if the systems have high bandwidth and are failed to perform under systems with low bandwidth which becomes one of the limitation. This problem can be overcome if the encoding technique is in a way such that the compressed bit rate will be compatible with low bit rate. To achieve this objective the image coding approaches has to compress the image effectively to the required data rate. Various image processing oriented services [1] requires higher accuracy with high processing data rate. In clinical image processing every image need to be processed for compression before it was streamed to remote place via a channel for further processing. There exists many image compression techniques but computational overhead and the lower retrieval accuracy are two major problems in such type of applications. The proposed work is undertaken to overcome these problems especially in clinical image processing and fill research gap for future researchers.

This research work objective was to develop an effective image coding system with less computational overhead and also achieves increased retrieval accuracy in medical image compression. Towards making the image compression more efficient, this work introduces a modified technique for image compression using neural networks.

II. Literature Review

Image compression is an evolving area in multi-disciplinary applications. This arena is growing exponentially, due to its numerous applications in digital imaging and encoding. Various applications need high effective image compression. In variety of applications, medical image processing is a rapidly evolving area. In case of medical image processing, medical samples are moved from one location to another location through a channel. In such transmissions, the practitioner needs exact information to perceive a perfect diagnosis results.

In earlier, various compression approaches are developed to perform medical image compression. The earlier developed most popular image compression techniques are Embedded Block Coding with Optimized Truncation (EBCOT) [2], Set Partitioning in Hierarchical Trees (SPIHT) [3], Joint Photographic Experts Group 2000 (JPEG-2000) [4], JPEG [5] and lifting scheme based compression techniques [6]. All the earlier compression techniques are classified as lossless and lossy. In lossy compression [7] the probability of accurate information retrieval at receiver side is very less which results in lower PSNR. Basically the faster encoding applications require this type of compression. Lossless compression is preferred in the scenarios where the information loss is not tolerable. Lossless compression method [8] is a scheme that permits the accurate data reconstruction from the compressed data at the receiver side. In [9] a lossless compression scheme is proposed based on the wavelet transform and an adaptive prediction. This compression scheme aimed to achieve maximum compression ratio. A lifting based lossless compression scheme is also proposed in [10] to attain a reduced information loss in the reconstructed image.

However these lossless and lossy image compression techniques become invalid in the case of medical images. In order to achieve a faster encoding and an increased accuracy, the image compression is carried out with the help of Artificial Intelligence (AI) techniques. “Artificial Neural Network (ANN)” [11, 12] is one of the most popular among various AI techniques through which there will be a superior performance in medical image compression when dealing with incomplete or noisy data. Though the ANN based approaches are accurate, the computational overhead is very high. A NN based medical image compression technique was proposed by Lanzarini et.al, [13] in which the compression and decompression ratio is fixed at 8:1 and the loss percentage is 2. This approach created back propagation network for the calculations of correspondent patterns of input and outputs.

Similarly, “Feed Forward Network (FFN)” based back propagation algorithm was proposed in [14]. This approach performs image compression by evaluating the activation values and coupling weights of hidden layer neurons. This method is observed as a better approach compared to the JPEG through the obtained PSNR values.

Recently, many more approaches were proposed by combining the Neural Networks (NN) with various techniques [15-17] to achieve an increased compression ratio. But, none of these approaches has obtained the required optimal compression ratio. In [18], a new image compression is developed by combining the bipolar coding with NN. Here the main purpose of bipolar coding is to achieve maximum similarity between the pixel values of the original and reconstructed image. The decimal values of image are converted into the equivalent binary code words through bipolar coding. This approach achieved an efficient compression ratio along with quality of image. A similar approach is developed in [19] for Genetic Algorithm (GA) with NN. The main focus of this method is on GA through which the small data is classified and mapped.

To achieve a faster encoding, a multilayer perceptron (MLP) algorithm based NN approach is developed by Gaidhane et al., in [20]. In this approach, the information which is below the threshold level is replaced by zero or removed to achieve the faster encoding performance. Thus the quality of image reconstructed at received side is observed to be poor. A similar approach was developed in [21, 23], named as “Vector Quantization” by which the generation of code vectors is takes place using the “self-organizing feature map” concept. Then the block set attached with the code vectors are designed by cubic surface to achieve an efficient perceptual fidelity of the decompressed images. A similar method is proposed by Allaf [22] based on NN for medical image compression. From the obtained results of convergence speed, PSNR and compression ratio, the proposed approach is observed to be optimal approach.

The Region(s) of interest (ROI) [24] methodology is used to achieve high compression ratios. In order to meet the requirements of less storage and minimum encoding time for medical imaging applications [25] and video related applications by preserving the diagnostic features in regions(s) of interest; the concept of heterogeneous (multiple) quality constraints are mostly used and give attractive results.

In the area of clinical image processing, the information loss is not accepted strictly due to the compact coding, because the lossy compression may removes some of the important information required for diagnosis, and also adds an extra artifacts which may give wrong diagnosis results [26]. Thus, for medical image processing applications generally lossless compression is preferred, because it results to more accurate diagnosis. The standard image compression techniques such as Discrete Wavelet Transform (DWT) based compression [2] are even not able to reconstruct the entire image because of the rounding process involved to round of the floating point values into integers.

ANN is a system modeled very loosely on the brain of human. The filed continues with so many names such as neuro computing, connectionism, machine learning system, parallel processing system, natural intelligent processing system and ANN. It is appropriate for special hardware or software in an attempt to simulate the multiple layers of neurons that is the normal processing elements. Connectivity to a variety of modules which represent the strengths of each neuron is linked to some of its neighbors. Neural networks with their extraordinary capability to derive meaning from complex or unclear data, identify patterns realized either humans or other computer techniques used to detect trends that are too complex. A trained NN can be treated as an “expert” who can give correct information based on the information given it for analysis. Many image compression algorithms [27-36] were proposed in the past but given the new circumstances, to provide estimates of the experts, the interest and the questions "what" can be used to answer and hence this research work is undertaken.

A. Advantages to Neural Network Coding

Adaptive learning: An ANN has ability to learn how to perform the tasks given based on the information provided for raining or starting experience.

Self-Organization: An ability to organize the information representation, received while learning process.

Real Time Operation: An ANN has a capability of parallel processing and the specialized hardware devices are manufactured and designed by taking the advantage of capability of ANN.

Fault Tolerance via Redundant Information Coding: ANN has the capability to retain even under the destruction partially or majorly.

This research developed a new learning based coding for proper coefficient selection such that the system accuracy precision will not be reduced. It is observed that in image compression, there are pixels in the image region which are less significant in representation, and the elimination of such coefficient will not affect much to the visual quality. This learning based coding will select such coefficient so as to minimize the number of coding coefficient achieving higher compression rate.

III. Research Methodology

In this work, a new Artificial Intelligence based image compression approach is developed as shown in Fig. 1 using the neural network concept.

A. *Proposed Image Compression Model*

Pre-processing Unit: The pre-processing unit read the medical image sample and acquires the intensities of grey pixels for further processing. These intensities are the output of this unit. The obtained grey intensities of pixels are processed as array for further decomposition unit.

Spectral Decomposition unit: This decomposition unit takes the array of gray intensities as inputs and extracts their multi-resolution features as outputs. These outputs are obtained through a spectral decomposition in a pyramidal fashion. This decomposition is carried out through a recursive process of low pass and high pass filters. The entire process of decomposition is termed as DWT.

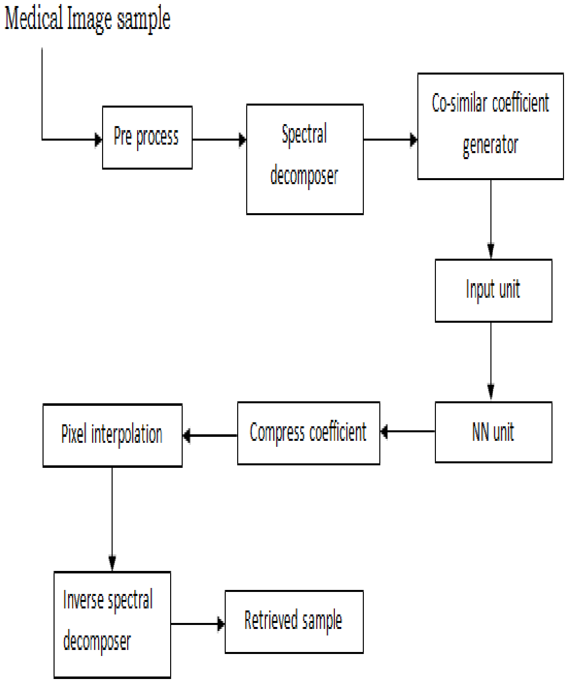


Fig. 1. Step by Step Representation of Proposed Image Compression Model.

Co-similar Coefficient Generator Unit: After obtaining the spectral coefficients from the decomposition unit, they are processed to extract co-similar coefficients. The spectral coefficients which exhibit similar properties are paired. These coefficients are termed as redundant coefficients. This is considered as a first level compression in which there is reduction of redundant information from the image. Further a NN is modeled for the obtained co-similar coefficients.

Input Unit: This unit considers the selected co-similar coefficients as inputs, normalizes them and then passed to neural network. The coefficients are extracted through column wise and then normalize to highest pixel value.

NN Unit: A matlab command ‘newff’ of matlab tool is used to realize this feedforward neural network (FF-NN) unit. This NN unit creates a FF-NN by extracting the min-max value for a given input coefficients through the least average learning algorithm. A sigmoid kernel functions for the creation of this network unit. This unit is created with a network coverage having an error of 0.1 and 50 epochs. The coefficient values are trained through this network and create a FF-NN.

Compress Coefficient unit: Further this compress coefficient unit is created to store the coded coefficients obtained after the feed forward neural network. This buffer is formulated through an array logic in which the coded coefficients are stored for further usage.

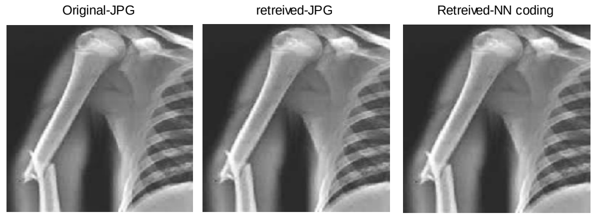
Pixel Interpolation unit: This unit is created to reconstruct the compressed image into its original size through interpolation logic. After obtaining interpolated coefficients, they are rearranged according to their order acquired from the encoder side.

Inverse Spectral Decomposition unit: The interpolated coefficients obtained from the above units are processed back to achieve its multi-resolution information through a successive low pass and high pass filters. The obtained recursive result is given as input for next level to reconstruct further resolution information. This entire process of inverse spectral decomposition is termed as inverse Discrete Wavelet transform (IDWT). The final output of this unit is a decompressed image file.

IV. Experimental Results

This section gives the comprehensive details about the evaluation and performance of the developed compression approach. Medical images of size 512 × 512 pixels have been taken for experimental purpose. Simulation results are obtained using Matlab software along with Neural Network toolbox to encode Medical images of size 512 × 512 at a rate of 0.25 bit per pixel. The experimental results are shown below. Fig. 2 presents a visual comparison for Encoding of given medial test image and the retrieved image under normal conditions.

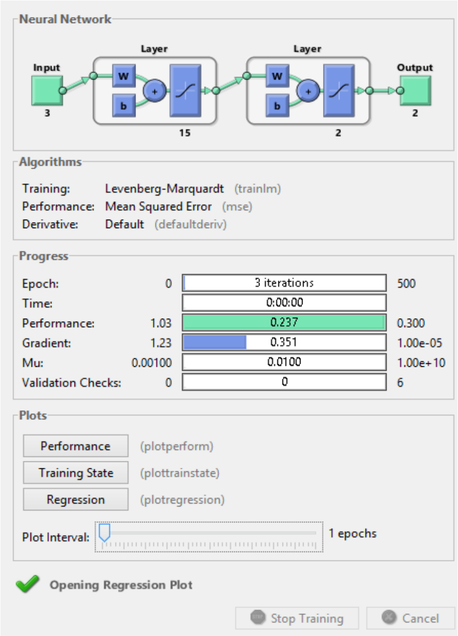
The evaluation and performance of developed NN coding is measured through parameters such Compression Rate (CR), Encoding Time (ET), Decoding Time (DT), Total Time (TT) for processing and Peak Signal to Noise Ratio (PSNR) and compared with the existing JPEG coding.



(a) (b) (c)

Fig. 2. (a) Original Image, (b) Retrieved Image after JPEG Coding (c) Retrieved Image after NN Coding.

The intermediate results obtained during the evaluation of NN coding on the test medical sample is shown in the below Fig. 3, Fig. 4, Fig. 5 and Fig. 6.



Algorithms

Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Derivative: Default (defaultderiv)

Progress

Epoch:	0	3 iterations	500
Time:		0:00:00	
Performance:	1.03	0.237	0.300
Gradient:	1.23	0.351	1.00e-05
Mu:	0.00100	0.0100	1.00e+10
Validation Checks:	0	0	6

Plots

Performance (plotperform)
Training State (plottrainstate)
Regression (plotregression)

Plot Interval: 1 epochs

✓ Opening Regression Plot

Fig. 3. Neural Network Training Tool (NN-Tool).

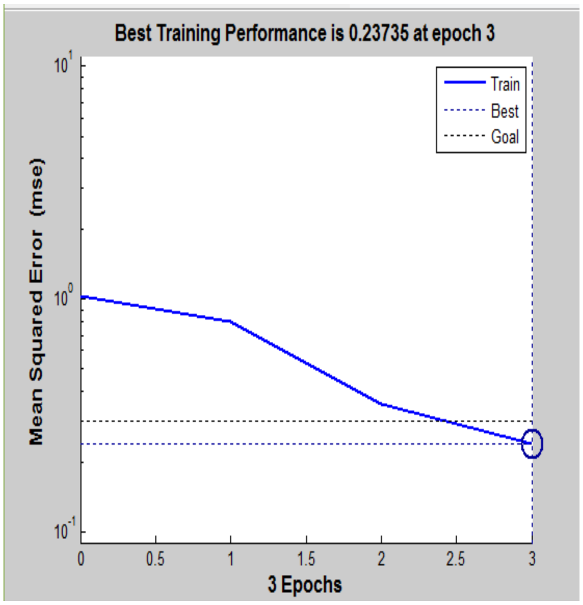


Fig. 4. Best Training Performance Curve.

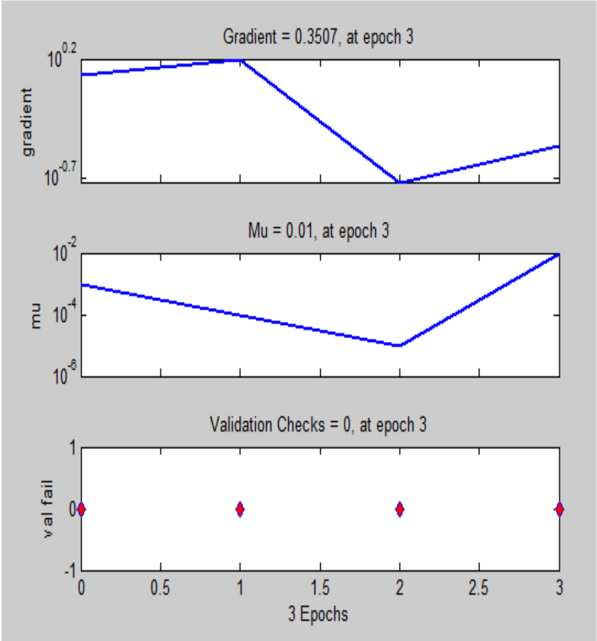


Fig. 5. Training State Curve.

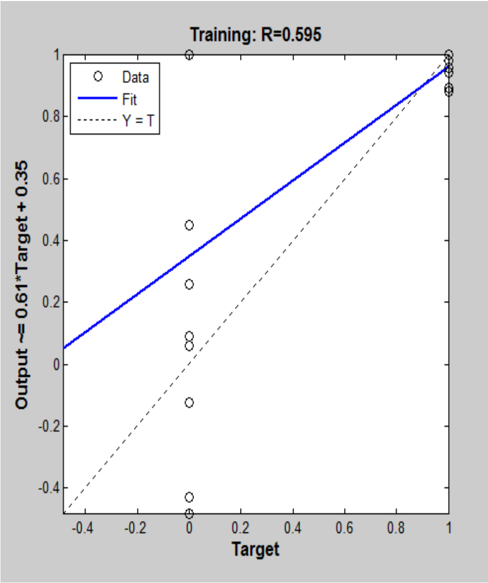


Fig. 6. Regression Curve.

The experimental results for the above test image sample in comparison with the earlier JPEG coding are shown in Table I.

In a similar, manner more samples are given for testing and the obtained results are shown below in Table II.

TABLE I. EVALUATION METRICS

Metric	JPEG-coding	NN-coding
Compression Rate (Bpp)	1.25	2.8234
Encoding Time Taken (sec)	5.125	6.8438
decoding Time Taken (sec)	9.0625	2.5469
Total processing Time Taken (sec)	14.1875	9.3906
PSNR (dB)	40.2530	50.2530

TABLE II. EVALUATION METRICS

Metric	Sample-1		Sample-2		Sample-3	
	JPEG	***NN***	***JPEG***	***NN***	***JPEG***	***NN***
Compression Rate	2.015	3.493	1.203	2.762	1.671	3.250
Encoding Time	5.218	6.812	4.640	7.437	5.250	6.968
Decoding Time	9.406	2.078	7.218	2.046	9.765	2.765
Total Processing Time	14.625	8.890	11.859	9.4844	15.015	9.734
PSNR (dB)	39.886	49.886	42.590	52.590	39.555	49.555

The quality of the sample images is assessed using peak signal to noise ratio (PSNR), mean square error (MSE) and spatial similarity index measure (SSIM). PSNR is usually expressed in terms of the logarithmic decibel scale.

$$PSNR(dB) = 10log_{10}\left(\frac{I_{peak}^2}{MSE}\right) \quad (1)$$

Where Ipeakis the peak values of the input video. MSE is a squared error loss. MSE measures the average squared error between true and estimated values. The mathematical formulation of MSE is given by,

$$MSE = \frac{1}{MXN}\sum(f - \hat{f})^2 \quad (2)$$

Where f - ground truth video

fˆ- interpolated video after extraction

The SSIM is given as,

$$SSIM = \frac{\sum_i \sum_i f(i,j) \otimes f_W(i,j)}{\sum_i \sum_i (f(i,j))^2} \quad (3)$$

For improved imperceptible data quality the similarity factor is closer to 1. The obtained values are presented in the following figures.

Fig. 7 shows the variation of MSE over noise density. The obtained MSE is higher with the increase in noise density. Higher noise variance causes obtained MSE to be high. The developed coding system shows a decrease in MSE due to proper coefficient selection in comparison to JPEG coding. MSE values for noise density remains the same which is of about 0.2 and 10-50% of decrease in MSE is obtained in case of proposed developed approach.

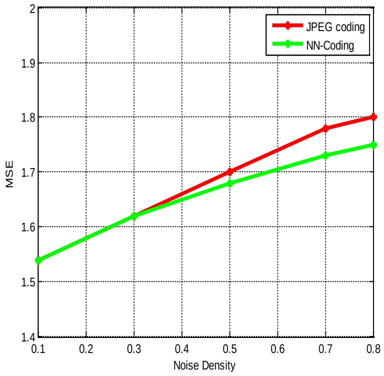


Fig. 7. Experimental Values with Noise Variation for MSE.

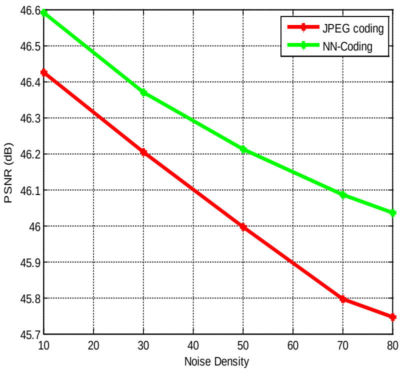


Fig. 8. PSNR Over Variation in Noise Density for Test Images.

Experimental results show that retrieved data quality is degraded due to raise in noise density. This impacts the content and overall quality of the image. PSNR in Fig. 8 shows the quality degradation results. About 2dB of improvement is achieved using the proposed developed approach in comparison with the conventional JPEG coding. SSIM values over varying noise density is computed and presented in Fig. 9 for the given test sample. SSIM for retrieved image using proposed approach is 0.6 in comparison to 0.48 attained using conventional approaches. The experimental metric values for a test image sample are presented in Table III.

A comparable analysis is carried out for variation in learning iteration made. The system is simulated for different offered data rate as a measuring parameter is carried out for various image samples. Fig. 10, Fig. 11 and Fig. 12 shows the experimental observations which are superior due to higher content similarity.

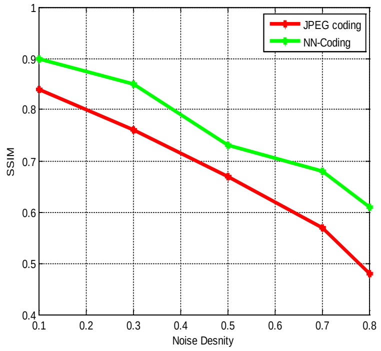


Fig. 9. Experimental Values of SSIM Over Noise Density Variation in Images.

TABLE III. EXPERIMENTAL VALUES FROM THE DEVELOPED SYSTEM OVER JPEG SYSTEM FOR AN IMAGE TEST SAMPLE UNDER DIFFERENT NOISE VALUES

Noise Variance	JPEG Coding System			Proposed Coding System		
	MSE	***PSNR***	***SSIM***	***MSE***	***PSNR***	***SSIM***
0.1	1.56	47.42	0.74	1.56	47.52	0.79
0.3	1.67	45.3	0.71	1.67	45.48	0.72
0.6	1.76	46.9	0.52	1.72	47.13	0.56
0.8	1.83	43.35	0.38	1.79	44.25	0.47

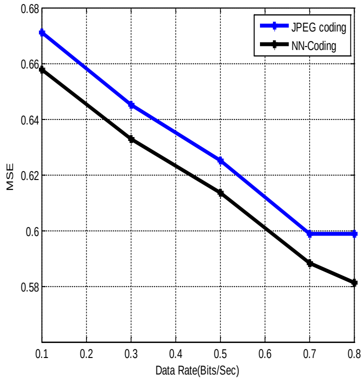


Fig. 10. MSE Observed for Variant Learning Iteration.

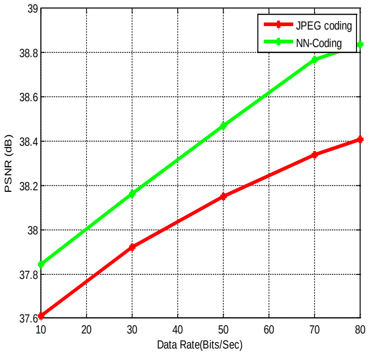


Fig. 11. PSNR Observed for Variant Learning Iteration.

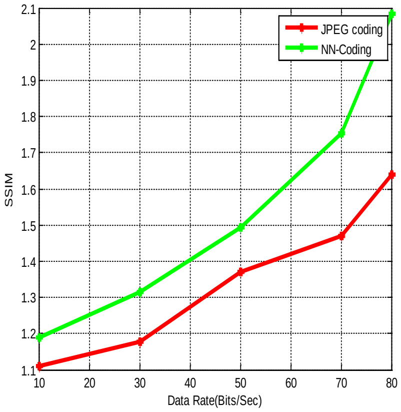


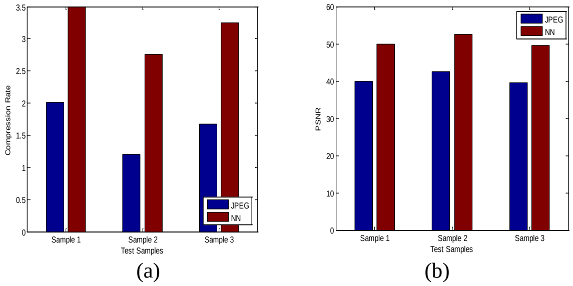
Fig. 12. SSIM Observed for Variant Learning Iteration.

The observed experimental values are presented in Table IV.

TABLE IV. EXPERIMENTAL VALUES FROM THE DEVELOPED SYSTEM OVER JPEG SYSTEM FOR AN IMAGE SAMPLE UNDER DIFFERENT LEARNING ITERATION

Learning Iteration	**JPEG Coding System**			**Proposed Coding System**		
	MSE	***PSNR***	***SSIM***	***MSE***	***PSNR***	***SSIM***
10	0.69	38.71	1.21	0.685	38.94	1.28
30	0.653	36.55	1.15	0.641	37.27	1.42
60	0.527	37.11	1.31	0.49	37.89	1.57
80	0.71	38.0	1.702	0.672	38.41	2.203

Fig. 13 shows the comparative analysis of various test samples for compression rate, PSNR followed by Fig. 14 which presents the encoding time and decoding time.



(a) (b)

Fig. 13. Comparative Analysis (a) Compression Rate, (b) PSNR.

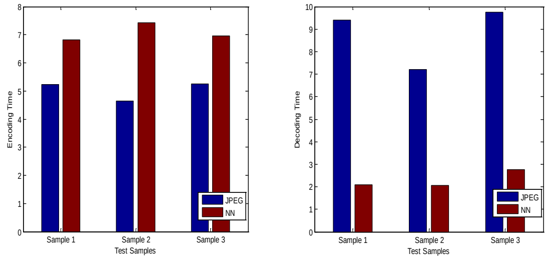


Fig. 14. Comparative Analysis (a) Encoding Rate, (b) Decoding Time.

Fig. 15 shows the total time comparative analysis of various test samples.

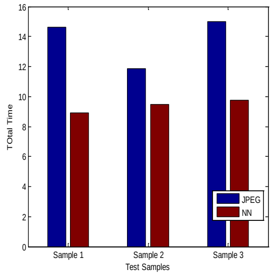


Fig. 15. Total Time Comparative Analysis.

V. CONCLUSION

This work presents an image compression approach for medical image compression using neural network approach. The coding is developed for the image pixel selection, where the learning approach of neural network is used for the selection of significant coefficients. The neighbor pixel value count is used for pixel selection, where the redundant coefficients are used for the coding of compressed data using weight optimization. One of the main limitations faced by the researcher is due to the hardware dependence of neural networks in conducting the experiment. The obtained result for the developed approach is compared with the conventional compression model of JPEG coding, and the observed quality metrics of PSNR and SSIM illustrates an improvement for the compressed data. However in the suggested coding the learning error convergence is observed to be more. This work can be further extended to achieve the objective of optimal rule formation; a new coding of image compression using fuzzy logic is suggested as a future work to this work.

References

[1] R. C. Gonzalez, “Digital Image Processing,” 2nd ed, pp. 343-362, 1992.

[2] D. Taubmann “High Performance Scalable Image Compression with EBCOT”, IEEE Transactions on Image Processing, Vol. 9, No. 7, Jul-2000.

[3] J. Jyotheswar and S. Mahapatra, “Efficient FPGA Implementation of DWT and Modified SPIHT for Lossless Image Compression”, Journal of Systems Architecture, Vol.53, pp.369–378, 2007.

[4] B. E. Usevitch, “A Tutorial on “Modern Lossy Wavelet Image Compression: Foundations of JPEG 2000”, IEEE Signal Processing Magzine, 1053-5888, Sep-2001.

[5] G. K. Wallace, “The JPEG Still Picture Compression Standard”, IEEE Transactions on Consumer Electronics, Vol. 38, No. 1, pp. xvii – xxxiv, 1992.

[6] P.Srikala and S. Umar, “Neural Network Based Image Compression with Lifting Scheme and RLC”, IJRET, Sep-2012.

[7] R. Matsuoka, M. Sone, K. Fukue, K. Cho and H. Shimoda., “Quantitative Analysis of Image Quality of Lossy Compression Images”, International Society of Photogrammetry and Remote Sensing, 2013.

[8] S. G. Chang and G. S. Yovanof “A Simple Block-Based Lossless Image Compression Scheme”, IEEE, 1997.

[9] Y-T Chen, D-C Tseng and P-C Chang “Wavelet-based Medical Image Compression with Adaptive Prediction”, IEEE, 2005.

[10] W Spires, “Lossless Image Compression via the Lifting Scheme”, Nov-2005.

[11] U Seiffert, “ANNIE—Artificial Neural Network-based Image Encoder”, Neuro Computing, Vol. 125, pp. 229-235, 2014.

[12] S. A. Alshehri, “Neural Network Technique for Image Compression “, IET Image Processing, Vol. 10, No. 3, pp. 222 – 226, 2016.

[13] L.L. Laura,A.C.M.T.V. Camacho, A. Badr, I.D.G. Armando and L.I.D Inform tica, “Images Compression for Medical Diagnosis using Neural Networks”, 1990.

[14] W. K. Yeo, David F. W. Yap, T.H. Oh, D.P. Andito, S. L. Kok, Y. H. Ho and M. K. Suaidi, “Grayscale Medical Image Compression using Feed Forward Neural Networks”, International Conference on Computer Applications and Industrial Electronics, pp 633-638, 2011.

[15] M. Liying and K. Khashayar, “Adaptive Constructive Neural Networks Using Hermite Polynomials for Image Compression”, Lecture Notes in Computer Science, Springer-Verlag, Vol. 3497, pp. 713-722, 2005.

[16] B. Karlik, “Medical Image Compression by Using Vector Quantization Neural Network”, ACAD Sciences press in Computer Science, Vol. 16, No. 4, pp. 341-348, 2006.

[17] Y. Zhou., C. Zhang, and Z. Zhang, “Improved Variance-Based Fractal Image Compression Using Neural Networks”, Lecture Notes in Computer Science, Springer-Verlag, Vol. 3972, pp.575-580, 2006.

[18] P. Tripathi “Image Compression Enhancement using Bipolar Coding with LM Algorithm in Artificial Neural Network”, International Journal of Scientific and Research Publications, Vol. 2, No. 8, Aug-2012.

[19] G.G Rajput and M.K. Singh, “Modeling of neural image compression using GA and BP: a comparative approach”, International Journal of Advanced Computer Science and Applications 2011.

[20] V. Gaidhane, V. Singh and M. Kumar, “Image Compression using PCA and Improved Technique with MLP Neural Network”, International Conference on Advances in Recent Technologies in Communication and Computing, 2010.

[21] A. Laha, N. R. Pal, and B. Chanda, “Design of Vector Quantizer for Image Compression Using Self-Organizing Feature Map and Surface Fitting”, IEEE Transactions on Image Processing, Vol. 13, No. 10, Oct-2004.

[22] AL-Allaf, “Improving the Performance of Back propagation Neural Network Algorithm for Image Compression/Decompression System”, Journal of Computer Science Vol. 6, No. 11, pp. 1347-1354, 2010.

[23] C. Karri and U. Jena, “ Fast Vector Quantization using a Bat algorithm for Image Compression”, Engineering Science and Technology, an International Journal, Vol. 19, No. 2, pp. 769-781, Jun-2016.

[24] N. Karimi, S. Samavi, S.M.R. Soroushmehr, S. Shirani and K. Najarian, “Toward Practical Guideline for Design of Image Compression Algorithms for Biomedical Applications”, Expert Systems with Applications, Vol. 56, No.1, pp. 360-367, 2016.

[25] S. Wong, L. Zaremba, D. Gooden, and H. K. Huang, “Radiologic Image Compression –A Review,” IEEE, Vol.83, pp.194- 219, Feb-1995.

[26] V. D. Raut and S. Dholay, “Analyzing Image Compression with Efficient Transforms & Multistage Vector Quantization using Radial basis Function Neural Network”, IEEE International Conference on Engineering and Technology (ICETECH), pp. 1 – 6, 2015.

[27] A.K.J. Saudagar, “A Case Study of Evaluation Factors for Biomedical Images Using Neural Networks”, Advances in Intelligent Systems and Computing, Vol 327, pp. 241-253, 2014.

[28] A.K.J. Saudagar and A.S. Syed, “Image compression approach with ridgelet transformation using modified neuro modeling for biomedical images”, Neural Comput & Applic, Vol. 24, pp. 1725–1734, 2014.

[29] A.K.J. Saudagar and O. AlShathry, “Neural Network Based Image Compression Approach to Improve the Quality of Biomedical Image for Telemedicine”, British Journal of Applied Science & Technology, Vol 4, pp. 510-524, 2014.

[30] A.K.J. Saudagar, “Minimize the Percentage of Noise in Biomedical Images Using Neural Networks”, The Scient. World J., Article ID 757146, 2014.

[31] E.H.S. Ahmed, N. Benamrane and A. taleb-ahmed, “Adaptive Medical Image Compression Based on Lossy and Lossless Embedded Zerotree Methods”, Journal of Information Processing Systems, Vol 13, pp. 40-56, 2017.

[32] B. Fan, “Selective Compression of Medical Images via Intelligent Segmentation and 3D-SPIHT Coding”, Theses and Dissertations, University of Wisconsin-Milwaukee, 2018.

[33] W. Cheng, H. Yifei and W. Weidong, “An End-to-End Deep Learning Image Compression Framework Based on Semantic Analysis”, Applied Sciences, Vol. 9, pp. 3580, 2019.

[34] H. Rafi, “Optimal Compression of Medical Images”, International Journal of Advanced Computer Science and Applications, Vol 10, No.4, pp. 133-140, 2019.

[35] A.S. Sushmit, S.U. Zaman, A.I. Humayun, T. Hasan and M.I.H. Bhuiyan, “X-Ray Image Compression Using Convolutional Recurrent Neural Networks”, pp. 1-4, 2019.

[36] L. Shuai, B. Weiling, Z. Nianyin and W. Shuihua, “A Fast Fractal based Compression for MRI Images”, IEEE Access, pp. 1-1, 2019.

An IoT based Approach for Efficient Home Automation with ThingSpeak

Mubashir Ali[1], Zarsha Nazim[2]
Department of Software Engineering
Lahore Garrison University
Lahore, Pakistan

Waqar Azeem[3], Khadija Javed[6], Maria Tariq[7]
Department of Computer Science
Lahore Garrison University
Lahore, Pakistan

Muhammad Haroon[4]
Department of Computer Science
HITEC University
Taxila, Pakistan

Aamir Hussain[5]
Department of Computer Science
Muhammad Nawaz Shareef University of Agriculture
Multan, Pakistan

***Abstract*—With passage of time, technology is rapidly growing. People and daily life processes are highly dependent on internet. The Internet of Things (IoT) is an area of magnificent impact, growth and potential with the advent and rapid growth of smart homes, smart agriculture, smart cities and smart everything. Internet of Things (IoT) construct an environment in which everything is integrated and digitalized. People depend on smart phones and want to do their daily routine tasks in easy and quick way. Ordinary homes consist of multiple digital appliances that are controlled or managed by individual remote systems. It's very hectic to use multiple individual remotes to control various component of homes. In current technological era, rather than home appliances, almost all type of home components available in digital forms. Various home automation systems with different specifications and implementations were proposed in literature. This research objective is to introduce an IoT based approach for efficient home automation system using Arduino and ThingSpeak. We have automated almost all essential aspects of smart home. Proposed system is efficient in terms of low power consumption, green building and increases the life of digital appliances. ThingSpeak cloud platform is used to integrate the home components; analyze and process the data. State of the art MQTT protocol is implemented for LAN communication. This paper will provide a path to IoT developers and researchers to sense, digitalize and control the homes in perspective of future IoT. Moreover, this work is serving as an instance of how life will be easier with the help of IOT applications.**

Keywords—Internet of Things (IoT); home automation; Arduino; ThingSpeak; sensors; cloud computing; mobile computing

I. Introduction

In current era, technology bring people and things towards adoption of internet. Life dependability on internet is massively increasing. The Internet of Things (IoT) became a domain of high potential, impact and learning [1]. Living cost is increasing day by day. The concentration of researchers is to implicate machinery to reduce this cost of living. IoT brings revolution by automation in agriculture [2], [3], sports [4], health [5], power management [6], industry [7] and assembly modeling [8], [9]. On the other hand, the increase demand of services also requires the data storage and exchange in well-organized way over the internet. IoT improvement has progressed commonly over the most recent couple of years since it has added another estimation to the universe of correspondence and data movements [10]. IoT has done tremendous achievement and everything is going to be more smart and intelligent in next few years so ordinary home system will also move to the platform of IoT [11]. By keeping in mind, the home automation system will allow the users to maintain and build the house that keep power consumption low as well as providing more control over electronic devices [12]. Automated homes will get the benefits of implemented devices and give permission to control it, either user is present or far away [13]. A green building is one that is capable to change according to the environment. It efficiently controls the available resources of building throughout the life cycle from location to design, development and ready to use to, maintenance, redesign to devastation [14]. In closed scope, smart buildings can be considered green buildings because they pursue the same goals as at home. Green building must be economical, ease to use, durable, maintainable and comfortable by requirements [15]. That is why it just not demands a close cooperation between design teams, engineers, architects throughout the project but flexible integration and communication of all home appliances and components. Home automation systems provide comforts for handicap people to use every device without moving. The internet of things has promised to offer the effective way to store and interchange data by connecting high speed networks [16] and electronic sensors with physical devices [17]. The IoT has created the revolution throughout the world and remarkably it has become integral part of life [18]. Home Automation uses several control frameworks to control home machines and tools. With the help of automation in homes, users have more control over homes. NodeMCU, Arduino and other microcontroller are used to make it easy to control home appliances. Multiple sensors like gas sensor, flex sensor, water sensor, temperature sensor, soil moisture sensor, etc. are integrated over microcontrollers to perform specific functionalities [19]. The changing status of sensors will show the real time utilization or variation of system. Appliances status could be seen over cloud platform. Different engineering challenges like Wi-Fi, TCP/IP [20], legacy systems, security and privacy concerns of IoT [21] will be explored before implementation of any IoT based system. This home automation system will provide great insight of embedded systems. Fig. 1 shows the concept of home automation system that how multiple appliances will be connected and controlled. The concept of connecting and monitoring the real home appliances with the help of IoT is discussed in this research paper.

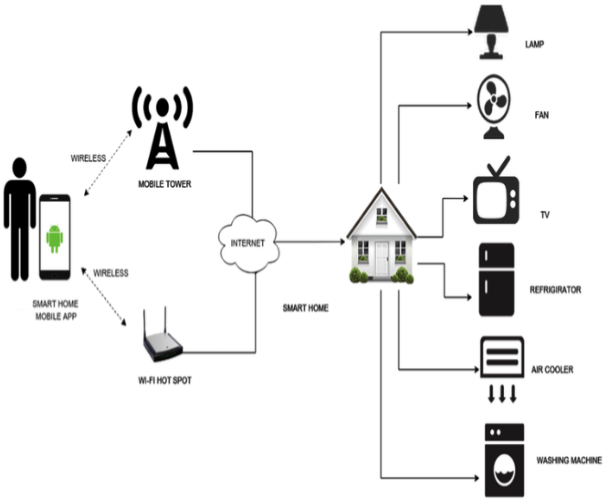


Fig. 1. Concept of Simple Home Automation [18].

This article is divided in to five sections. Section II extensively reviews the literature to read the available implementation of home automation systems and IoT concerns related to automation of homes. Section II also shows the need of this work by highlighting the motivation considered use cases. Section III provides detailed understanding with proposed system design what type of hardware devices, software tools, cloud and networking infrastructure needed to develop this system. Moreover, this section also elaborates experimental setup and environment. Section IV discusses the results with the help of diagrams and charts. Finally, Section V concludes the research by highlighting the contribution and briefs the future direction.

II. Literatuer Review

Various home automation systems were proposed in literature with different specifications and functionality. [22] proposed a home automation by implementing zigbee with Arduino to control the home appliances. This system controls small home appliances by using various till date technological sensors. Users are able to check the status of their home appliances using web server. A web application is designed to control and manage the system. Paper proposed by [23] shows how intelligent home automation is operated and controlled. In this paper the intelligent home automation system with low cost is presented by implementing Arduino UNO microcontroller. There are two main modules that are software communication module and hardware interface module. Arduino UNO microcontroller is used which works as micro web servers and interface of hardware modules and different sensors also used to sense the environment. [24] developed a home portal structure for interconnecting home components with IEEE 1394 AV framework and X10 control line interface with Internet. This gave remote access limits from Web for cutting edge AV mechanical components like Digital Video Camera, Digital VCR related with IEEE 1394 framework and home machines like TV, work zone light, electric fan related with X10 controller. A Java based home automation structure by using World Wide Web [25]. The home devices were controlled from ports of embedded structure board related with PC based server at home. Author in [26], in 2005 proposed Internet based remote control system where home digital devices are related with slave center point. The slave center points talk with expert center point through RF and pro center has successive RS232 interface with PC server. The center points rely upon PIC 16F877μc. This system is controlled by web page application. Author in [27] proposed a framework for controlling home electrical components over the Internet by using Bluetooth remote advancement to give an association from the machine to the Internet and Wireless Application Protocol (WAP) to give a data interface between the Internet and a phone. Another implementation of smart home is proposed by [28] using Arduino mega, Relay, RF module, WIFI module, cloud and mobile application. They are controlling fan and lights by sensing the environmental factors and manually by user. Another energy optimized home automation is proposed by [29] to minimize the energy use in resource limited environment. This system is based upon different digital devices like Multiband antenna, HVAC, Thermal Management, Energy efficient sensors. Energy optimization is elaborated by using different charts and graphs.

Multiple systems for home automation were developed and implemented in literature with different scope from complex to simple systems. A system controlling simple devices like fan or light is known as simple home automation while the system controlling heavy devices like automatic intelligent doors is known as complex system. In many current available systems, mostly fail to cover the basic functionality of home automation. Most of the systems do not provide user friendly environment to control the homes. Appropriate cloud selection is another perspective for secure and real time monitoring even the user is outside from home. So there is a need of an efficient home automation system that deals with above raised concerns in current high-tech era. Here, we are proposing an IoT based home automation system using Arduino with ThingSpeak that address the upraised issues. All other specification of this system, hardware, software or tools, networking architecture, cloud selection, mobile application are elaborated in next section. At initial level, this system implements following seven use cases.

1) Door Lock Control and Monitoring
2) Curtains Control
3) Light Control and Monitoring
4) Fan Control and Monitoring
5) Power Supply Control
6) Fire Control
7) Automatic Water Tank Filling
8) Environmental Parameters

The proposed solution is not just cost effective but also it's easy and reliable when it comes in the terms of implementation and programming. All the hardware is integrated over microcontroller. Other sensors and devices sends data to microcontroller. These all devices and sensors used in this system helped in monitoring and controlling home appliances. This system will provide the real time feedback as user will be able to check what is happening at home.

III. System Design

The proposed system has low cost and efficient monitoring by utilizing IoT based devices. Different modules are used with Arduino UNO microcontroller. The home automation system is offering the features like monitoring the temperature, humidity, fire, gas and water level in tank. It similarly provides the switching functionalities that directs different kind of home appliances and linked with the system used for automation. It is essential need of today's era to improve our life condition. It provides advanced way of life by controlling doors, windows and curtains according to environmental parameters, all home appliances from turning on/off lights, fans to power supply and automatically water tank filling. ThingSpeak cloud is used to provide real time monitoring and controlling. Arduino can perceive surroundings with the help of input signals of different sensors and acts towards surroundings via actuators. Fig. 2 illustrates the working scenario of proposed system. Sensors collect data from home appliances and pass to microcontroller board that directs data to ThingSpeak. A real time notifications and status of different devices shown to user via mobile application. User can easily command and manage the home even from outside the home.

A. Hardware Requirements

Following hardware devices and sensors are used to build the proposed system:

1) Arduino UNO
2) Servomotor
3) Stepper Motor
4) Light Dependent Resister (LDR)
5) Power Relay Board
6) DC Motor
7) ACS712 Current Sensor
8) Magnetic Sensor
9) Flame sensor
10) Water Level Sensor
11) Temperature Sensor
12) DHT11 Humidity Sensor
13) Soil Moisture Sensor
14) Gas Sensor

a) Arduino UNO: Fig. 3 shows the Arduino UNO that is microcontroller with Microchip ATmega328P based upon open source technology [30]. It works as a control board and contains different set of pins for connecting other boards or devices with Arduino. Board contain 6 analog and 14 digital pins and programmed by using Arduino integrated development environment. Inbuilt WIFI facility is available on board for connecting with internet. In our system, all the sensors are integrated over Arduino that sense the data. Arduino transfer the data over ThingSpeak that offers real-time updates for user via mobile application or web interface.

Fig. 4 illustrates the proposed design for Arduino microcontroller that integrates all sensors and other devices on board. As shown in figure, all sensors are directly connected with Arduino that sense the data and pass to Arduino board. The microcontroller collects data from sensors and transfers it to ThingSpeak cloud. Fans, lights and main power supply has high voltage so these devices are connected with a power relay board that control the voltages and pass only bearable voltage to Arduino for operations. Servomotor are connected with Arduino and doors/windows that managed by user to initiate the commands from mobile application. Slide retrofit curtain system is implemented with curtains and movement of curtains managed by user from mobile application. The main reason here to use the Arduino is that, it is low cost in term of price as well as computation and programming. The integration method of sensors and other devices over Arduino is explained further in sub headings.

b) Servomotor: A servomotor shows in Fig. 5 is designed to control the positioning of specific devices. It is integrated with sensor to direct the actuator to precisely control the linear or angular position, acceleration and velocity [31]. It belongs to special class of motors that are used to build the closed loop control systems. It's widely used in automated systems, CNC systems and robotics. In our system, it is implemented on doors and windows to lock and control the movement through Arduino.

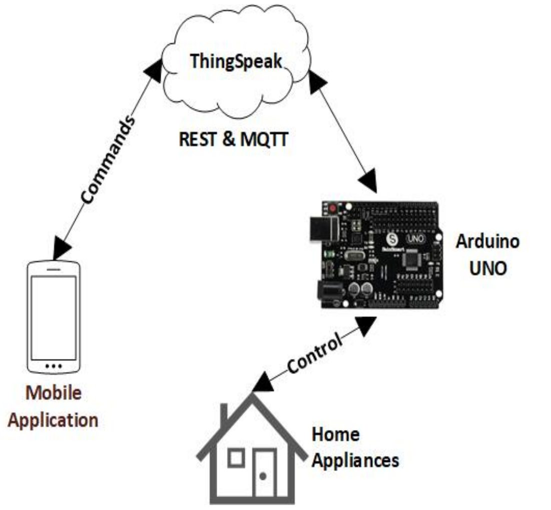


Fig. 2. Proposed System.

Fig. 3. Arduino UNO.

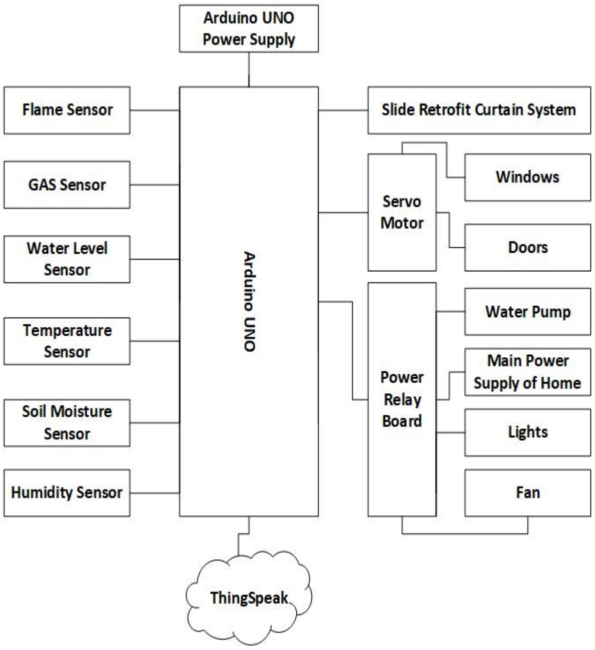


Fig. 4. Proposed Design for Arduino.

Fig. 5. Servomotor.

c) Stepper Motor: Stepper motor is used to control curtain in automatic manner. It enables the automatic opening and closing of curtains according to specific times like morning or evening plus manual control is possible via mobile application. Servomotor can also be used for curtain control but it cannot move the curtains with exact torque.

d) Light Dependent Resister (LDR): Fig. 6 shows the light depended register. LDR is often used in circuits where it's important to identify the existence of light level. In this research paper, we are using LDR to automate the light to control the switches and checking the present condition of appliances.

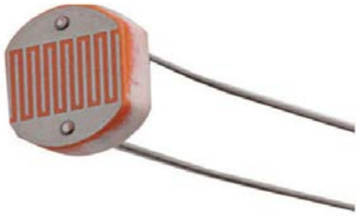


Fig. 6. LDR.

e) Power Relay Board: The most useful thing that can do with Arduino is to control voltage of appliances like light, fans, heaters, AC and others. Arduino operates at 5V and can't control voltage directly but can be done by using 5V relay to switch 120-240V. Some relays use electromagnet to operate automatically the switch nonetheless others use solid state relays. Fig. 7 shows he single channel 5 voltage relay board.

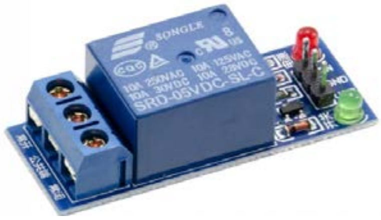


Fig. 7. Single Channel 5V Relay Board.

Relays are used when it's necessary to detach low power signals. When certain event occur relays automatically turn on, for example when temperature gets higher than 25%. Other multiple sensors are integrated over Arduino to operate the system. ACS712 current sensor based upon hall effect principal is used to measure the both direct and alternative current. Magnetic sensor is used to determine the variance in magnetic fields of circuits. Flame sensor is integrated to detect the presence of fire or flame. The response of flame detector is fast and efficient rather than heat detector. Water level sensor is integrated over water tank to measure the level of water. After circuit level, it actuates signal to water motor for ON and OFF. DHT11 is a basic humidity sensor for efficient sensing. It is low cost and does efficient monitoring. It uses capacitive sensor to check out the environmental midair and releases advanced indication on information pin. Soil moisture sensor is used to check the soil parameters of current environment and generate real time results for decision making. Gas sensor is used to detect the presence of specific gases and their level in environment. It is used to check the gas leakage in house.

B. *Tools and Protocols*

Following Software and tools are used to build the proposed sy system,

a) Arduino IDE: Arduino Integrated Development Environment (IDE) is used for functional sensor integration. Flame sensor attached with Arduino with 3 pin input interface and enabled with digitalRead() function that detect the presence of fire. Servomotor is connected with Arduino with 3 pins female connector via 3 jumper wires as shown in circuit diagram. Servomotor can rotate with 180 degree and we have set the rotation value in rotateLoop() user defined function according to our requirement. LDR is integrated with Arduino and analogRead() function is activated to read the current values of LDR sensor. Lights will manually or automatically operated based upon LDR sensor values even if a person is far away from home. Stepper motor is used to control the exact movement of curtains manually or automatically by setting time. motorLoop() function is implemented with conditional structure of time values and manual control. The input values of fan switches are controlled manually via mobile application trigger. The environmental parameters are recorded via environmental sensors like humidity, soil moisture and temperature sensor. Water level sensor is connected with Arduino and place in water tank. It will detect the level of water with analogRead() function and water pump will be started on specific value and off on specific value.

Fig. 8 elaborates the circuit diagram of proposed prototype. It is used to show the actual integration of system that how multiple sensors and components are connected with back bone Arduino. Mostly sensors are operating on 5V so power relays are used to manage the voltage differences.

b) Wireless Sensors Communication: Wireless communication of sensory data is required in IoT systems in effective and secure way [32]. Multiple Wireless Sensor Network protocols are available for data communication and transfer with different security features and other parameters. ThingSpeak communication API is used for sensory data communication. Data is transmitted over ThingSpeak channel in private or public manner. ThingSpeak communication API is based upon REST and MQTT protocol. Typically, the read and write operating time span is 15 seconds.

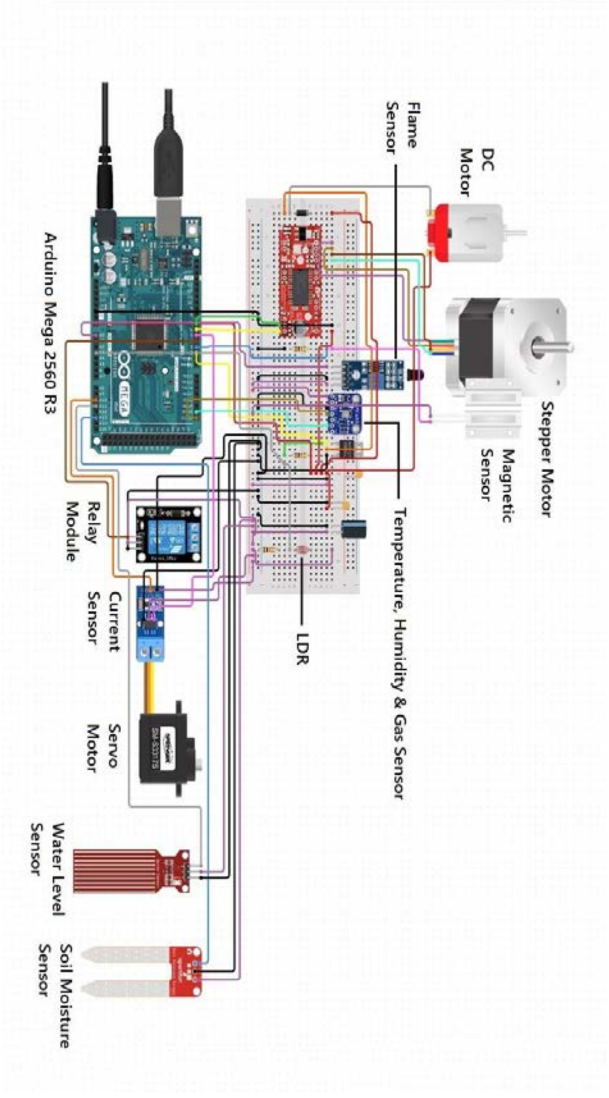


Fig. 8. Circuit Diagram.

c) Ionic Framework: Ionic framework is Cordova and AngularJS based cross platform mobile application development tool [33]. It provides easy cloud integration by implementing cloud APIs. Ionic have many distinguished features like cross platform, JS Components, Angular, Secure, Cordova, Ionic CLI, Elegant Designs, Native Experience, High Performance, Web Components, Interactive Paradigm, Automated Builds, Splash Screens, etc.

d) MATLAB: MATLAB is multipurpose computing integrated development environment with hundreds of scientific libraries. It has various distinguished features of data visualization, plotting functions and comparative graphs. ThingSpeak provides built-in feature of MATLAB for data visualization [34]. MALTAB is used by more than 3 billion users with numerous background of economics, statistics, science and engineering. We have used MATLAB for data analytics and triggers are activated against analyzed data.

C. *ThingSpeak*

ThingSpeak is cloud platform specifically designed for IoT analytic services with wide range of data visualization options. It supports live stream data visualization. MATLAB is integrated with it that makes it highly recommended for IoT systems [35]. It provides easy configuration with channel analytics. It collects the sensed data from IoT systems, preprocess and analyze the data and trigger a reaction according to set instructions. MATLAB helps to build predictive triggers and models to automatically react in certain scenarios.

IV. Results and Discussion

Fig. 9 shows the simulated model of proposed home automation. At initial stage, kitchen and one room is automated along with main door. Gas sensor and Flame Sensor is integrated in kitchen at appropriate location. Water tank is also place in kitchen with water level sensor. Environmental sensors are integrated both in kitchen and room. Stepper motor is only integrated with room curtains and servo motor is fixed at main door. All the sensors and appliances are attached with main Arduino and real time sensed data is transferred over ThingSpeak. Data processing and analytics are performed on cloud and actions are activated as a triggers from mobile application.

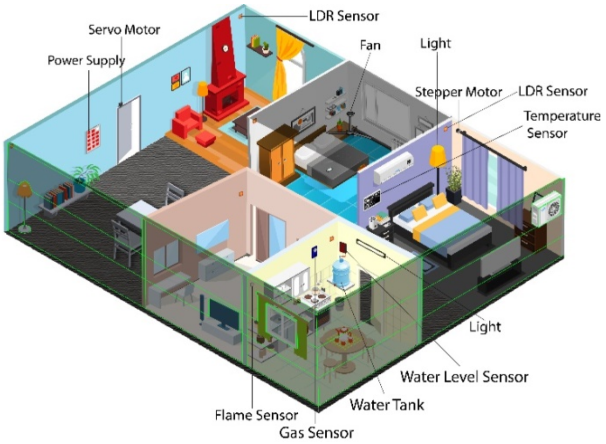


Fig. 9. Simulated Model of Proposed Smart Home.

Fig. 10 shows the implementation of proposed approach. As discussed earlier and elaborated in circuit diagram, all the sensors are integrated with Arduino. The sensed data is transferred over ThingSpeak. Mobile application fetches the real time statistics from cloud. All the triggers are imitated from mobile application. Triggers refer to manual control of lights, fans, curtains and door.

Fig. 11 shows the dashboard of custom designed mobile application for home automation. The first activity of mobile app authenticates the user via channel id and password. Main dashboard activity is appeared after validation of the credentials. Dashboard shows the real time statistics from ThingSpeak. The current status of all appliances and sensors are visible via mobile application. User can manually operate any appliance via mobile application. Furthermore, user will be able to check the current environmental factors of home even if away from home. Main door, room curtains, lights and fans can be operated via mobile application. Water tank option show the current level of water in the tank. Fire alarm is activated if flame sensor detects high intensity flame within range.

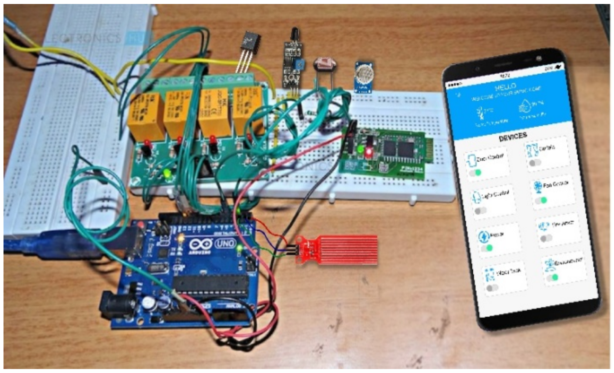
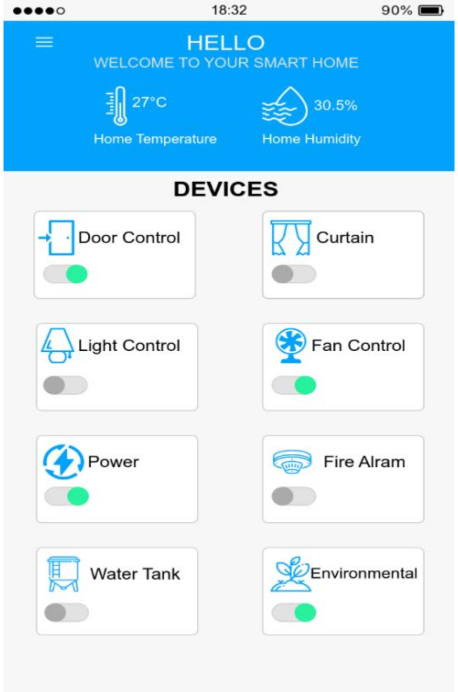


Fig. 10. Smart Home Implementation.



●●●●○ 18:32 90%

HELLO
WELCOME TO YOUR SMART HOME

27°C Home Temperature
30.5% Home Humidity

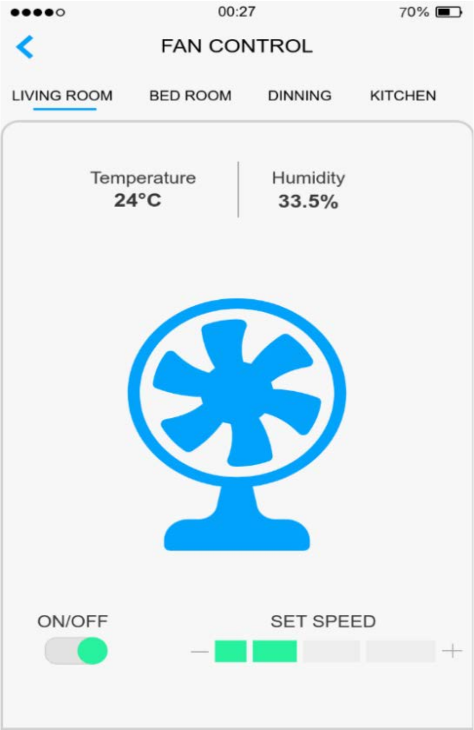
DEVICES

Door Control
Curtain
Light Control
Fan Control
Power
Fire Alram
Water Tank
Environmental

Fig. 11. Dashboard of Home Automation.

Fig. 12 shows the fan control option of proposed home automation system. The fan control option of dashboard leads to this activity. Initially the digital fans of living room and kitchen is automated. Digital fan supports multiple speed levels. User can easily on/off or set speed at multiple levels. Same as light, curtain and door is manually controlled by user.

The graph in Fig. 13 illustrates the profile of environmental factors. Environmental factors include temperature and relative humidity. In graph, x-axis shows the time slots while the y-axis shows temperature and humidity values. Graph shows the comparative temperature and humidity of kitchen and living room. This profile of temperature and humidity is taken form ThingSpeak.



●●●●○ 00:27 70%

FAN CONTROL

LIVING ROOM | BED ROOM | DINNING | KITCHEN

Temperature 24°C
Humidity 33.5%

ON/OFF
SET SPEED

Fig. 12. Fan Control.

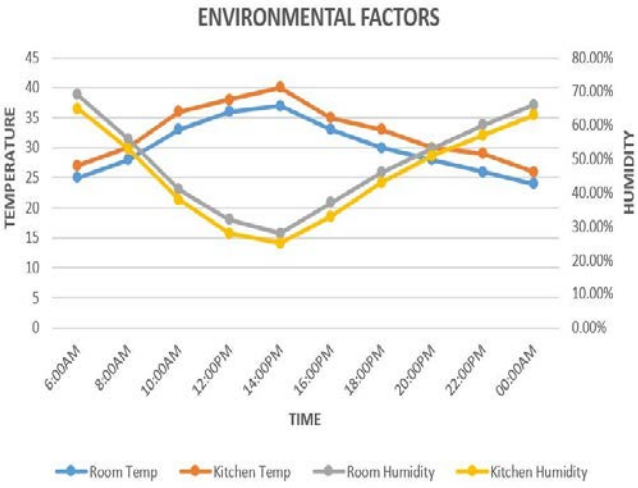


Fig. 13. Temperature and Humidity Profile.

V. Conclusion

This paper provides a state of the art method of home automation with ThingSpeak platform. ThingSpeak provides improved security, data management and data visualization. Wiring and switching cost is reduced by utilizing wireless networks. Power consumption also condensed inside the building when loads condition is off. The sensed data is analyzed at cloud and real time statistics provided via mobile application. A prototype is implemented to elaborate the performance and functionality of proposed approach. Fans, lights, curtains and door are automated. Home appliances can be easily controlled via mobile application. Furthermore, the proposed system provides the real time statistics of environmental factors.

In future, we will improve this system by reducing delay time, adding speech recognition, system automation by history learning and by security features. Furthermore, biosensors will be integrated within home to monitor and control air for better health.

References

[1] J. E. Ibarra-Esquer, F. F. González-Navarro, B. L. Flores-Rios, L. Burtseva, and M. A. Astorga-Vargas, “Tracking the evolution of the internet of things concept across different application domains,” Sensors (Switzerland), vol. 17, no. 6, pp. 1–24, 2017.

[2] A. A. R. Madushanki, M. N. Halgamuge, W. A. H. S. Wirasagoda, and A. Syed, “Adoption of the Internet of Things (IoT) in agriculture and smart farming towards urban greening: A review,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 4, pp. 11–28, 2019.

[3] M. Ali, N. Kanwal, A. Hussain, F. Samiullah, A. Iftikhar, and M. Qamar, “IoT Based Smart Garden Monitoring System using NodeMCU Microcontroller,” Int. J. Adv. Appl. Sci., vol. 7, no. 8, pp. 117–124, 2020.

[4] M. Ali, S. Hafeez, M. K. Paracha, and T. Liaqat, “IOT Based Architecture for Basketball Supervision,” LGU Res. J. Comput. Sci. IT, vol. 3, no. 4, pp. 30–36, 2019.

[5] S. Majumder, T. Mondal, and M. J. Deen, “Wearable sensors for remote health monitoring,” Sensors (Switzerland), vol. 17, no. 1, 2017.

[6] M. Ali and M. K. Paracha, “An IoT Based Approach for Monitoring Solar Power Consumption with ADAFRUIT Cloud,” Int. J. Eng. Appl. Sci. Technol., vol. 4, no. 9, pp. 335–341, 2020.

[7] D. Trotta and P. Garengo, “Industry 4.0 key research topics: A bibliometric review,” in 2018 7th International Conference on Industrial Technology and Management, ICITM 2018, 2018, vol. 2018-January, pp. 113–117.

[8] C. Wang, Z. Bi, and L. Da Xu, “IoT and cloud computing in automation of assembly modeling systems,” IEEE Trans. Ind. Informatics, vol. 10, no. 2, pp. 1426–1434, 2014.

[9] M. K. Paracha, M. Ali, A. Mehmood, and M. Qamar, “IoT Based Approach for Assembly Modeling System with Adafruit Cloud,” Int. J. Multidiscip. Sci. Eng., vol. 11, no. 1, pp. 5–12, 2020.

[10] M. Islam and S. Reza, “The Rise of Big Data and Cloud Computing,” http://www.sciencepublishinggroup.com, vol. 7, no. 2, p. 45, Sep. 2019.

[11] H. Shi, N. Vo, and J. Szajman, “Sensitivity analysis and optimisation to input variables using winGamma and ANN: A case study in automated residential property valuation,” Int. J. Adv. Appl. Sci., vol. 2, no. 12 (Part 2), pp. 19–24, 2015.

[12] M. Asadullah and A. Raza, “An overview of home automation systems,” in 2016 2nd International Conference on Robotics and Artificial Intelligence (ICRAI), 2016, pp. 27–31.

[13] S. Palaniappan, N. Hariharan, N. T Kesh, V. S, and A. Deborah S, “Home Automation Systems - A Study,” Int. J. Comput. Appl., vol. 116, no. 11, pp. 11–18, Apr. 2015.

[14] Z. Jiang, H. R. E.-2009 I. P. & E. Society, and undefined 2009, “Design, modeling and simulation of a green building energy system,” ieeexplore.ieee.org.

[15] C. Jin, … G. D. C. on E. T. and C., and undefined 2011, “Economic analysis of Green building technology based on incremental cost,” ieeexplore.ieee.org.

[16] S. Belhaj and S. Hamad, “Routing protocols from wireless sensor networks to the internet of things: An overview,” Int. J. Adv. Appl. Sci., vol. 5, no. 9, pp. 47–63, Sep. 2018.

[17] J. Shah and B. Mishra, “Customized IoT Enabled Wireless Sensing and Monitoring Platform for Smart Buildings,” Procedia Technol., vol. 23, pp. 256–263, Jan. 2016.

[18] S. G Tzafestas, “Synergy of IoT and AI in Modern Society: The Robotics and Automation Case,” Robot. Autom. Eng. J., vol. 3, no. 5, Sep. 2018.

[19] B. Kim, S. Hong, Y. J.-2008 F. I., and undefined 2008, “The study of applying sensor networks to a smart home,” ieeexplore.ieee.org.

[20] A. A. Alghamdi, “Information security and steganography technique for data embedding using fuzzy inference system,” Int. J. Adv. Appl. Sci., vol. 6, no. 3, pp. 12–16, Mar. 2019.

[21] M. A. Khan and K. Salah, “IoT security: Review, blockchain solutions, and open challenges,” Futur. Gener. Comput. Syst., vol. 82, pp. 395–411, May 2018.

[22] J. Bangali and A. Shaligram, “Design and implementation of security systems for smart home based on GSM technology,” Int. J. Smart Home, vol. 7, no. 6, pp. 201–208, 2013.

[23] T. Chakraborty and S. K. Datta, “Home automation using edge computing and internet of things,” in Proceedings of the International Symposium on Consumer Electronics, ISCE, 2018, pp. 47–49.

[24] T. Saito, I. Tomada, Y. Takabatake, J. Ami, and K. Teramoto, “Home gateway architecture and its implementation,” in 2000 Digest of Technical Papers. International Conference on Consumer Electronics. Nineteenth in the Series (Cat. No.00CH37102), pp. 194–195.

[25] A. R. Al-Ali and M. AL-Rousan, “Java-based home automation system,” IEEE Trans. Consum. Electron., vol. 50, no. 2, pp. 498–504, May 2004.

[26] A. Z. Alkar and U. Buhur, “An internet based wireless home automation system for multifunctional devices,” IEEE Trans. Consum. Electron., vol. 51, no. 4, pp. 1169–1174, Nov. 2005.

[27] N. Sriskanthan, F. Tan, and A. Karande, “Bluetooth based home automation system,” Microprocess. Microsyst., vol. 26, no. 6, pp. 281–289, Aug. 2002.

[28] K. Mandula, R. Parupalli, C. A. S. Murty, E. Magesh, and R. Lunagariya, “Mobile based home automation using Internet of Things(IoT),” in 2015 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2015, pp. 340–343.

[29] L. Salman et al., “Energy efficient IoT-based smart home,” 2016 IEEE 3rd World Forum Internet Things, WF-IoT 2016, pp. 526–529, 2017.

[30] “Arduino - Introduction.” [Online]. Available: https://www.arduino.cc/en/guide/introduction. [Accessed: 20-Sep-2019].

[31] J. Chen, X. Zou, and F. Wang, “Research and Design of DC Servo Motor Position Control System Based on LabView,” in 2010 International Conference on E-Product E-Service and E-Entertainment, 2010, pp. 1–5.

[32] M. Guoe, J. Y. Shan, and I. Yong, “Evaluation of sensor network capability in a practical problem,” Int. J. Adv. Appl. Sci., vol. 14, no. 7, p. 18, 2017.

[33] B. Dunka, E. A. Emmanuel, and D. O. Oyerinde, “Hybrid Mobile Application Based on Ionic Framework Hybrid Mobile Application Based on Ionic Framework Technologies,” Int. J. Recent Adv. Multidiscip. Res., vol. 04, no. January 2018, pp. 3–4, 2017.

[34] “ThingSpeak - MATLAB & Simulink.” [Online]. Available: https://www.mathworks.com/products/thingspeak.html. [Accessed: 08-May-2020].

[35] S. Pasha, “Thingspeak Based Sensing and Monitoring System for IoT with Matlab Analysis,” Int. J. New Technol. Res., vol. 2, no. 6, 2016.

Discrete Cosine Transformation based Image Data Compression Considering Image Restoration

Kohei Arai
Graduate School of Science and Engineering
Saga University, Saga City
Japan

***Abstract*—Discrete Cosine Transformation (DCT) based image data compression considering image restoration is proposed. An image data compression method based on the compression (DCT) featuring an image restoration method is proposed. DCT image compression is widely used and has four major image defects. In order to reduce the noise and distortions, the proposed method expresses a set of parameters for the assumed distortion model based on an image restoration method. The results from the experiment with Landsat TM (Thematic Mapper) data of Saga show a good image compression performance of compression factor and image quality, namely, the proposed method achieved 25% of improvement of the compression factor compared to the existing method of DCT with almost comparable image quality between both methods.**

***Keywords*—Discrete Cosine Transformation; data compression; image restoration; Landsat TM**

I. Introduction

Image data compression methods can be divided into two types, information loss-less and information lossy. The former uses image redundancy for getting high data compression ratio. On the other hand, the later ensures no image degradation by the data compression, data compression is not so high though. JPEG imagery data compression based on DCT is one of the information lossy data compression methods and is popular and widely used data compression method. Although the data compression ratio is satisfactory good, image quality degradation is severe in comparison to the other information lossy data compression methods. Due to the facts, there are block noise, mosquito noise, color distortion noise, etc. in the JPEG data compression method.

Image data compression method proposed here is based on the well-known JPEG compression method. By using image restoration methods, the aforementioned noises are removed as much as we can in the proposed image data compression method. This is the basic idea of the proposed method which allows comparatively high data compression ratio and relatively small image degradation by the data compression.

Research on image restoration is divided into those related to restoration methods and those related to methods for estimating restoration parameters from degraded images. Image restoration methods can be roughly classified into linear restoration filters and nonlinear restoration filters [1]. The former starts with the classic Wiener filter and parametric Wiener filter, which only restore the best approximation image on average, and evaluates the difference between the restored image and the original image not on the space of the original image but on the observed image. A general inverse filter, a least-squares filter with constraints, and a projection filter and a partial projection filter that may be significantly affected by noise in the restored image have been proposed [2]. However, the former is insufficient for optimization of evaluation criteria, etc., and is under study.

On the other hand, the latter is essentially a method for finding a non-linear solution, so it can take only a method based on an iterative method, and various methods based on iterative methods have been tried. There are various iterative methods, but there are a stationary iterative method typified by the successive excess relaxation method (SOR method) and an irregular iterative method typified by the conjugate gradient method [3], [4], [5]. In general, the former requires a large number of iterations, but the accuracy is high, and the latter has excellent convergence, but the problem of accumulation of rounding errors is a problem. When applied to image restoration, it is necessary to pay attention to noise resistance.

On the other hand, the maximum entropy method has been proposed as an image restoration method because it can take into account constraints (or prior knowledge) and resistance to noise [6]. In addition, as a parameter estimation method, methods using stationary iteration methods such as Newton's method and quasi-Newton's method and non-stationary iterative methods such as the conjugate gradient method have already been proposed [7], [8]. Furthermore, an annealing method has been proposed [9].

From the viewpoint of image compression, the estimation of the degradation operator (restoration parameter) by image compression on the transmission side can be performed with high accuracy because the images before and after compression can be referred to. By encoding this restoration parameter and sending it to the receiving side together with the compressed image, the receiving side can restore the image deteriorated by the compression based on the decoded restoration parameter [10]. This is the basis of the compression method with image restoration proposed in this paper.

To show the effect of this method, image compression based on orthogonal expansion is taken as an example here. This paper reports that a high-quality reconstructed image can be obtained by devising the encoding of the reconstructed parameters.

The following section d4escribes research background followed by theoretical background. Then the proposed method is described followed by experiment. After that conclusion is described together with some discussions.

II. Research Background

Facsimile data compression by rearranging picture elements is proposed [11]. Data compression for archiving of Advanced Earth Observing Satellite: ADEOS data is well reported [12]. Method for image compression with a cosmetic restoration, on the other hand, is proposed [13] together with a method for image compression with cosmetic restoration [14].

Meanwhile, a study of data lossy compression using JPEG/DCT and fractal method is conducted and well reported [15]. Also, preliminary study on information lossy and lossless coding of data compression for archiving ADEOS data is conducted and well reported [16].

Method for video data compression based on space and time domain seam carving maintaining original quality when it is replayed is proposed [17]. Data hiding method which robust to run-length data compression based on lifting dyadic wavelet transformation is proposed [18]. On the other hand, method for image portion retrieval and display for comparatively large scale of imagery data onto relatively small size of screen which is suitable to block coding of image data compression is proposed and evaluated [19].

Prediction method of El Nino Southern Oscillation event by means of wavelet based data compression with appropriate support length of base function is proposed and validated with the actual data [20]. Meanwhile, method for data hiding based on Legall 5/2 (Cohen-Daubechies-Feauveau: CDF 5/3) wavelet with data compression and random scanning of secret imagery data is proposed and evaluated effectiveness and efficiency [21].

III. Proposed Method

A. Transmission of Compressed Data and Restoration Parameters

Compression methods that allow image quality degradation are roughly classified into predictive coding, orthogonal transform coding represented by Fourier transform, and approximate coding represented by vector quantization. Of these, the orthogonal transform coding is a compression method that can easily estimate the restoration parameter on the transmission side. In this method, image compression is realized by reducing a large dimension of an order after orthogonal transformation, but image quality is deteriorated because high-order information is lost. Consider that the image quality degradation is restored in an integrated manner. Fig. 1 shows the outline of the method.

In the past, only compressed data obtained by compressing original image data by orthogonal transform coding was transmitted. However, the method proposed this time creates the data necessary to configure the restoration filter on the transmitting side and creates the compressed data. Transmit with.

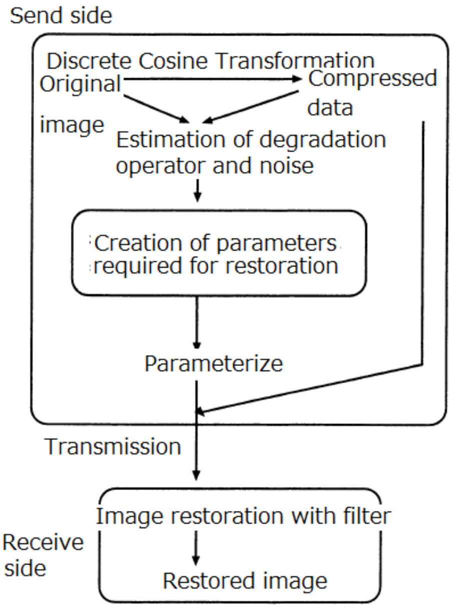


Fig. 1. Process flow of the Proposed Image Data Compression Method.

On the other hand, on the receiving side, a filter is constructed from reconstruction parameters for constructing a transmitted reconstruction filter, and a deteriorated image is restored.

B. Creating a Restoration Filter

The image quality degradation model is represented by a convolution operation as shown in equation (1).

$$g(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x-\xi, y-\eta) f(\xi,\eta) d\xi d\eta + n(x,y) \quad (1)$$

Here, g represents a compressed image, h represents a deterioration operator, f represents an original image, and n represents noise. Also, if each Fourier transform is *G, H, F, N,* then.

$$G(\mu,\nu) = H(\mu,\nu)F(\mu,\nu) + N(\mu,\nu) \quad (2)$$

Can be expressed as If the restoration filter is B (μ, v), the restoration image F (μ, v) can be expressed by the following equation.

$$F (\mu, v) = B (\mu, v) G (\mu, v) \quad (3)$$

It suffices if *G, H, .F,* and *N* are all known in equation (2), but the original image data and the compressed image data exist on the transmitting side, but the restoration filter and noise are unknown there. In equation (2), *G* and *F* are known, but *H* and N are unknown. Since H is difficult to find, we assume *H (μ, v)* = 1 and consider a model where noise causes image quality degradation. Then, the noise *N* is the difference between *G* and *F*.

Then, equation (2) becomes equation (4).

$$G (\mu, v) = F (\mu, v) + N (\mu, v) \quad (4)$$

It can be expressed like this. Wiener filter is considered as an example of the restoration filter.

$$B(\mu,\nu) = \frac{H^*(\mu,\nu)}{|H(\mu,\nu)^2 + S_n(\mu,\nu)/S_f(\mu,\nu)} \quad (5)$$

H (μ, v) = 1, Sn (μ, v) = N (μ, v), Sf (μ, v) = F (μ, nu) (6)

$B(\mu,v) = 1/\{1+N(\mu,v)/F(\mu,v)$ (7)

It becomes, therefore, the following equation,

F (μ, v) = B (μ, v) G (μ, v)

= F (μ, v) G (μ, v)/{ F (μ, v)+ N (μ, v)}

= F (μ, v) G (μ, v)/ G (μ, v)

= *F (μ, v)* (8)

Thus, the original image can be restored. At this time, since the deterioration operator is collected in the frequency spectrum of the S / N ratio, it can be restored on the receiving side only by adding this to the compressed image. The restoring filters other than the Wiener filter have many parameters of the degrading operator and are considered unsuitable for image compression.

C. Restoration Filter Parameterization

As an orthogonal transform, discrete cosine transform is used as an example. There have already been proposals on the spectrum estimation after conversion and the configuration of the restoration filter accompanying it [7]. Here, since the Wiener filter is used as the restoration filter, a new parameter configuration for the restoration filter is devised. The filter is expressed in equation (9).

F (μ, v) =1/{1+ F (μ, v)/F (μ, v)} (9)

Here, when looking at the portion of *N (μ, v) / F (μ, v),* it can be seen that this is the reciprocal of the S / N ratio of the original image and the noise. Therefore, it is first conceivable to parameterize the S / N ratio for transmission. Looking at the S / N ratio in the image, it is as shown in Fig. 2.

Fig. 3 shows the concept of S / N ratio parameterization. In addition, when inverse quantization is performed, the maximum value (MAX) and the minimum value (MIN) of the low frequency component are required, so that these must also be transmitted. After all, what is going to be transmitted is

- SN ratio of quantized low frequency components.
- The maximum value (MAX) used for quantization of low frequency components Minimum value (MIN).
- Coefficient of equation of regression plane showing SN ratio of high frequency component.

In this figure, the difference between Fig. 4 (original image) and Fig. 5 (compressed image) is used as noise, and the S / N ratio is shown for the red image.

It represents the S / N ratio of the higher frequency component toward the center of the image. Also, the blacker the pixel, the lower the S / N ratio, and the whiter the pixel, the higher the S / N ratio.

From these observations, it can be seen that white points and black points are mixed in the low-frequency components, so that values having a considerably large absolute value are mixed, and the values vibrate violently. Conversely, it can be seen that the value of the high frequency component changes relatively smoothly. From this, it is considered difficult to parameterize the low-frequency component, so consider transmitting the low-frequency component as it is, and parameterizing and transmitting only the high-frequency component.

The above is not true for general images. However, it is not necessary to discuss that the image quality is better when the high-frequency component is approximated and sent than when the high-frequency component is deleted after the discrete cosine transform as in the existing method. What can be considered in the parameterization is a polynomial approximation. In the case of polynomial approximation, the degree of the polynomial and the coefficient of each term are transmitted. This can be determined by a method based on regression analysis.

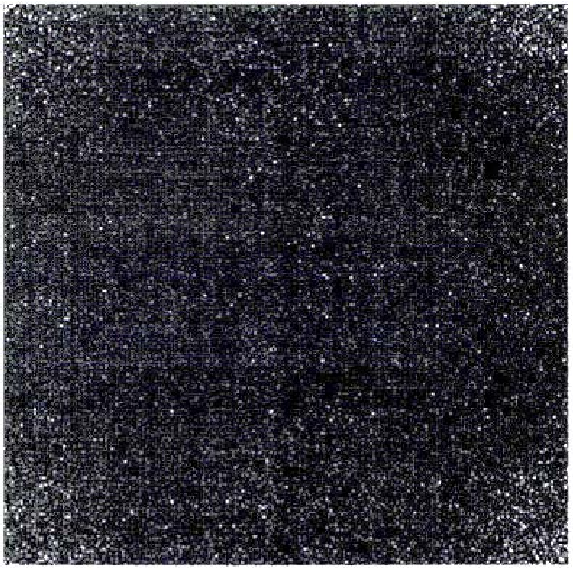


Fig. 2. S / N Ratio in the Frequency Domain.

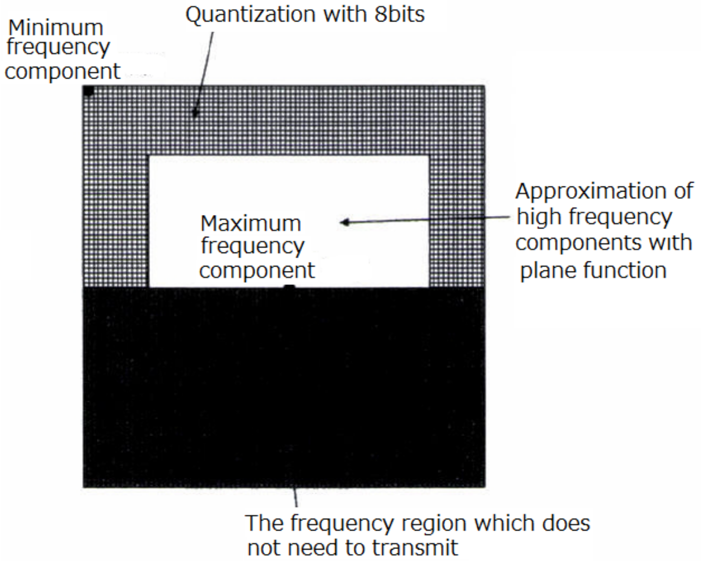
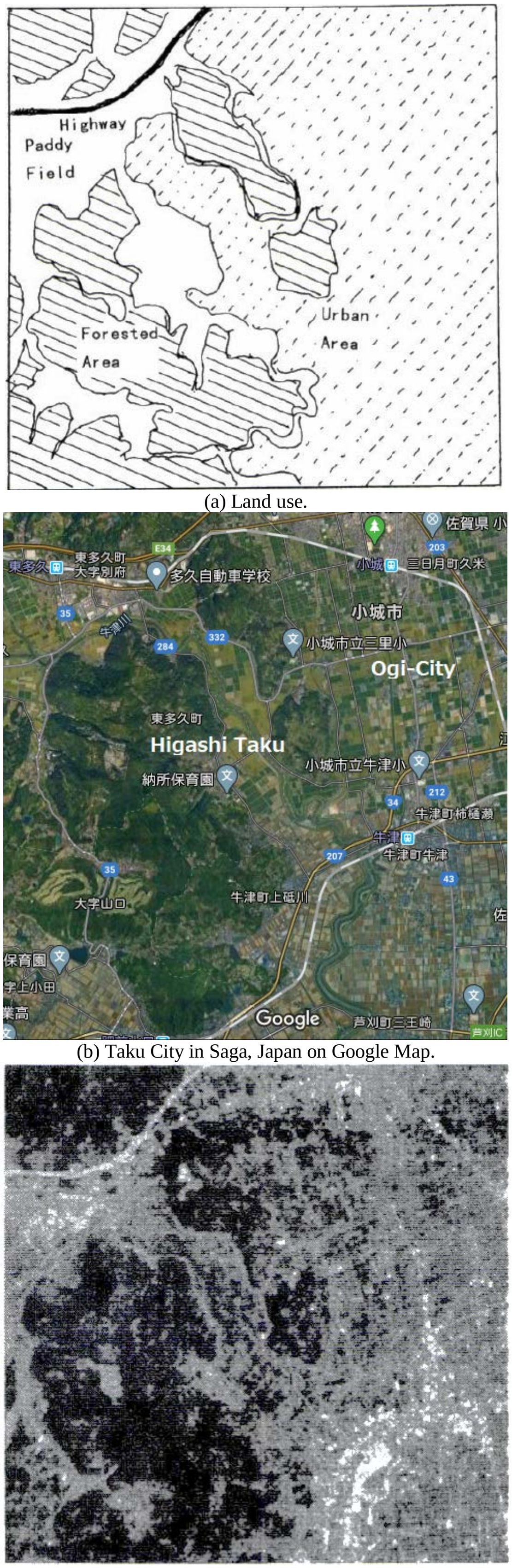


Fig. 3. Coding Method for the Parameters of the S/N Ratio.



(a) Land use.

(b) Taku City in Saga, Japan on Google Map.

(c) Landsat TM Image of Saga Japan.

Fig. 4. Original Image of Taku City in Saga Observed from Landsat.

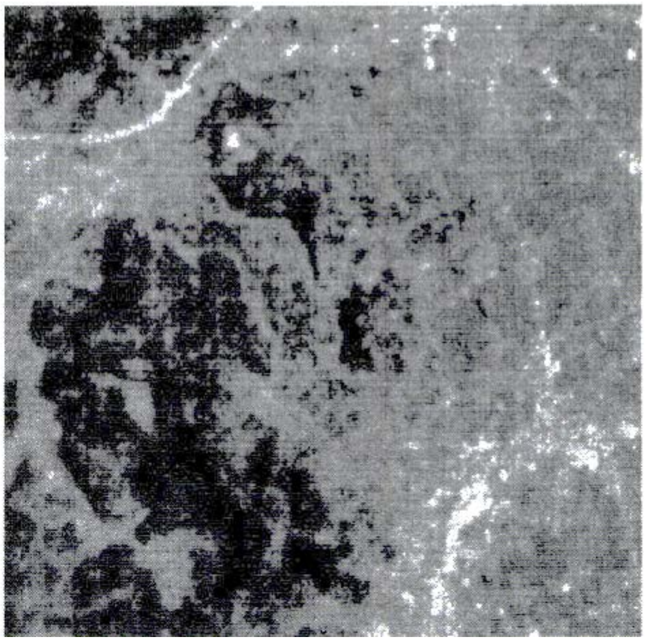


Fig. 5. Compressed Image with DCT Compression (Compression Factor=69).

When finding the regression plane, if data (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) are obtained:

$$y = a_0 + a_1x_1 +\ldots + a_{n-1}x_{n-1} + a_nx_n \quad (10)$$

The sum of squares of the error

$$J = \Sigma \{y_{i-} (a_0 + a_1x_{1i} +\ldots + a_{n-1}x_{(n-1)\,i} + a_nx_{ni})\}^2 \quad (11)$$

These are determined so that is minimized. If both sides are differentiated by a_0, ..., a_n and set to 0, the following normal equation is obtained:

$$\frac{\partial J}{\partial a_0} = -2\sum\{y_i - (a_0 + a_1x_{1i} + \cdots + a_nx_{ni})\} \quad (12)$$

$$\frac{\partial J}{\partial a_1} = -2\sum x_{1i}\{y_i - (a_0 + a_1x_{1i} + \cdots + a_nx_{ni})\} \quad (13)$$

$$\frac{\partial J}{\partial a_n} = -2\sum x_{1n}\{y_i - (a_0 + a_1x_{1i} + \cdots + a_nx_{ni})\} \quad (14)$$

By solving this, the coefficients a_0,..., a_n are obtained, and only the coefficients are transmitted. When actually performing approximation, a three-dimensional plane is considered, so calculation is performed with n = 2. Furthermore, since the S / N ratio of the low frequency component is a floating point number, quantization is performed with 8 bits per element in order to reduce the capacity as much as possible. That is, one element is represented by one byte.

The quantization method first finds the maximum value from the low frequency components, sets it to MAX, and sets the minimum value at the boundary between the low frequency component and the high frequency component to MIN. Next, quantization is performed so that MAX becomes 255 and MIN becomes 0.

Furthermore, since this S / N ratio is a value in the frequency domain, there is a property that the same value appears at a position symmetric with respect to the highest frequency component. In particular, with respect to the imaginary part, a value whose polarity is reversed appears at a position symmetrical with respect to the highest frequency component. By utilizing this property, the capacity can be further reduced by half.

These series of parameterization processes are performed on the transmission side, the low frequency component of the S / N ratio is inversely quantized on the reception side, and for high frequency components, a plane equation is obtained from the transmitted coefficients. It is possible to calculate the actual value.

Fig. 6 shows the result of approximation of Fig. 3.

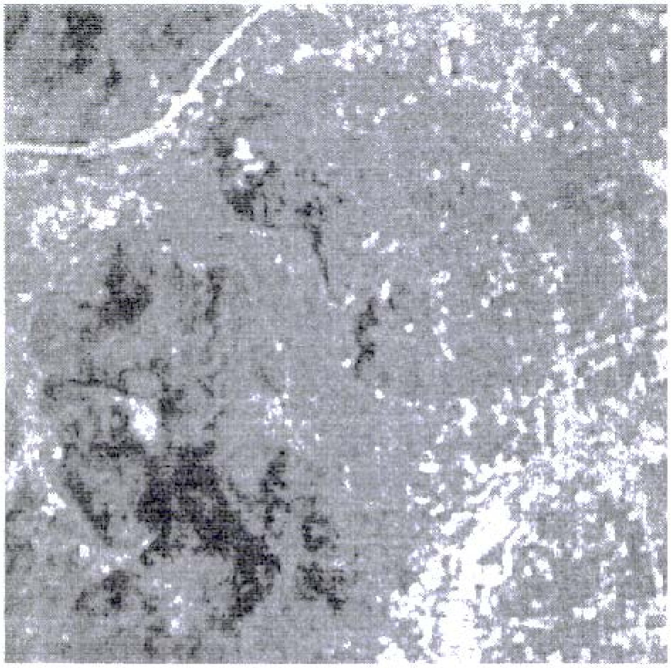


Fig. 6. Restored Image with Approximation of S/N Ratio.

IV. Experiments

A. Original Image

Fig. 4 shows the original image. The sample image used this time is a Landsat / TM image near Ogi, Taku in the western part of Saga city, and is a PPM image using images of blue, green, and red wavelengths. The image size is 256 × 256 pixels. There are various formats of image data, but there are PPM format (color image) and PGM format (black and white image) as uncompressed formats. The luminance value of one pixel is represented by 8 bits for PGM and 24 bits for PPM (8 bits for red, 8 bits for green, and 8 bits for blue). Image data is recorded with 1 byte per pixel for PGM and 3 bytes per pixel for PPM, each with a header of about 15 to 30 bytes.

In PGM, since one pixel is one byte, the number of pixels of an image becomes the capacity of image data almost as it is. If the image size is 512 × 512, the data capacity will be about 262144 bytes, which is a considerable capacity. If this is a color image, the number of bits per pixel will change from 8 bits to 24 bits, so it will triple to 786,432 bytes, further expanding the capacity. If the moving image is a color moving image, a large amount of still images will be included, so that the capacity will be further increased.

Fig. 5 shows an image compressed by the discrete cosine transform. This is a compressed version. The Q factor is specified by an integer value in the range of 0 to 100, with 100 being the best image quality and 0 being the worst. The Q factor = 10 specified here is a considerably high compression ratio.

In this case, the capacity of the original image was 196720 bytes, the capacity of the compressed data was 2869 bytes, and the compression ratio was about 69 times. Figure 6 shows the result of creating and restoring a Wiener filter without creating an SN ratio at all. It can be seen that the details have been restored.

B. Compression with Image Restoration

Fig. 6 shows the results of an attempt to restore the image by constructing a Wiener filter from the approximated S / N ratio. Although it was a slightly blurred image, it was able to be restored in considerable detail.

The image quality after restoration was improved compared to that before restoration, and it was found that restoration of high frequency components was possible to some extent. The data capacity of the S / N ratio is 3816050 bytes without approximation, and has a capacity of about 3.8 Mbytes. In this data format, the SN ratio for each frequency component is represented by a floating point number and output as text data.

As a result of the approximation, the data capacity of the SN ratio was 5053 bytes, and the compression was remarkable. After all, when the compressed image and the decompression parameters (encoded data of S / N ratio) were combined, the capacity became 7922 bytes and the compression ratio became 24.83 times. At this time, it was also found that when trying to obtain the same image quality by compression with discrete cosine transform, the compression ratio could only be obtained about 20 times. This is slightly less than JPEG compressed with a Q factor of 98. Comparing the image quality, it is not worse than the one compressed with Q factor = 98.

V. Some Discussions

The image was compressed at a fairly high compression rate by the discrete cosine transform, and a filter was created to correct the degraded image and restore a good quality image. Then, the data (S / N ratio) necessary to construct the filter was parameterized and considered to be transmitted together with the compressed data.

As a result, it was possible to obtain an image that was somewhat blurry but was quite close to the original image. The original image used this time has a capacity of about 196 Kbytes, the data capacity of the compressed image (Q factor = 10) is 2869 bytes, and the capacity of the restoration filter is about 3.6 Mbytes (red 1.2 bytes, green 1.2 Mbytes, blue 1.2 Mbytes) Bytes), but with the approximation of the SN ratio, the capacity of the restoration filter could be compressed to about 53 Kbytes.

If this is added to the capacity of the compressed image of 2869 bytes, the compression rate will be about 7 times at about 56K bytes, which is slightly less than that of JPEG compressed with Q factor = 98, and the image quality is not worse than that. The Wiener filter created this time was able to completely restore the original image, but even if the S / N ratio for constructing the Wiener filter was approximated, an image close to the original image could be obtained.

The compression ratio was relatively effective, about 25 times, and the effect on the image quality by approximation of the S / N ratio was also small. It was found that if the same image quality was to be obtained by a compression method involving discrete cosine transform, the compression ratio would be about 20 times, and the compression effect would be reduced by 25 %. This is the result of a subjective evaluation experiment of image quality by a one-to-one comparison method based on the Thurston method.

Prepare a compressed image in which the Q factor in JPEG compression is changed in 10 steps from 10 to 90 (the compression ratio changes from about 90 times to 5 times), and compare the compressed image proposed this time with the one-to-one comparison This is the result of evaluating the quality of image quality for 40 subjects. Fig. 7 shows an image with a compression ratio of 20 at this time and an image quality determined to be comparable to that of the proposed method.

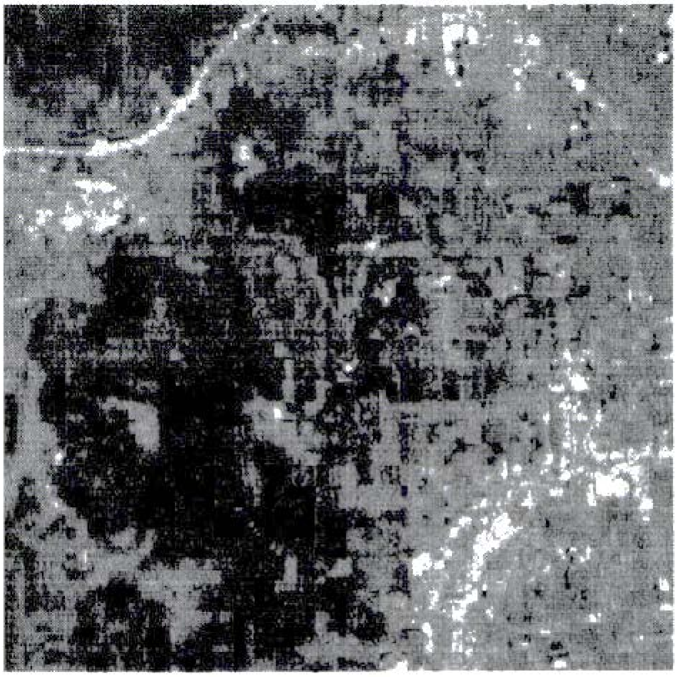


Fig. 7. Compressed Image with JPEG Compression (Compression Factor=69).

VI. CONCLUSION

Discrete Cosine Transformation: DCT based image data compression considering image restoration is proposed. An image data compression method based on the compression (DCT) featuring an image restoration method is proposed. DCT image compression is widely used and has four major image defects. In order to reduce the noise and distortions, the proposed method expresses a set of parameters for the assumed distortion model based on an image restoration method.

The results from the experiment with Landsat TM (Thematic Mapper) data of Saga show a good image compression performance of compression factor and image quality, namely, the proposed method achieved 25% of improvement of the compression factor compared to the existing method of DCT with almost comparable image quality between both methods.

VII. FUTURE RESEARCH WORKS

Further research works are required for the applicability of the proposed data compression method with the other remote sensing images.

ACKNOWLEDGMENT

The author would like to thank Prof. Dr. Hiroshi Okumura and Prof. Dr. Osamu Fukuda of Saga University for their valuable comments and suggestions.

REFERENCES

[1] Ono and Suzuki, "Easy-to-understand JPEG / MPEG2 Implementation", p.1-p.93 Ohmsha 1995.

[2] Supervised by Harashima, `` Image Information Compression '', Ohmsha 1994.

[3] Gregory K. Wallace, gThe JPEG Still Picture Compression Standardh, Submitted in December 1991 for publication in IEEE Transactions on Consumer Electronics.

[4] Co-authored by K.R.Rao / P.Yip, translated by Hiroshi Yasuda and Hiroshi Fujiwara, "Image Coding Technology-DCT and Its International Standards", Ohmsha 1993.

[5] Takagi, Shimoda, and Arai, "Image Analysis Handbook", University of Tokyo Press 1991.

[6] K. Arai, T. Yamasaki and Y. Terayama, A method for image compression with a cosmetic restoration (COSRES Coding), Proceedings of the Australasian Conference on Remote Sensing, Mar., 1994.

[7] K. Arai, Yamazaki, Terayama, "Data Compression with Image Restoration", 1994 Annual Conference of the Institute of Television Engineers of Japan, pp.197-198, July 1994.

[8] N. P. Galatsanos, R. T. Chin, gDigital Restoration of Multichannel Imageh IEEE Trans. on Acoust., Speech and Signal Processing, ASSP-37, pp.415-421, Mar., 1989.

[9] Fujimoto, Fujita, Yoshida, "Reconstruction from Multiple Degraded Images Based on Image Probability Model" IEICE Technical Report, EID97-137, IE97-162, pp.15-22, Feb., 1998.

[10] Masao Ikuzawa, `` Introduction to Statistics for Psychology, '' Minerva Shobo, 1975.

[11] Y,Yasuda and Kohei Arai, Facsimile data compression by rearranging picture elements, Proc.of the 7th Picture Coding Symposium, S405, 45-46,1977.

[12] Kohei Arai, Data compression for archiving of ADEOS data, Proc.of the IGARSS'89, D3-5, 914-918, 1989.

[13] Kohei Arai, T.Yamasaki and Y.Terayama Method for image compression with a cosmetic restoration, Proc.of the Australasian Conference on Remote Sensing, 1993.

[14] Arai,K., T.Yamasaki and Y.Terayama, A method for image compression with cosmetic restoration, Proceedings of the 9th Australasian Conference on Remote Sensing, Mar.1994.

[15] S. Sobue, K.Cho and Kohei Arai, A Study of Data Lossy Compression Using JPEG/DCT and Fractal Method, Proceedings of the ISTS Symposium in Gifu, May 1996.

[16] Kohei Arai, Preliminary Study on Information Lossy and Lossless Coding of Data Compression for Archiving ADEOS Data, IEEE Trans. on Geoscience and Remote Sensing, Vol.28, No.4, pp.732-735, Jul.1990.

[17] Kohei Arai, Method for video data compression based on space and time domain seam carving maintaining original quality when it is replayed, International Journal of Research and Reviews on Computer Science, 2, 4, 1063-1068, 2011.

[18] Kohei Arai, Yuji Yamada, Data hiding method which robust to run-length data compression based on lifting dyadic wavelet transformation, Proceedings of the 11th Asian Symposium on Visualization, ASV-11-08-11, 1-8, 2011.

[19] Kohei Arai, Method for image portion retrieval and display for comparatively large scale of imagery data onto relatively small size of screen which is suitable to block coding of image data compression, International Journal of Advanced Computer Science and Applications, 4, 2, 218-222, 2013.

[20] Kohei Arai, Prediction method of El Nino Southern Oscillation event by means of wavelet based data compression with appropriate support length of base function, International Journal of Advanced Research in Artificial Intelligence, 2, 8, 16-20, 2013.

[21] Kohei Arai, Method for data hiding based on Legall 5/2 (Cohen-Daubechies-Feauveau: CDF 5/3) wavelet with data compression and random scanning of secret imagery data, International Journal of Wavelets Multi Solution and Information Processing, 11, 4, 1-18, B60006 World Scientific Publishing Company, DOI: I01142/SO219691313600060, 1360006-1, 2013.

AUTHOR'S PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 55 books and published 620 journal papers as well as 450 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.html

DVB-T2 Radio Frequency Signal Observation and Parameter Correlation

Bexhet Kamo[1], Elson Agastra[2], Shkelzen Cakaj[3]
Faculty of Information Technology, Polytechnic University of Tirana
Sheshi Nene Tereza 1, 1004, Tirana, Albania

***Abstract*—In this paper, field test measurement are described and statistically correlated to obtain useful information about radiofrequency (RF) behavior of Digital Video Broadcasting - Terrestrial, second generation (DVB-T2) channels. Monitored radiofrequency data parameters are analyzed from statistical perspective and for finding, if any, linear correlation between them. Practical series of field measurements in the surrounding of Korça city in Albania are performed for consecutive 48 hours with sample data each second. The obtained results show the main issues that need to be considered in monitoring service reception quality which is not strongly related to the received channel power level but to the Modulation Error Rate (MER) parameter.**

***Keywords*—DVB-T2; radio coverage; statistical correlation RF data; field measurements**

I. Introduction

Spectrum monitoring and analysis of the provided services and relative radio frequency parameters are of key importance in spectrum use planning and optimization. Spectrum use occurs 24h/day every day, likewise, spectrum monitoring should also be performed on the same continuous or statistically correct basis to ensure that the spectrum is used as intended [1].

Basic idea of this research is to create a base of real RF digital signal measurements and troubleshooting on wireless communications. This is necessary as to complete the theoretical analysis of RF signal propagations on real and complex environments [2][3].

Radiofrequency received signal parameters such SNR (Signal to Noise Ratio) are extensively used for modulation classification, signal recognition and receiver decision on correction factor to apply [4][5].

The main objective of this research is to evaluate, create a test case of a real wireless communication systems and correlate different signal parameters with each other. The purpose of this correlation is to describe the presence or not of sufficient information on RF parameters (such as SNR) to be used for classification and decision making in receiver part.

To validate the proposed approach, the wireless communication system, needs to be stable and continuous in time and not influenced by user usage. In this case, using Terrestrial Digital Video Broadcasting communications, guaranties 24 hours/day continuous signal transmitting with invariant configurations during all the monitoring time.

The VHF and UHF bands are not immune to the effects of anomalous propagation. High atmospheric pressure areas over water can give rise to ducting. Areas of unusually high ionization in the ionosphere are also prone to causing anomalous propagation effects. The result is interference from distant systems, normally considered too distant to warrant great efforts in coordination. These effects are usually transitional and while statistical data on them is available, it is only by monitoring that the implications of these anomalies on wireless systems can be judged. The appropriate interference cure will be case-specific and good monitoring data will greatly aid identifying the causes of the problem.

On DVB-T2 broadcast services and coverage area planning, spectrum monitoring process can perform measurements to check requirements. However, it should be noted that it is not possible to measure the coverage in a given area directly, but it is possible to verify the results predicted by planning tools.

With digital television systems is possible to perform many different type of measurements on either the digital bitstream prior to modulation, or on the modulated signal itself [6]. For verifying predicted service area coverage for DVB-T2 services, the measuring process will monitor modulated signal parameters such as Radio Frequency channel power level; MER (Modulation Error Rate); CNR (Carrier to Noise Ratio); BER pre-LDPC (Bit Error Rate before Low Density Parity Check block). Service providers should clearly define their understanding of coverage meaning as simple as exceeding certain field strength or more complex as end user quality of service perception. In the later case, a more complex and statistical data elaborating is required as to define a better insight of the measurement process and on the quality of service provided.

The contribution of this research is related to the monitoring process of DVB-T2 received signal at a given serviced area. Statistical analysis of the monitored signal parameters and possible correlation for measured data are performed to better understand radio frequency chain behavior. In this case, a real service area (operating in Albania since 2018 [7]) has been monitored for 48 consecutive hours with measurements every 1s for a total 172800 sample data. Performed monitoring is done on UHF channel 57 with 762 MHz central frequency and DVB-T2 modulator parameters as described in Table I.

This paper is organized as follows: Section 2 describes methodology and measurements equipment setup. Section three describes test case analysis, statistical information extraction and confidence level. Detailed analysis and numerical useful data extraction are described in section four. Conclusions and recommendations for future work will be emphasized in section five.

TABLE I. DVB-T2 MODULATRO PARAMETERS FOR UHF CHANNEL 57

TX modulator parameters [6][7][8]	**DVB-T2**
COFDM Mode	32k-ext
Guard Interval	1/32
Carrier Modulation	256 QAM
Pilot pattern	PP4 (TR-ACE)
FEC	3/4
Net bitrate	42.385 Mbps
Required C/N (Rice channel estimation)	20.7 dB

II. OBSERVATION METHOD

Based on the ITU-R standards and recommendations a broadcast radio frequency communication is successful if some minimal quality parameters are guaranteed for certain period of time and locations, defined as percentage of the probability of availability of the communication. For this purposes and based on ITU-R recommendation [8], the monitoring site is configured as shown in Fig. 1 and parameters described below.

- RX antenna positioned on 10m agl (above ground level)
- Three closed locations are considered for each test point to identify receiver maximum channel power
- Measurement mode: channel power
- Channel bandwidth: 8 MHz
- RBW: 40 kHz
- Detector type: r.m.s.
- Sweep time: 1s
- Unit: dBm

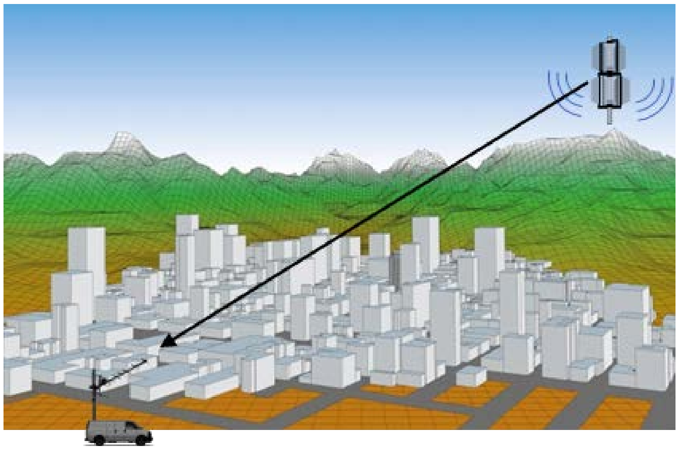


Fig. 1. Sketch of Monitoring Configuration.

The receiving antenna is a calibrated one with 11dBi gain in the given frequency channel. Spectrum/signal analyzer was configured to automatically save sample data every second and transfer them to the attached processing unit for further analysis.

The chosen monitoring site is line of sight with the transmitting station.

III. TEST CASE ANALYSIS

To validate the predicted coverage analysis of the given territory, and to have a better insight on RF system performance for DVB-T2 network, a monitoring campaign is performed in the region of Korça, Albania by the authors [7]. At each test point in the given area, a live measurement of each RF parameter is recorded automatically every 1s for consecutive 48 hours in June 2019, for a total of 172800 sample data for each test site. A visual representation of the recorded data for consecutive 48 hours in one site are shown in Fig. 2.

Form the recorded data, some statistical information as mean value (μ) and relative standard deviation (σ) for each parameter are obtained. However, as recorded data are random variables, if sample changes over time, also evaluated mean and standard deviation will change too. To better present mean μ and standard deviation σ, estimation of this parameters as $\hat{\mu}$ and $\hat{\sigma}$ as statistical variables with a confidence interval (μ_{CI} and σ_{CI}) are calculated based on a percentage of confidence level. So, the expected value and relative standard deviation are evaluated, and results are shown in Table II.

The standard deviation for channel signal power (~ 0.7 dB) is beyond the expected one of 3.3 dB as suggested in ITU-R recommendation [8] [9] for fixed reception.

Due to the nature of the observed parameters and in the scale they are presented (logarithmic), the obtained confidence level can be neglected and is comparable with instrument measuring tolerance.

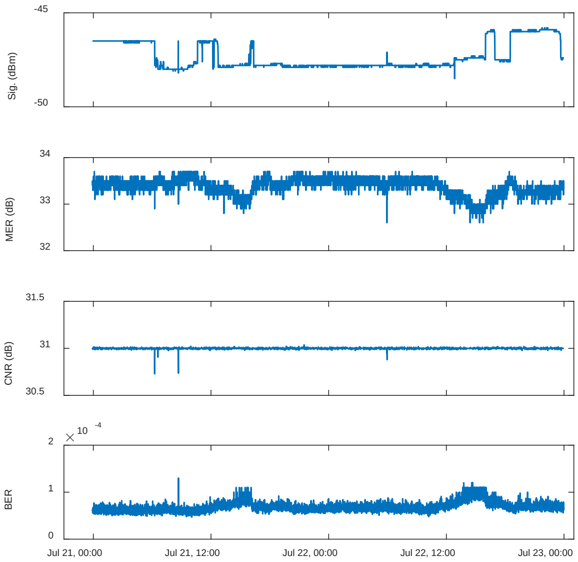


Fig. 2. Time Variation of RF Recorded Data.

TABLE II. EXPECTED VALUE AND CONFIDENCE INTERVAL (CI) WITH 90% CONFIDENCE LEVEL

	$\hat{\mu}$	$\square_{ci}$	$\hat{\sigma}$	$\square_{ci}$
Sig. (dBm)	-47.3173	±0.0029	0.7134	±0.0021
MER (dB)	33.3659	±0.0008	0.2040	±0.0006
SNR (dB)	30.9997	±0.0002	0.0538	±0.0002
BER ($\times10^{-4}$)	0.7253	±0.0021	0.5183	±0.0015

Observed data can be presented as Normalized Probability Distribution Function (PDF) and relative Cumulative Distribution Function (CDF) as shown in Fig. 3. This information is useful to refer observations to the ITU-R coverage parameters and compare to predicted values later.

As the interests is for RF channel power level as minimum received value for which the received signal level is guaranteed on certain percentage of time (ex. 90% of time), the definition of cumulative distribution function we are interested on is the opposite of standard CDF which sees the maximum value of the statistical data. In this case, is convenient to define the cumulative probability distribution function (CPDF) as in equation (1) for RF channel power level, MER and CNR and as CDF for BER:

$$CPDF = \left(1 - CDF\right) \times 100\% \tag{1}$$

The above definition is congruent with the usage of the statistical information describing the physical phenomena. In this case, the worst case is presented for RF channel power as a minimum value, the same is true for MER and CNR and the opposite is true for BER where the worst case is presented with the maximum value.

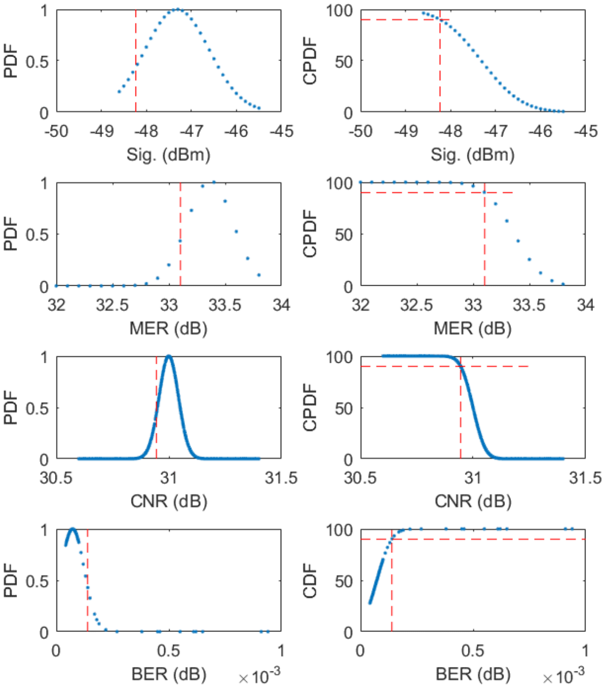


Fig. 3. Normalized Probability Density Function (PDF) and Cumulative Probability Distribution Function (CPDF) for Recorded Data. Dashed Line Highlights 90% Probability Limit as Example.

An alternative view of the same interpretation can be evaluated from Fig. 3 where graphics regarding Signal level; MER and CNR will be read from right to left and the BER from left to right.

In this case is useful to define the coverage as percentage of time (mostly as 50% or 90% of time). So, from the observed data is possible to evaluate RF parameters with the required confidence level as highlighted in Fig. 3. In this case, if the worst-case scenario is requested for 90% of time availability, graphics presented in Fig. 3 are useful to construct data on Table III.

From Table III, 90% of samples (90% of time) have RF channel power level greater than -48.23dBm; MER greater than 33.10dB; SNR greater than 30.93dB and BER lower than 1.39×10^{-4}. The same conclusion can be obtained from mean and standard deviation relative to each parameter, and statistically define the required 90% confidence level. The later can be useful to analytically compute confidence level with less sample data.

TABLE III. MEASURED RF PARAMETERS WITH 90% TIME PROBABILITY

	Relationship	**Limit for 90% of time**
Sig. (dBm)	Grater than	-48.23
MER (dB)	Grater than	33.10
SNR (dB)	Grater than	30.93
BER ($\times10^{-4}$)	Smaller than	1.39

IV. RF DATA CORRELATION

Monitored data analyzed in the previous paragraphs just presents the measure findings but no relationships are still computed between RF parameters.

In this case, statistical parameter evaluations are of critical importance, not only for the design process phase, but also on the live measurements to confirm or correct the predicted behavior on the network design phase [10][11][12][13].

All RF parameters are essential for quality DVB-T2 signal reception and decoding. In this section some statistical manipulation procedures are defined as to better understand the relationship of monitored parameters. As will be discussed later on, an increment of received channel power, not necessary will result in better CNR or MER and less BER.

For this purpose the Pearson correlation coefficient is used as a measure of correlation for two of any couple of random RF parameters. The definition we will use for Pearson correlation coefficient as presented in [14].

For the intent of this material, let A and B be two of any RF parameters, each with N scalar observations, then the Pearson correlation coefficient $\rho(A, B)$ is defined as in equation (2) where μ_A and σ_A are the mean and standard deviation of A, respectively, and μ_B and σ_B are the mean and standard deviation of B.

$$\rho\left(A,B\right) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{\overline{A_i - \mu_A}}{\sigma_A}\right)\left(\frac{B_i - \mu_B}{\sigma_B}\right) \tag{2}$$

From this definition, a correlation coefficient matrix of two random variables which is the matrix of correlation coefficients for each pairwise variable combination is defined as in equation (3).

$$R = \begin{pmatrix} \rho(A,A) & \rho(A,B) \\ \rho(B,A) & \rho(B,B) \end{pmatrix} \qquad (3)$$

Since any of variables are always directly correlated to themselves, the diagonal entries are just 1.

Computing Pearson correlation matrix coefficients as in equation (3) for any of pairwise of monitored RF parameters, results in data presented in Table IV.

Analyzing Table IV, data entry a very weak correlation of channel power level and BER (ρ = -0.0125, close to 0) which means that these variables are uncorrelated, but this does not mean that are independent [14].

To have a better insight of statistical behavior for RF parameters, a correlation behavior is computed and presented in graphic way for each pairwise of RF parameters and shown in Fig. 4. The later permits for visual inspection and relationship and correlation of RF parameters.

From telecommunication theory, is expected to have a strong correlation of signal channel power and carrier to noise ratio, but as can be seen from Fig. 4 it is not met in this case. Also, the same data is reported as Pearson correlation coefficient in Table IV, where the correlation coefficient is too weak (0.0041, close to 0) which means that these two variables are uncorrelated, but this does not mean that are independent.

From visual analysis, some linear correlation of BER with MER are observed, in this case negative correlation as reported in Table IV (-0.6132). The negative value indicates that one variable increases its value, the other decreases its relative value and vice-versa. Also, this is what is expected, as better MER (higher values) will result in less de-modulations errors and in less BER. This is not true for CNR and BER where a correlation coefficient of -0.0077 is observed. The negative sign of Pearson correlation coefficient is coherent with the expectations, but not its absolute value.

For the DVB-T2 modulation parameters used on this implementation, and in the analyzed location, is always possible to decode the DVB-T2 transport stream without any error for the quality of received RF parameters.

TABLE IV. CORRELATION COEFFICIENTS FOR PAIRS OF RF DATA

	Sig.	**MER**	**CNR**	**BER**
Sig.	1	-0.2567	0.0041	-0.0125
MER	-0.2567	1	0.0033	-0.6132
CNR	0.0041	0.0033	1	-0.0077
BER	-0.0125	-0.6132	-0.0077	1

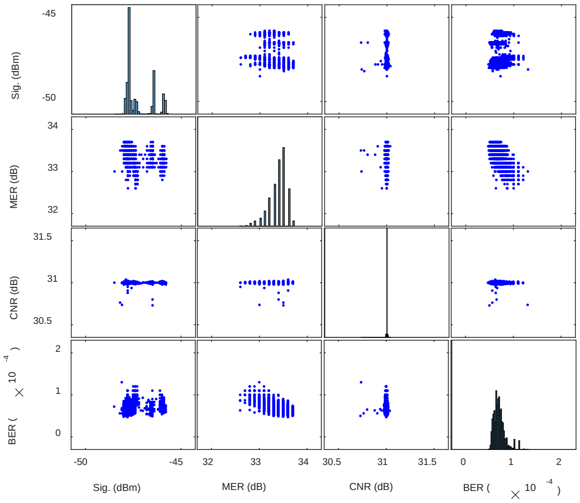


Fig. 4. Visual Correlation of RF Measured Data.

V. CONCLUSIONS AND FUTURE WORK

In this paper is presented radio frequency parameter monitoring for DVB-T2 network. The measuring parameters and relative standard deviations are coherent with that suggested by ITU-R for fixed reception especially the standard deviations of monitored data. Weak correlation of RF parameters are observed which indicates that during monitoring process all the RF parameters need to be temporary recorded and analyzed. Using only few of these parameters for signal/modulation classification are not sufficient. Also, received channel power is not sufficient to characterize the transmitting channel or reception signal quality. For the later, is better evaluating MER as has a higher correlation coefficient with BER and consecutively with reception signal quality from end user perspective.

As future work will be integrating and correlating monitored data for more channel frequencies at the same location and time as to better understand propagation channel behavior and to use this information for channel modeling corrections.

REFERENCES

[1] ITU-R, “Handbook on National Spectrum Management,” 2015 Edition.

[2] Q. Chen, A.L. Gerig, U. Techavipoo, J. A. Zagzebski, T. Varghese, "Correlation of RF signals during angular compounding," IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control, 2005, 52(6), 961–970.

[3] A. P. G. Hoeks, T. G. J. Arts, P. J. Brands, R. S. Reneman, "Comparison of the performance of the RF cross correlation and doppler autocorrelation technique to estimate the mean velocity of simulated ultrasound signals," Ultrasound in Medicine & Biology, 1993, 19(9), 727–740.

[4] A. Hazza, M. Shoaib, S.A. Alshebeili, A. Fahad, "An overview of feature-based methods for digital modulation classification," 1st International Conference on Communications, Signal Processing, and Their Applications (ICCSPA) 2013.

[5] S. Hassanpour, A. M. Pezeshk and F. Behnia, "A robust algorithm based on wavelet transform for recognition of binary digital modulations," 2015 38th International Conference on Telecommunications and Signal Processing (TSP), Prague, 2015, pp. 508-512.

[6] ETSI Technical Report (ETR 290), “Digital Video Broadcasting (DVB); Measurement Guidelines for DVB Systems,” May 1997.

[7] B. Kamo, E. Agastra and S. Cakaj, “DVB-T2 Coverage Area in Albanian Allotments using existing Analog TV Transmitting Antennas,” 2018 26th IEEE International Conference on Software, Telecommunications and Computer Networks (IEEE-SoftCOM), Split, 2018, pp. 1-5.

[8] ITU-R Recommendation SM.1875-2, “DVB-T coverage measurements and verification of planning criterua,” August 2014.

[9] I. Eizmendi, G. Prieto, G. Berjon-Eriz, I. Landa and M. Velez, "Empirical DVB-T2 Thresholds for Fixed Reception," in IEEE Transactions on Broadcasting, vol. 59, no. 2, pp. 306-316, June 2013.

[10] K. Ruščić and A. Skenderović, “Measurements and propagation model tuning in DVB-T2 network,” Proceedings ELMAR-2014, Zadar, 2014, pp. 1-4.

[11] A. Martian, M. Dambeanu, C. Oprea, C. Vladeanu and I. Marghescu, "DVB-T2 radio coverage analysis in Romania," 2017 25th Telecommunication Forum (TELFOR), Belgrade, 2017, pp. 1-4.

[12] M. Slimani et al., "Results of the DVB-T2 Field Trial in Germany," in IEEE Transactions on Broadcasting, vol. 61, no. 2, pp. 177-194, June 2015.

[13] B. Ruckveratham and S. Promwong, "Performance evaluation of DVB-T2 propagation for fixed reception," 2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Chiang Mai, 2016, pp. 1-5.

[14] Joaquim P. Marques de Sá, “Applied Statistics Using SPSS, STATISTICA, MATLAB and R,” 2nd edition, Springer-Verlag, 2007

Causes of Failure in the Implementation and Functioning of Information Systems in Organizations

José Ramón Figueroa-Flores[1], Elizabeth Acosta-Gonzaga[2], Elena Fabiola Ruiz-Ledesma[3]
Instituto Politécnico Nacional
UPIICSA, CDMX
México

***Abstract*—When implementing or starting up an information system, there are usually a number of causes that can lead to its failure. Today, there are few companies that do not rely on technology to carry out their business processes. Wanting to have a competitive advantage over its competitors and the changing global business, puts pressure on the implementation of information systems implementation projects, be it an ERP (Enterprise Resource Planning), a CRM (Customer Relationship Management) or Big Data projects to manage a central repository of all internal and external data that a company can manage. Although it is an illusion for the company to start a project to implement an information system, its failure can lead to its key business processes not being carried out correctly. This article has the purpose of exposing the most common causes when implementing an information system, but also during the operation of the system, which can lead to organizational chaos and to take measures that no company wishes to take. A real case of failure is exposed during the implementation of an information system in an important Mexican company. The research team was allowed to interview general and systems area managers as well as employees. In addition, a survey was carried out among 30 people between managers and heads of department who followed closely on the implementation process of the global operations and technology system within of the company. The most influential factors were a deficient administration, a bad definition of the project and inappropriate consultancy.**

***Keywords*—Information systems; outsourcing; resistance to change; organizational culture; decision making; information systems implementation failures**

I. Introduction

Market globalization and internationalization has risen the competitive pressure on business, which has driven companies to participate in projects that may be critical to their development and even for their survival [1]. These projects, such as the implementation of information technologies for information systems, have one thing in common: they are to be managed, carefully planned, staffed, organized, monitored, controlled and evaluated [2].

Information Systems (IS) are specifically designed to provide a series of benefits to the enterprises and can even become an essential factor for their success by offering a competitive advantage. However, during the implementation or functioning of a computer system, some companies find out that they are not helping them to reach their goals or that they do not have the expected performance, which drives them to make costly changes that can set the company back.

There is a series of causes or factors that can affect the implementation and functioning of an information system. From a technical view, an information system's implementation can be a success, but functionally it can be perceived as a failure. During the implementation phase, the scopes that the information system is to have are to be defined accordingly to the company's objectives; that is to say, if these are not aligned to the company's mission and vision, it is very probable that a series of problems manifest during the functioning phase, leading to failure. A computer system's success or failure in implementation depends on counting with both the adequate information technologies and collaborators, such as a suitable administration during the project's development and implantation.

The objective of this study is to expose the most common causes that affect the implementation and functioning of a computer system, reviewing what other authors say about these causes and complementing them from personal experience, so that they can finally be compared with a real case which happened within a small Mexican company which, for this paper's purposes, shall be called "TVT". The analysis will reveal whether it was the most common causes which lead to the implementation failure of a computer system in TVT, or if there are other causes that must be considered.

II. Causes that Affect the Implementation of an IS

A. Incorrectly Defining the Project

Before beginning, the development or implementation phases of a computer system, the first question that a company must answer is whether it really needs it or not. At times, not knowing what they want the system for is the main cause of its failure. Many failures in a computer system have been attributed to defects in the project's requirements. For example, McKinsey's study about large-scale IT projects informed that the factors associated with causes and requirements were the most common causes for IT project failures [3]. Companies usually define short, medium and long term goals as well as a vision for their future; if the objectives and vision that are being pursued are well defined, they must adjust to the development requirements of the information system so that, at the end of the implementation phase, it will meet the company's expectations.

B. Inadequate Technology

Nowadays there are many information technologies (software, hardware, communications, etc.) that aid in the correct implementation of an information system. Amongst all its variety, one must choose the adequate tool so that one may reach the company's goals. A bad selection of these technologies could lead to failure in a short term since the system's implementation. One of the most typical consequences of a bad choice of technology is the underestimated or overestimated complexity of the tool, which can cause several users not to use it. This situation will be addressed later in greater detail.

With the constant evolution of technology and the creation of new technological tools with new functionalities according to our changing world, some become popular and transcend, while others have short utility spans. The chosen tool must not only be good and in vanguard, but also compatible with the company's goals. A typical mistake is to think that the best tool will give us the best system, but the high prices compared to the benefits are often not justifiable.

In the latest years, Big Data technologies have gained much popularity and importance amongst companies, as they often bring multiple benefits. These technologies are part of a new generation specifically designed to extract the highest data volumes at the lowest cost. They allow for the collection of a wide array of data types, as they allow capture, discovery and/or analysis at high speed [4]. In order to keep up with the vanguard, many companies have started to implement this kind of technologies; however, more often than not they end up never exploiting their full potential. Most of these Big Data implementations can be found in companies that do not need to manage high data volumes nor do they have large sources or do not require high speed processes, and the only thing they are doing is migrating their actual intelligence processes to a new technology. Even if Big Data technologies are less costly for their distribution and support business model (compared to business intelligence, which stem from a licencing model), the implementation can be excessively expensive due to the high prices of consultancy and collaborators that are involved during this phase. This, added up with the little profit that these companies can get out of such a system, the benefits will be small and probably similar to those obtained with tools that were already giving them their desired results.

C. Inappropriate Consultancy

It is precisely an inappropriate consultancy that can lead to an incorrect choice of technology. It is very common for companies to acquire and pay for services of third companies to be in charge of the implementation of IT, often called systems consulting or IT support. These organizations often lack enough staff to take care of it and are also often reluctant to hire large quantities of employees for projects that might not be a part of their usual business processes. An IT consultancy company provides the required advice in order to increase the success probabilities in the implementation of a system and in taking advantage of its technology. Most times, however, IT consultancy services do not grasp the functional and non-functional requirements in the analysis phase, leading to them not being portrayed in the final system's general functionality. Added to this, IT consultants, taking advantage of technological fads and the little experience of some companies, are likely to offer their customers expensive systems that they do not need.

To make sure that one has selected the correct consultancy services and technologies for their implementation in an information system, several concept tests must be carried out on different IT companies and tools in order to know which ones adjust better the project's definitions. It is also ideal to conform a work team with highly qualified staff, so the management levels in the organizations must be aware of all consultants that the IT company is to enter into their project.

Oftentimes, it is also important that the IT consultancy company knows how to integrate the technologies inside the organization with swift development methodology instead of the traditional cycle. In the traditional cycle, organizations often become desperate to see tangible results, and, according to their needs, after large timespans, which can lead to frustration and the cancellation of the implementations. In the last decade, software development has been characterized by two main approaches: the development of agile software, which has the objective of achieving a greater speed and flexibility during the development process, and user-oriented design, that places the final user's needs and objectives at the core of the software development centre so to deliver software with adequate usability [5]. Even if the benefits of implementing agile methodologies and work schemes are known, it is true that a poor execution of these can be counterproductive, but that is a topic for another study.

D. Inadequate and Incomplete Training

Suring the implementation of an information system, it is important that the collaborators who will be the system's final users, be it an ERP or a CRM, are trained to operate it correctly. It is a common practice to dedicate several hours on this effort so there are no questions referring to the functioning of some task or module. Oftentimes, fearing that the collaborators might neglect key activities at the company where the system is to be implemented, their immediate bosses and management assign less time than recommended to the training, which can result in the collaborators not employing the system adequately.

It is also important to point out members of the staff that are most fit to operate the system. Training, as it is, rarely produces competent employees [6]. The collaborators must have the necessary skills to understand the system's functioning; otherwise, no amount of training they can be subjected to will make them able to fully take advantage of the system's functionalities.

E. Resistance to Change

The user's resistance to the implementation of new information systems has been identified as an important cause of failure for new systems, and this must be understood and managed. Historically, information systems implementation projects have been plagued by failures for which user resistance has been identified as the main obstacle. The user's resistance is the first challenge for the implementation of a new information system on the greater scale [7]. Among the factors that generate said resistance are fear of the unknown, of failure, of losing authority inside the company, of not being able to learn the new abilities and knowledge that are required during training, and fear of the incorporation of a new talent that might prove superior. Other causes are the lack of information about the project and questioning of the workload that the usage of a new system might incur.

To face all these causes of resistance, there must be a multidisciplinary team within the project, and workers must be able to contribute with ideas about the processes and activities, organizing periodic follow-up meetings, and keeping a positive attitude and open communication within the team. The project's importance must be set clear along with the advantages that will be obtained with its use, such as a positive change in the company's technological evolution.

F. Uneficient Management

The change initiative within an organization so to set an information system implementation project into motion comes from high management. A good manager or systems director must have enough skills to use and maintain a specific kind of technology that will aid the company's business processes [8] The project's planning must be aligned with the company's goals for the functioning of a system, and, from there, the high management team must be able to choose the correct elements in terms of resources, time, help, and technology to carry out a successful implementation. Experienced strategic partners should be put in leadership positions so to make the right decisions according to the defined objectives, and not to personal interests. High management problems usually surface when they fail to notice that the implementation of a system equates to changes in the business. Oftentimes they are not committed to change, don´t define clear business goals and incur in unhealthy practices such as nepotism.

III. Causes that Affect the Functioning of an IS

A. Lack of Commitment at the Management Level

In the strategic process, which includes formulation, execution and control of the strategies in the company, the characteristics of the management style can be appreciated. Managers, in most cases, imply chances in the company's organization, due to a strategy that strives towards approaching new institutions that the company needs to adapt to. While facing said changes, resistance to change is likely to arise both in the individual and organizational levels [9]. It is important to assume that resistance to change is going to be present during the functioning of an IS, but if high management does not set an example by adapting to the upcoming changes, lower-level collaborators are not likely to do it. An information system that produces reports or graphics of the company's situation must be used by high management with the confidence that the information is real, once that the project has been concluded, tested, and demonstrated to be working to perfection.

B. Lack of Performance Indicators

An indicator is a piece of data or a set of data that help measure the evolution of a management system. Indicators are means of evaluating to what extent the strategic objectives are being met. They are useful for they produce information that helps in analysing the performance and detecting deviations in the meeting of objectives. There are several kinds of indicators, such as fulfilment, evaluation, efficiency, effectiveness, and management.

The results of measured performance indicators can be used not only to enhance processes, products, manufacturing, app programming, staff, activities, etc., but also to advise decisions in company management [10]. If an IS does not count with performance indicators, it will be hard to have a it clear whether the system is meeting the objectives that were defined during the planning of the implementation project. Having performance indicators helps the company to decide if it is profitable to continue with the actual functioning of the system, or whether the strategies have to be rethought and changes are to be made to adequate to a new operation.

C. Lack of Change in Organizational Culture

Organizational culture is the set of values that the collaborators of a company share. These values persist through time and can be noted through behaviour patterns, signs, symbols, idioms and other forms of behaviour [11]. A company's organizational culture can be affected during the functioning of an IS, but for good. A culture of use of technology in these times helps to the adequate implementation of a system [12]. An organization that is used to manual processes without technology as a part of their day-to-day routine will hardly accept the integration of a new system to assist in their everyday activities, even if it will allow them to save time to dedicate to other activities that will probably generate more value. The change of cultural paradigm must be synched with the accepting of an IT system, as it helps create cohesion between the company and the system, thus tracing the right path for things to turn out in the best possible way [12].

Managers must be aware that culture at the organizational level can strongly influence the adoption of an IS, and not all aspects of the culture can be controlled in their totality; on the other hand, culture is always on the making (it cannot be spontaneously created) and it forms and reforms itself through social relationships [13].

IV. Methodology

A research was carried out directly in the Mexican company TVT, in which the research team was allowed to interview general and systems area managers as well as employees to find out their thoughts about the failure of there is. In addition, a survey was carried out among 30 people between managers and heads of department who followed closely on the implementation process of the global operations and technology system within TVT. A third company studio also provided data that helped measure the negative impact that the implementation of an IS had within TVT.

V. Case of Study

A. About the Information System that was Attempted to Implement

A technological evolution project was carried out within the Mexican company "TVT", which involved all its business operations areas from their key business processes to their support ones. The project consisted in various phases in which several modules were to be released according to an already planned strategy. These modules included the incorporation of a corporative-level CRM, the integration of all their legacy systems into a functional one, business intelligence projects that included area reports for management and strategy, new sale interfaces in their online portal, portal services for speeding up customer service and a corporative Data Warehouse whose model had been acquired with a well-known IT company, among others. The project started on 2010 and was expected to fully conclude on 2014. However, after years of delays, the project was officially concluded on 2017, unsuccessfully. TVT invested approximately 10 million Mexican pesos throughout the implementation of the global project, from which, only 29% of the modules were released to the productive environment. In Fig. 1 the relationship per year (from 2011 to 2014, years of start and finish according to the global project plan) one can observe the modules that were released versus the ones that were supposed to be released. In all the years, the number of releases was always below planned. The project's financial losses are estimated in more than 4 million Mexican pesos.

B. *Causes that Lead to Failure*

When TVT noticed that it was unfeasible to continue to support a project that demanded too much money and offered little to the company's interests, it decided to cancel the project, which caused for many people to be unemployed including high command people which many employees point as the main cause of the failure. To carry out the global project, TVT hired the services of a leading consultancy company in their IT area, as well as hundreds of internal employees that usually stayed for short periods within the company.

After the failure, TVT hired the services of a leading statistics company to find the causes for their financial loss. Even if some of the modules were released and are still operative within TVT, many of which were thought to be put in production and were to be used for strategic processes within the company weren´t.

The statistics company that TVT hired to find the causes for their global project's failure identified the following causes along the interviews that it applied to TVT's employees:

- Bad planning from the systems director: Their systems director had no control over the planning of the global project, resulting in him not defining development strategies that would lead to the delivery of partial results.
- Nepotism within management positions and subdirection: The high commands recruited friends and family members that were not sufficiently qualified for their posts. The collaborators that were capable were often relegated to the background and not taken into account for the planning.
- Construction of projects or modules that depended on third-party conclusion: Consistently with the management's bad planning, many projects that depended on the conclusion of other modules were started before their previous stages had even concluded their analysis phases.
- Bad analysis of requirements: A lot of time was invested in the analysis phase of each project, but the analysts did not understand TVT's business. When the SCRUM agile framework was enabled, the time in the analysis phase was reduced, but the functional requirements were not being met. There was no follow-up to the agile framework; each team did what they understood for methodology, there was no training or involvement on the managers' part for their teams to carry it out correctly.
- Excessive salary of the systems employees: For many collaborators within TVT, the systems employees' salaries were too high and this generated envy, which resulted in the rest of the staff refusing to attend to the systems employees, thinking that if they made a higher amount of money it was because they could do the same work without business context. Here the principal cause was that there was no control over salary information, and it generated jealousy among the employees.
- Corruption in the purchasing of technology: It was rumoured, and basically of general knowledge, that de acquisition of technology in tenders and concept tests were stained with corruption that the providing companies generated towards the systems directives. It is spoken of expensive gifts and trips abroad.

At the end of the global technology and operation project, the systems director was removed from his charge and transferred to another company from the same group TVT belonged to. The subdirectors and some other managers were fired, and new collaborators were hired for their positions to try and reorganize the project's course, but none of them found a way of salvaging it. Thus, the project was terminated, and a new global strategy project that included new technology and a different focus was started in its stead.

In the 30 interviews that were carried out for this paper, the employees often mentioned that the sensation of failure and little functionality of the global project were constantly present over the years that it lasted, and that it worsened over time. Many employees within TVR did not trust the course the project was taking after the second year, and many began questioning whether their collaboration was truly useful or if it was even going to bring any benefits to TVT.

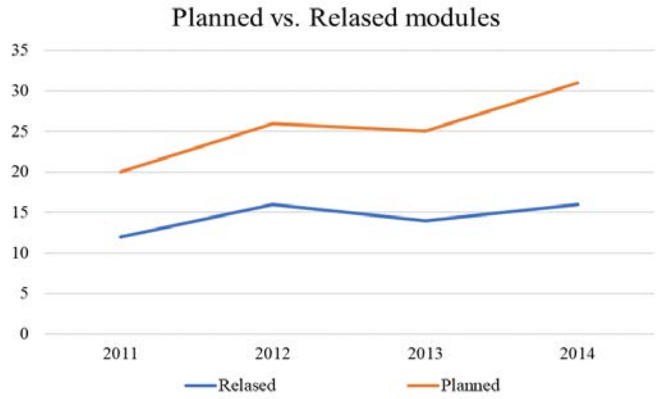


Fig. 1. Relation between Planned Modules and Released Modules within the Productive Environment. Own Elaboration.

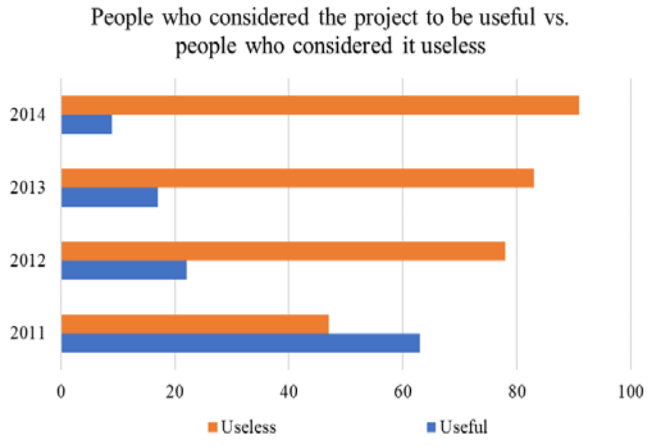


Fig. 2. Percentage of People who think the Project is useful Versus the Ones that Considered it useless. Own Elaboration.

In Fig. 2, we can appreciate that in 2011 the majority thought the global technology project to be useful to the company; by 2012, however, most people deemed it useless. Among the reasons they identified for their feeling, are that none of the modules that were supposed to be active worked correctly in a productive environment, and that there were delays in most of them, which caused discomfort to the employees who were promised the service, management and operative areas included.

As it has been said, many staff members considered that the systems employees were being overpaid for the few services they provided, and there was not an effective strategy that defined the direction in which the global project went, for modules that depended on other modules that had not yet been finished were being started. In a desperate attempt of delivering results, the SCRUM agile management framework was included with the rest of the projects, but for many it was what sunk what was left of the global project. The same interviewed people thought that there was not an accurate business focus and that the objectives were not being set clear for the conclusion of the global project.

VI. Discussion

It is important to position ourselves at the moment of the IS's implementation, since, for this case of study, the functioning stage never fully arrived. Though the most common cause of failure during implementation is resistance to change [7], for TVT this factor was discarded since the employees showed that they were open to collaborating and were even eager to take part take part in the project.

An investment was made in technology training courses for all the collaborators so that they would easily adapt to the new operation modules. The average age of the employees was of 32 years, and there was a solid organizational culture that was well aligned with the technology that was to be implemented.

Here the most influential factors were a deficient administration, a bad definition of the project and inappropriate consultancy. All the causes that the statistics company found are related to these factors, with the lousy administration and management as the main causes. The beforementioned causes are usually the most determining for failure, as mentioned by [3]. A good manager or systems area director must have enough skills to use and keep a particular kind of technology to aid in the business processes within the company [8]. In this case, however, the managers and directors usually delegated activities to people who lacked knowledge in technology and put their untrained "trusted people" in high command positions. The fact that several modules whose development depended on previous, unfinished modules were started can be due to the company not defining clear short and medium term goals; the activities were instead loosely defined and people were hired even though their services were not needed, as the projects were in standby. Contrary to Brhel's statement in their article [5], the SCRUM framework did not come to the rescue of the global project, as it was incorporated without clear knowledge and correct management of its handling and lacked a correct consultancy during its incorporation. The failure of introducing the agile framework SCRUM is due to it being carried out incorrectly, recurring to bad practices that harm the agile development's reputation.

VII. Conclusion

Obtaining a competitive advantage nowadays must be a key strategy for companies, which makes it worth to invest in technology to help achieve it. TVT invested (and had no qualms about it) a lot of money to get said advantage over its closest competitors, which in Mexico include a leading transnational company. The problem is not having a laid-out action plan to achieving this advantage, and the planning not being oriented towards the company's business goals, thus not having a defined strategy and planning that will lead to the desired results. The lack of a responsible management department that is aware of the goals they are to reach is most counterproductive in the implementation of an IS; even worse is not noticing the real cause of failure in time. In TVT they noticed too late, and there was nobody who could rescue a project that was condemned to fail from its management. It is important to keep metrics or indicators during both the planning and development of such a project, so that when problems surface they can be corrected and avoid major diversions from the layout. Among all the causes cited in this paper for the failure in an IS's implementation and functioning, it is probable that someone who works in the systems area has faced at least one in their career.

Acknowledgment

This work was supported by Instituto Politécnico Nacional, Maestría en Informática [grant numbers SIP20201101, SIP20200832].

References

[1] L. Raymond y F. Bergeron, Impact of Project Management Information Systems on Project Performance, vol. 2. New York: Springer, 2015.

[2] M. J. Liberatore y B. Pollack-Johnson, "Factors influencing the usage and selection of project management software", IEEE Trans. Eng. Manag., vol. 50, núm. 2, pp. 164–174, 2003.

[3] K. Chari y M. Agrawal, "Impact of incorrect and new requirements on waterfall software project outcomes", Empir. Softw. Eng., vol. 23, núm. 1, pp. 165–185, 2018.

[4] D. Vesset et al., "IDC's Worldwide Big Data and Analytics Software Taxonomy, 2017", IDC Analyze the future website. [En línea]. Disponible en: https://www.idc.com. [Consultado: 20-may-2020].

[5] M. Brhel, H. Meth, A. Maedche, y K. Werder, “Exploring principles of user-centered agile software development: A literature review”, Inf. Softw. Technol., vol. 61, pp. 163–181, 2015.

[6] G. J. Gery, “Training vs. Performance Support: Inadequate Training is Now Insufficient”, Perform. Improv. Q., vol. 2, núm. 3, pp. 51–71, 2008.

[7] H.-W. Kim y A. Kankanhalli, “Investigating User Resistance to Information Systems Implementation: A Status Quo Bias Perspective”, en MIS Quarterly, vol. 33, núm. 3, 2009, pp. 567–582.

[8] R. Hodson, “Good Jobs and Bad Management: How New Problems Evoke Old Solutions in High Tech Settings”, en Industries, Firms, and Jobs: Sociological and Economic Approaches, First., G. Farkas y P. England, Eds. New York: Springer Science+Business Media New York, 1988, pp. 247–280.

[9] M. R. Ochoa-Aliaga, “Enfoque de liderazgo gerencial para el compromiso del factor humano en el desarrollo del proceso estratégico”, Rev. Cienc. y Cult., núm. 2, pp. 112–114, 1997.

[10] A. Selmeci, I. Orosz, G. Györök, y T. Orosz, “Key Performance Indicators used in ERP performance measurement applications”, en 2012 IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics, SISY 2012, 2012, pp. 43–48.

[11] A. Gordillo-Mejía, D. Licona-Padilla, y E. Acosta-Gonzaga, Desarrollo y aprendizaje organizacional mediante el uso de TIC’s, Segunda Ed. México: Editorial Trillas, 2013.

[12] E. Claver, J. Llopis, M. Reyes González, y J. L. Gascó, “The performance of information systems through organizational culture”, Inf. Technol. People, vol. 14, núm. 3, pp. 247–260, 2001.

[13] S. Jackson, “Organizational culture and information systems adoption: A three-perspective approach”, Inf. Organ., vol. 21, núm. 2, pp. 57–83, 2011.

Handwritten Arabic Characters Recognition using a Hybrid Two-Stage Classifier

Amjad Ali Al-Jourishi[1], Mahmoud Omari[2]
Computer Science Department
Amman Arab University
Amman, Jordan

***Abstract*—Handwritten Arabic character recognition presents a big challenge to researchers in the field of pattern recognition. Arabic characters are characterized by their highly-cursive nature and many of them have a similar appearance. For example, the only difference between some of the alphabet characters is the existence of a number dots above or below the main character shape. This paper proposes a system for isolated off-line handwritten Arabic character recognition using the Discrete Cosine Transform (DCT) as the feature extraction method and a two-stage hybrid classifier. The two stages are a Support Vector Machine (SVM) and a neural network (NN). The first stage is a two-class SVM classifier which classifies a character either a character with dot(s) or without dot(s). The output of this stage is used to extend the feature vector of the character by the class value to give it an extra unique feature. The extend feature vector is fed to a multi-class neural network model to classify the character. The proposed approach is tested on a database of Arabic handwritten characters called AlexU Isolated Alphabet (AIA9K) containing 8,737 character images. The experimental results of the first stage classifier showed a high recognition accuracy rate of 99.14%. The proposed two-stage hybrid classifier obtained an average recognition accuracy rate of 91.84% over all Arabic Alphabet characters.**

Keywords—Arabic character recognition; Support Vector Machine (SVM); neural network (NN); hybrid classifier

I. Introduction

The optical character recognition OCR is an important field for offline handwriting recognition systems. Offline handwriting recognition systems are unalike online handwriting recognition systems [1] [2]. In certain contexts, the ability to handle large amounts of handwritten script data is priceless. An example of these applications is the automation of copying script in old documents taking into account the complex and irregular nature of writing [3]. Arabic optical character recognition is still primitive and slowly developing compared to other languages [4].

The main challenge in the Arabic script recognition systems originate from the cursive nature of the characters. Moreover, some characters have two to four different forms depending on its position in the word. Several characters are connected with complementary parts above, below, or inside them. In addition, there are many similarities among the Arabic characters with regard to their structure and morphology that makes it difficult to recognize, particularly those characters that have dots. Therefore, to distinguish some characters from each other, Arabic Language uses a variety of dots, one, two or three dots, above or below the main shape of the character. These characters are: (ب, ن, ت, ث, ج, خ, ف, ق, غ, ض, ظ, ش, خ, ي, ز). The elimination of any of the dots will cause misinterpretation of that character. In addition, some people handwrite these dots as dashes, which brings more difficulties for a recognition system.

The Arabic alphabet is used for writing different languages such as Persian, Urdu, and Jawi [5]. The Arabic alphabet consists of 28 letters and most of them are written in a cursive manner. There are several shapes for most of the Arabic letters depending on its position within the word. Those different shapes correspond to the different placements of the character within a word, such as at the beginning, in the middle, at the end.

In automated optical character recognition systems, the choice of feature extraction method could be the most important issue for obtaining high recognition accuracy [6]. AlKhateeb, R., J., Ipson, & El-Abed [6] proposed an approach for recognizing handwritten Arabic words that utilizes Discrete Cosine Transform (DCT) as the feature extraction method. The resulted features are used to train a neural network for classification. Lawagali, Bouridane, Angelova, & Ghassemlooy [7] compared the effectiveness of using DCT and Discrete Wavelet Transform (DWT) in capturing the features of handwritten Arabic characters. The authors built a new dataset containing 5600 characters covering all Arabic characters. To compare the two feature extraction methods, a neural network model was built and implemented. The results of the experiment results showed that the use of DCT-based feature extraction method outperformed DWT.

Furthermore, distinguishing the Arabic handwritten text is a difficult task due to the fact that Arabic characters have complex formality, and writing style from one person to another is highly variable. The aim of this research is to confirm the feasibility of using multi-stage classifier for recognizing offline handwritten isolated Arabic characters. We believe that each stage of the classifier allows partial recognition and reduces overall misclassification errors.

The rest of the paper is organized as follows. Section 2 gives brief overview of related work. The proposed technique is presented in Section 3. Section 4 shows the details of the experiments and results discussion. Finally, Section 5 closes with a conclusion.

II. Related Work

Several techniques have been proposed for offline Arabic handwritten recognition [8]. The techniques vary in the type of classifiers being used. Some of them uses a single classifier and other tries to benefit from more than one classifier by constructing a multi-stage hybrid classifier. Most of the techniques implement neural networks in addition to some other classifiers. In this section, we present a review of the related work that uses single and multi-stage classifiers in building the recognition system for offline handwritten Arabic characters. In addition, we present the results of those techniques that have used the same dataset being used here in this paper.

Torki, Hussein, Elsallamy, Fayyaz, and Yaser [10] presented a comparative study of the window-based descriptor on the application of handwriting recognition of Arabic alphabets. It shows a detailed empirical assessment of the different descriptors with many classifiers. The purpose was to evaluate different window-based descriptors as feature extraction methods. They used AlexU Isolated Alphabet (AIA9K) datasetat with defferent descriptors in literature, namely, HOG, SIFT, SURF, LBP, and GIST. The paper presented a comparative evaluation of four common classifiers on the chosen descriptors, namely, Logistic Regression, Linear SVM, Nonlinear SVM, and Artificial Neural Networks. The proposed system obtained a recognition accuracy rate of 72.64% for NN and 70.05% for SVM with SURF descriptors.

Alijla and Abu Kwaik [11] proposed a recognition system for online handwriting of isolated Arabic characters, suitable for hand-held applications. The proposed system uses feedforward and backpropagation neural networks as the main classifier. The system employs online feature extraction methods including Number of Segments and Letter Direction. The system also used Density, Aspect Ratio and Character Alignment as the offline features and arranged the characters into four groups according to the number of segments in the Arabic character. The system is designed with four neural networks, one for each group of characters. The system achieved a recognition accuracy of 95.7% on a dataset of untrained writers.

Ali, Shaout, and Elhafiz [12] proposed two phase classifier to recognize offline handwritten Arabic characters. The two-phase system is based on dividing the characters into two groups according to their similarity. In the second phase, a specific classifier for each character group is used to classify the character within a group. The proposed system uses NN for both classification phases. The feature extraction method used in the system is the Principal Components Analysis (PCA) and extracted a feature vector of 95 values. The proposed system applied on a private dataset and achieved a recognition accuracy rate of 93%.

Abed & Alasad [13] suggested an approach for the identification of isolated Arabic characters using error back propagation neural networks (EBPANN). The neural network was optimized to recognize 12 characters which achieved a recognition accuracy rate of 93.61%.

Al-Boeridi and Ahmad [14] demonstrated the performance of a hybrid Off-line handwriting recognition system (OFHR) for Malay Bank Cheques written in Malay language. The proposed recognition system used two individual classifiers, namely, NN and SVM. The authors concluded that these two classifiers gave an exceptional result. But at the same time, this hybrid method is difficult to implement and takes longer to obtain satisfactory results. The experimental results show that NN has a higher recognition rate at 99.06% and SVM at 97.15%.

Al-Jubouri and Abusaimeh [15] proposed two-stage classifiers to recognize handwritten Arabic characters. The first stage uses the Support Vector Machine classifier which classifies the characters into two groups namely: characters with dot(s) and characters without dots. The second stage uses a neural network classifier. The experiment conducted on a dataset of 2927 character images from the IFN-ENIT dataset with no character segmentation. The proposed approach used Discrete Wavelet Transform (DWT) and curvelet feature extraction methods. The experiment result showed a recognition accuracy rate of 92.2%.

Younis [16] presented a deep neural network to solve the problem of recognizing offline handwritten Arabic characters based on a Convolutional Neural Network (CNN) models. The deep CNN has been tested on two datasets, AIA9K and AHCD. The accuracy for the two datasets were 94.8% and 97.6%, respectively.

III. Proposed Technique

Classification is a general categorization in which the body and key objects are identified and recognized. The main objective of using SVM in the proposed system is to separate the characters with dots and those without dots. This separation of characters into two classes, makes is easier for the second stage of NN classifier to recognize the individual character. The distinction between characters significantly reduces the error rate in recognizing some characters within the system. In other words, the probability of characters being similar in shape will be reduced when the classification is augmented with a good feature extraction method, such as DCT [17].

The choice of feature extraction method is the most important step to achieve high recognition accuracy in automatic recognition systems. One of these methods is the 2D Discrete Cosine Transform (DCT), which is a transform method for converting image data into its primary components by calculating a set coefficients and store them in a 2D matrix. These coefficients are categorized as low-frequency values located in the top left corner and high-frequency located at the bottom right corner of the 2D matrix. Thus, the ability of DCT to pack the energy of the image to a few low-frequency coefficients is considered as one of its main characteristics [18].

The Support Vector Machine classifier is one well-known classifiers and have been extensively used in many industrial applications [19]. SVMs gained considerable interest in the research community and proven to have many characteristics useful in Machine Learning applications.

Neural network is one classifier that is used extensively in many applications of pattern recognition, including image recognition, speech recognition, and text recognition [20]. This paper focuses on using multi-class NN within a two-stage Arabic character recognition system. Any multi-class problem can be defined by Three-tuples (S, T, C), where F represents an n-dimensional feature space, T is a training dataset which is a subset of S, and C is a set of class labels [21]. Each element, e, in T is associated with a class label c where the number of class labels is greater than 2. In the training phase, the NN is trained on T to produce a model function F that maps any given feature vector $x \in S$ such that $F(x)=c$, where $c \in C$.

A multi-class NN classifier maps the input feature from the feature space into the output space. The NN classifier consists mainly of three types of layers; input, output, and hidden. Neural networks are characterized by their topology, and this is determined by the learning algorithm and the neurons characteristics. The NN has been applied to solve the problem of recognizing both printed and handwritten Arabic characters. Various methods for classification augmented by various feature extraction methods have been proposed. In this paper, a multi-layer perceptron backpropagation (BP) NN [22] is used for training and then for classification of handwritten Arabic character. The input layer of the NN is fed with the training feature set T, while the output layer produces the class of the tested input.

This research explores the classification capabilities of both the SVM and the NN to produce intelligent off-line Arabic handwritten character recognition system. The major steps in the proposed classification system is shown in Fig. 1, which includes feature extraction step and two-stage classifier, explained in the following subsections.

A. *Feature Extraction Phase*

The Discrete Cosine Transform DCT [23] is used as a feature extraction method for the alphabet character images. Using DCT as a feature extraction technique can remove the redundancy from the image data and earn a more effective representation of the character image by a set of numerical values [24]. In handwritten text, the features represent the useful information extracted from the characters. This information is then used to classify characters and assist in the classification process. The DCT transforms an image from the spatial domain to the frequency domain. This transformation can help reduce redundancy and focus on the power of the image in a very limited frequency range. Hence, the DCT converts the data of the image into elementary frequency components (i.e., coefficients). The coefficients matrix resulted from applying the 2D DCT function contains low-value coefficients located at the bottom right corner and the high-value coefficients at the upper left corner. The high-value coefficients are the most important ones as they can be used to represent the image and can also be used to reconstruct the original image with some image quality loss. Thus, DCT is used in the JPEG lossy image compression algorithms [25].

The input for the DCT is 32x32 pixel black and white image of a character. The 2D DCT produces a 32x32 two-dimensional matrix of data coefficients. These coefficients are considered accurate representation of the original image; however, the transformation has made it easier to get rid of redundant information. The number of DCT coefficients representing the image are reduced to a smaller set of possible values that hold most of the energy in the image. The feature vector of an image is generated by extracting the higher coefficients values in the matrix resulting from applying the 2D DCT. These coefficients constitute the minor diagonal elements of the matrix. The coefficients are read from the matrix in a zigzag fashion and storing them in a one-dimensional feature vector as shown in Fig. 2.

Extensive experiments were carried out using MATLAB [26] to find those DCT coefficients that are the most representative features of a character image. The coefficients chosen were those ones that are sufficient to reconstruct the original image when performing the inverse DCT, rather than all coefficients of the image. The total number of these coefficients that represent the minor-diagonal elements of the 32x32 pixels image is 560. This number is determined by empirical testing to reconstruct perceivable characters with a minimum number of coefficients. These features are utilized for training and testing phases of the system.

B. *Training Phase*

The training phases for the SVM is shown in Fig. 3. The purpose of training is to produce a SVM model that can, later on, differentiate between letters with dot(s) and letters without dot(s) of the alphabet characters. During training, the SVM is fed with the feature vectors of all characters. As mentioned before, the feature vector represents the *n* DCT coefficients representing a character.

During the training phase of the neural network, as shown in Fig. 4, the inputs are two manually separated subsets corresponding images for letters with dot(s) and letters without dot(s). The images are fed to feature extraction step that uses DCT at which vectors of features are generated for both types of letters. To distinguish between those two datasets, the feature vector is extended with an extra value, here, this values is either 1 or 2, corresponding to letters with dots and letters without dots, respectively. Now the feature vector length is *561.* The extended feature vectors are then fed to the neural network running a feedforward back-propagation algorithm for training.

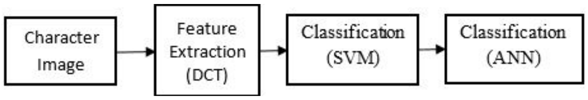


Fig. 1. Major Steps of the Classification System.

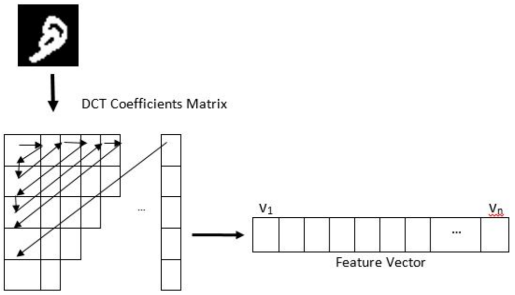


Fig. 2. Rearranging DCT Coefficients into One-Dimensional Feature Vector.

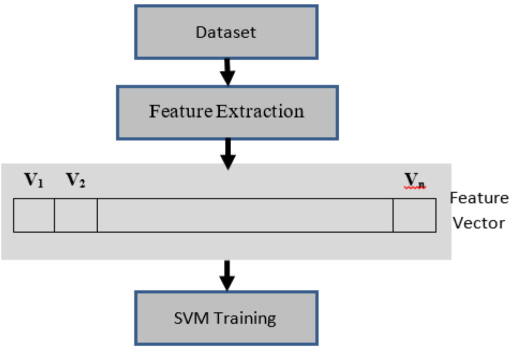


Fig. 3. The SVM Training Phase.

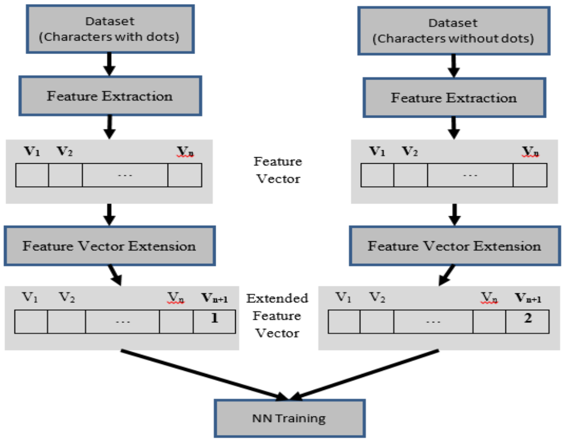


Fig. 4. The NN Training Phase.

C. Testing Phase

During the testing phase of the system, the SVM is used to classify the character either a character with dot(s) or without dot(s). When the SVM classifies a character, the output class value of 1 or 2, corresponding to a character with dot(s) and a character without dot(s), is appended to the original feature vector of the corresponding character. After appending the feature vector for the character being classified, the NN is fed with that feature vector to give the final class of the character as shown in Fig. 5.

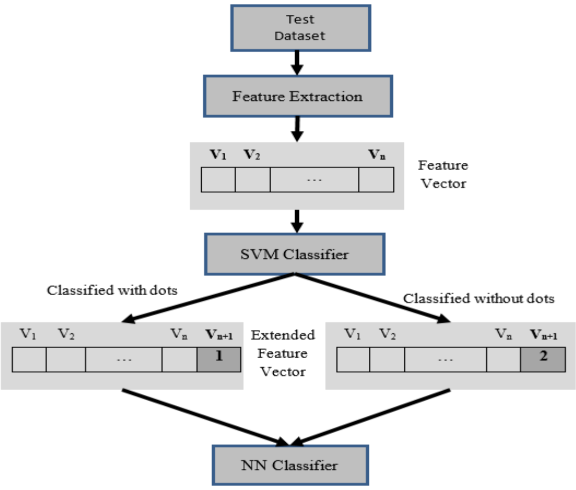


Fig. 5. The Proposed System Recognition Phase.

IV. Experiments and Results

The dataset used in the experiments is a novel dataset called AlexU Isolated Alphabet (AIA9K). The database was built and proposed by researchers at the University of Alexandria/Egypt [9]. The database contains 8,737 valid samples of the 28 Arabic alphabet letters. The extracted images of handwritten Arabic characters were written by 107 volunteer Arabic writers among the students in the Faculty of Engineering at Alexandria University. Each writer wrote the Arabic characters three times on a form. All the Arabic characters were scanned from the forms using a scanner at a resolution of 300dpi.

To verify the proposed approach, three experiments were implemented and carried out using MATLAB version R2016a [26]. The first experiment was performed to train and test the classification accuracy of the SVM classifier. The dataset is divided into 60% for training and 40% for testing. The second experiment intended to test the performance of a standalone neural network classifier using the original dataset. The third experiment is conducted to measure the performance of the proposed two-stage classifier. In this experiment, the dataset is divided into 70% for training, 15% for validation, and 15% for testing.

The first experiment was conducted to test the recognition accuracy of the SVM classifier. The rule of the SVM model classify individual as either character with dot(s) or without dot(s). The recognition accuracy of this model is shown in Table I. The results are very promising, and the overall recognition accuracy achieved is 99.14%. It is noted that, more than one-third of the characters are correctly recognized. The lowest recognition accuracy obtained was for the letter “Daad” (ض).

The second experiment was conducted to test the performance of a standalone neural network in which the network was trained and tested on the original dataset. The recognition accuracy for individual characters for this experiment is shown in Table II. It can be seen that the best recognition ratio of 96.57% obtained for the alphabet character “Alif” (ا), while the worst recognition accuracy of 79.75% obtained for the alphabet character “Thaa” (ث). The reason for this low recognition accuracy of the character “Thaa” (ث) is due to the great similarity in the way people writes this character compared to other alike characters. The overall recognition accuracy of the standalone neural network classifier of all alphabet characters is 88.5%.

The third experiment is implemented to test the performance of the proposed approach. First, the feature vectors of the test dataset are fed to the SVM model which produces either one of the two aforementioned classes, either with dot(s) or without dot(s). Following that, the recognized class value is appended to the feature vector of that particular character. The newly appended feature vector is then fed to the NN stage for final classification. The maximum recognition rate result obtained is 97.51% while some characters were difficult to recognize, as they are incorrectly recognized by the SVM stage. As shown in Table III, the character “Miim” (م) and “Baa” (ب) have the highest recognition rate of 97.51%, while the character “Thaa” (ث) has the lowest recognition rate of 79.13%. The overall recognition accuracy achieved for this two-stage hybrid approach is 91.84%.

TABLE I. RECOGNITION ACCURACY RATES OF THE SVM CLASSIFIER

Char.(Name)	Accuracy (%)	Char.(Name)	Accuracy (%)
(Alif) ا	99.69	(Daad) ض	96.26
(Baa) ب	100.00	(TAA) ط	96.88
(Taa) ت	100.00	(Dhaa) ظ	98.75
(Thaa) ث	100.00	(Ayn) ع	98.73
(Jim) ج	99.69	(Ghayn) غ	98.75
(Haa) ح	99.69	(Faa) ف	99.69
(Khaa) خ	99.69	(Qaaf) ق	100.00
(Daal) د	100.00	(Kaaf) ك	99.38
(Dhal) ذ	99.38	(Laam) ل	100.00
(Raa) ر	99.07	(Miim) م	100.00
(Zayn) ز	98.13	(Nuun) ن	100.00
(Siin) س	96.56	(Haa) هـ	98.75
(Shiin) ش	98.75	(Waaw) و	100.00
(Saad) ص	98.12	(Yaa) ي	100.00

TABLE II. RECOGNITION ACCURACY RATES OF STANDALONE NEURAL NETWORK CLASSIFIER

Char.(Name)	Accuracy (%)	Char.(Name)	Accuracy (%)
(Alif) ا	96.57	(Daad) ض	83.49
(Baa) ب	96.26	(TAA) ط	85.31
(Taa) ت	87.23	(Dhaa) ظ	81.93
(Thaa) ث	79.75	(Ayn) ع	87.97
(Jim) ج	84.47	(Ghayn) غ	89.41
(Haa) ح	88.75	(Faa) ف	87.23
(Khaa) خ	86.25	(Qaaf) ق	82.55
(Daal) د	86.65	(Kaaf) ك	91.59
(Dhal) ذ	89.41	(Laam) ل	95.92
(Raa) ر	86.29	(Miim) م	96.26
(Zayn) ز	89.10	(Nuun) ن	87.23
(Siin) س	87.50	(Haa) هـ	87.50
(Shiin) ش	89.10	(Waaw) و	95.95
(Saad) ص	91.22	(Yaa) ي	85.98

TABLE III. RECOGNITION ACCURACY RATES OF THE PROPOSED TWO-STAGE CLASSIFIER

Char.(Name)	Accuracy (%)	Char.(Name)	Accuracy (%)
(Alif) ا	97.20	(Daad) ض	90.97
(Baa) ب	97.51	(TAA) ط	92.81
(Taa) ت	88.98	(Dhaa) ظ	90.65
(Thaa) ث	79.90	(Ayn) ع	96.52
(Jim) ج	95.65	(Ghayn) غ	91.90
(Haa) ح	93.75	(Faa) ف	88.74
(Khaa) خ	89.69	(Qaaf) ق	83.80
(Daal) د	91.61	(Kaaf) ك	93.77
(Dhal) ذ	92.83	(Laam) ل	96.55
(Raa) ر	88.79	(Miim) م	97.51
(Zayn) ز	92.21	(Nuun) ن	88.92
(Siin) س	89.69	(Haa) هـ	93.75
(Shiin) ش	89.85	(Waaw) و	96.08
(Saad) ص	92.48	(Yaa) ي	89.41

It is clear that the best classification accuracy obtained in the proposed approach is for those characters that were well recognized by the SVM, which affects positively the final NN classifier. This proves the effectiveness of the proposed approach in recognizing characters over the standalone NN classifier.

V. CONCLUSIONS

This paper proposed an isolated Arabic offline handwritten alphabet character recognition system. The proposed system employs the DCT as the feature extraction method and utilizing both a Support Vector Machine and a neural network, in a two-stage hybrid arrangement. The reason behind using two-stage classifier is to overcome the main limitations of using traditional single-stage classifier. The first stage SVM classifier achieved a recognition accuracy of 99.14%, which classifies the characters into one of two classes, namely, characters with dot(s) and characters without dot(s). The notion behind this approach is to make it easy for the neural network stage to classify each character after being discriminated as either with dot(s) or without dot(s). The experimental results showed that the recognition accuracy of the neural network classifier stage depends highly on the accuracy of the first stage classifier. That is, when there is a misclassification in the first stage, subsequently, affecting the results of the final stage. Despite this, the recognition accuracy of the proposed two-stage hybrid approach achieved 91.84%. Furthermore, the experimental results showed that the two-stage hybrid classifier approach outperforms a standalone neural network classifier. Further investigation is need to enhance the proposed approach by employing different feature extraction methods as well as applying this hybrid approach on different datasets, and possibly different types of classifiers.

REFERENCES

[1] R. Plamondon and S. N. Srihari, "On-line and Off-line Handwriting Recognition: A Comprehensive Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 2, no. 1, pp. 63-84, 2000.

[2] R. Renuka, V. Suganya and K. B. Arun, "Oline Handwritten character Recognition using Digital pen for static Authentication," in the international conference on computer compunction and informatics, Coimbartore, India, 2014.

[3] A. Belaïd and N. Ouwayed, "Segmentation of Ancient Arabic Documents," in Guide to OCR for Arabic Script, London, Springer-Verlag, 2011, pp. 103-122.

[4] M. T. Parvez and S. A. Mahmoud, "offline Arabic Handwritten text recognition," ACM computing surveys, vol. 45, no. 2, pp. 23-33, 2013.

[5] S. Naz, A. I. Umar, R. Ahmed, S. F. R. Muhammad Imran Razzak and F. Shafait, "Urdu Nasta'liq text recognition using implicit segmentation based on multi-dimensional long short term memory neural networks," SpringerPlus 5, vol. 5, no. 1, 2016.

[6] J. H. AlKhateeb, J. R., J. J., S. S. Ipson and H. El-Abed, "Word-based handwritten arabic scripts recognition using dct features and neural network classifier," in 5th International Multi-Conference on Systems, Signals and Devices, 2008.

[7] A. Lawagali, A. Bouridane, M. Angelova and Z. Ghassemlooy, "Handwritten Arabic Character Recobnition: Which Feature Extraction Method," International Journal of advanced science and technology, vol. 34, september 2011.

[8] R. K. Patel M.S, "Offline Arabic handwriting recognition: a survey," International Journal of Engineering Research & Technology (IJERT), vol. 28, pp. volume 3, Issue 19, 2015.

[9] M. Torki, M. E. Hussein, A. Elsallamy and M. F. S. Yaser, "Window-Based Descriptors for Arabic Handwritten Alphabet Recognition: A

Comparative Study on a Novel Dataset," Arxiv.org, p. arXiv:1411.3519, 2014.

[10] B. Alijla and K. Kwaik, "OIAHCR: Online Isolated Arabic Handwritten Character Recognition Using Neural Network," The International Arab Journal of Information Technology, vol. 9, no. 4, pp. 343-351, July 2012.

[11] O. B. Ali, A. Shaout and M. Elhafiz, "Two stage classifierfor Arabic Handwritten Character Recognition," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 12, pp. 646-650, 2015.

[12] M. A. Abed and H. A. A. Alasad, "High Accuracy Arabic Handwritten Characters Recognition Using Error Back Propagation Artificial Neural Networks," International Journal of Advanced Computer Science and Applications, vol. 6, no. 2, pp. 145-152, 2015.

[13] O. Al-Boeridi, S. M. S. Ahmad and J. Paw, "A scalable hybrid decision system (HDS) for Roman word recognition using ANN SVM: study case on Malay word recognition," Neural Computing and Applications, vol. 26, no. 6, p. 1505–1513, 2015.

[14] M. Al-Jubouri and H. Abusaimeh, "Offline Arabic Handwritten Isolated Character Recognition System Using Support vector Machine and Neural Network," Journal of Theoretical and Applied Information Technology, vol. 95, no. 10, pp. 2315-2322, 2017.

[15] K. S. Younis, "Arabic Handwritten Character Recognition based on Deep Convolutional Neural," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 3, no. 3, pp. 186-200, 2017.

[16] A. Al-Haj, "Combined dwt-dct digital image watermarking," Journal of Computer Science, vol. 3, no. 9, pp. 740-746, 2007.

[17] P. G. Ken Cabeen, "Image Compression and the Discrete Cosine Transform," 2019. [Online]. Available: https://www.math.cuhk.edu.hk/ ~lmlui/ dct.pdf. [Accessed 20 5 2019].

[18] J. S.-T. Nello Cristianini, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge university press, 2000.

[19] M. E. S. R. Nidal Lamghari, "Comparison of Neural Network Parameters for Classification of Arabic Handwritten Isolated Characters," International Journal of Computer Applications, pp. Vol. 178(9),42-49 , 2019.

[20] R. Rojas, Neural Networks: A Systematic Introduction, Springer Science & Business Media, 2013.

[21] M. A. Nielsen, " How the backpropagation algorithm works," in Neural Networks and Deep Learning, Determination Press, 2015, pp. 200-220.

[22] A. Nasir, "How I Came Up With the Discrete Cosine Transform," Digital Signal Processing, vol. 1, no. 1, pp. 4-5, 1975.

[23] A. Lawgali, A. Bouridane, M. Angelova and Z. Ghassemlooy, "Handwritten Arabic Character Recognition: Which Feature Extraction Method," International Journal of Advanced Science and Technology, January 2011.

[24] A. M. Raid, W. M. Khedr, M. A. El-dosuky and W. Ahmad, "Jpeg Image Compression Using Discrete Cosine Transform - A Survey," International Journal of Computer Science & Engineering Survey (IJCSES), vol. 5, no. 2, pp. 39-47, 2014.

[25] MATLAB, Version R2016a, Natick, Massachusetts: The Mathworks Inc., 2016.

[26] M. Torki, M. E. Hussein and A. Elsallamy, "Window-based Descriptors for Arabic Handwritten Alphabet Recognition: A Comparative Study on a Novel Dataset," arXiv preprint arXiv:1411.3519, 2014.

Critical Factors Affecting the Intention to Adopt Big Data Analytics in Apparel Sector, Sri Lanka

Hiruni Bolonne[1]
Department of Accounting
University of Sri Jayewardenepura
Colombo, Sri Lanka

Piyavi Wijewardene[2]
Odiliya Virtual Technologies (Pvt) Ltd
Colombo, Sri Lanka

***Abstract*—Big data has become a potential research area in apparel industry due to vast amount of data generated in a short period of time. However, the inability to adapt to the challenging and digital environment has pulled out the weaker from the industry while growing the adopters more and more powerful players. As the insights generated out of data becoming core competitive advantages, now it is pertinent to identify which factors would affect the intention to adopt big data analytics in an apparel sector organization. The three contexts of the Technology-Organization-Environment (TOE) framework along with Technology Acceptance Model (TAM) were used as foundational frameworks to explore the influence on the attitude towards using of users which would ultimately affect to the intention of adopting big data analytics. The findings generated from the study denotes that factors considered in both TOE framework and TAM model except organizational context having a positive correlation towards the user's attitude of using which would ultimately lead the organization in enhancing its intention to adopt big data analytics. Finally, the research concludes that the variable, attitude towards using plays a positive mediating role between the direct relationship of critical factors affecting the intention to adopt big data analytics. It is hoped that findings of this research would enrich the existing literature while affecting practitioners to involve in adopting big data analytics by prioritizing investments accordingly.**

Keywords—Critical factors; TOE framework; Technology Acceptance Model (TAM); attitude towards using; intention to adopt; Big Data Analytics (BDA); apparel sector; Sri Lanka

I. Introduction

The evolution of the concept, big data has occurred and expecting it to occur in future to develop more and more new and powerful computational technologies to respond and exploit the various types of data stored in the vast repositories which is continuing to accumulate data at a non-stop rate [1]. Hitt and Kim [2] have concluded that organizations which have already accepted the data-driven decision-making patterns and adopted accordingly, were able to enhance their productivity rates by 5-6%. Since it has been several years from the introduction of big data analytics to the corporate world, still the majority of the industries operate at an early stage of adoption [3], while the rest struggle in the dark seeking for the ways to fully understand the functions and capabilities of the concept, big data [4]. Many research papers on this trending area depict a variety of influencing factors which would impact the intention of adopting big data and would go beyond merely focusing on the end user acceptance. When comes to Sri Lankan context, there is a dearth of studies which identifies the key potential factors affecting the organizations to develop an environment for big data analytics. To address the research gap, this paper grounded in Technology-Organization-Environment (TOE) framework of Tornatzky and Fleischer [5], in combination with the Technology Acceptance Model (TAM) seeking to legend greater clarity on the critical factors that would influence big data adoption to broaden the knowledge and to assist the Sri Lankan apparel sector organizations with the adoption of big data in deriving towards the competitive advantage over its rivals which exist both within and outside the territory.

The remainder of this article proceeds as follows. Section II elaborates the research problem. Section III depicts the theoretical background. Significance of the study is discussed under Section IV and Section V covers the research design and methods. Section VI focuses on the analysis and discussion. Section VII concludes the overall idea of the study while final section pointing out the limitations of the study and the future research directions where further research could be carried out.

II. Research Problem

In Sri Lanka, the apparel industry is known to be among the largest contributors to economic growth with a contribution over $5 billion to gross domestic production which also accompanies a labour force of over 500,000. The Joint Apparel Association Forum (JAAF) is sharing a strongly confident view that the industry inherits a continuous potential to grow in spite of the rapidly evolving and increasing competition which the global market for apparel is subjected to [6]. Neighbour nations of Sri Lanka in the South Asian region are much bigger and expanding their industries while Sri Lanka are yet to make greater strides to face the competition [7].

The real challenge is to be competitive and yet increase significantly the share of value and volume in the market [7]. Speaking about the radical state of uncertainty around the globe, there is heightened demand for volatility, geopolitical risks, natural disasters, terrorist attacks, social media disruptions which are all taking their toll on global supply chains [7]. The changes are inevitable and will disrupt the textile supply chains [7]. The present circumstance of the Sri Lanka's apparel industry doesn't sound performing at a good rate due to COVID-19 pandemic crippling over the key customers which has led increased cancellation of orders which is a massive blow to the industry [8]. Brandix CEO [8] predicted about the possibility of a price war once the demand shrinks as the biggest player, obviously China working well along with semi lockdown of Bangladesh and with the work of both Vietnam and Indonesia which will have a direct impact to the Sri Lanka's positioning among the other apparel export markets.

The world is on the verge of a change and this change will have implications on the way Sri Lanka does business in the future and the ways, in which Sri Lankan apparel players design, source, manufacture and deliver, must be redesigned, if they truly need to keep the textile industry alive [7]. Therefore Sri Lanka needs to create a knowledge hub with emerging technologies and innovations such as artificial intelligence and advancements in material science coupled with digitalization, big data and analytics to create the perfect platform for modern day business [7].

When looking at the Sri Lankan context, a limited number of studies can be found and Premaratne [9] has analyzed the Sri Lankan lifestyles in data mining in the contexts of education and health. This indicates that the existing literature lacks the sufficient power to conduct quantitative evaluations and to identify and analyze the relationships among the influencing factors. The focus of the study is to identify the critical factors which influence the attitude of using big data analytics and the effect of them on the intention to adopt by an apparel sector organization. Therefore, the research questions of this study are:

- What determinants are responsible for explaining the variation of attitude towards using big data analytics and the weight of each determinant?
- To what extent is the attitude towards using big data analytics affecting the intention to adopt big data analytics?
- Is there a mediation effect of attitude towards using on the direct relationship between critical factors and the intention to adopt of big data analytics?

To address the highlighted research gap, this study specifically focuses on three research objectives as mentioned below:

- To explore the key determinants responsible for explaining the variation of attitude towards using big data analytics and to calculate the weight of each factor
- To examine the relationship between the variables; the attitude towards using big data analytics and the intention to adopt big data analytics
- To assess the mediation effect of attitude towards using on the direct relationship between critical factors and the intention to adopt big data analytics

III. Theoretical Background

A. Definition of Big Data

A definition which was proposed by Gartner [10], defined big data as *"high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making"*. Lukoianova & Rubin [11] further improved the definition by emphasizing the feature, veracity which refers to the accuracy of big data.

B. The Technology–Organization–Environment (TOE) Framework

Tornatzky & Fleischer [5], framed the factors that could impact an organization in embracing innovation or technology were falling under three contexts known as technological (existing and new technologies), organizational (organization size, scope, managerial structure) and environmental (industry competitors, industry size, regulatory environment).

1) Technological Context

- Data-related Infrastructure Capabilities

An organization's readiness and the ability of using big data analytics will be reflected by the extent to which the organization has better infrastructure capabilities and also the integration of underlying data which is required by the supportive data infrastructures can be considered as one of the complex undertakings [12].

- Data Management

Data needs to fulfil the main fundamentals such as reliability, completeness, timeliness, consistency and accuracy for its usage and consumption within the big data analytics domain [13], [14] & [15]. Unless the fundamentals in data are not achieved, usage of data analytics will be hampered and limited as the trust in data will be lost by the users and as well by decision makers [16].

- Privacy & Security

Big data also known to be changing the landscape of security technologies towards networking and forensics and in circumstances where there are no right security and encryption, then it leads to serious threats from big data [17]. If the intelligence of the assaulter is far beyond the installed security limits, then the encryption techniques wouldn't be sufficient either to defend or to shield the log files [18].

2) Organizational Context

- Vision and Strategy

An initiative regarding to big data analytics should pop up based on the business requirements and therefore, to direct the implementation, it needs to be backed by a strategic business vision [19]. If not, the big analytics systems are business centric, it will lead to failure. Always there needs to be a business problem to generate positive results from big data analytics.

- Sponsorship and Governance

Watson et al., [20] emphasized the necessity to receive the consistent support and sponsorship from business executives in securing the operating resources required throughout the implementation process. He [20] also pointed out this as a mostly expected outcome as a significant cost, time and many other resources required to be invested throughout the process and to bring it to alive.

- Organizational Structure

Shiwei, et al., [4] emphasized that the well-built organized structure of an organization has been supportive to the adoption of big data such as cross-organizational collaboration structure, data analytics departments and staff configurations.

- Talent Strategy & Capability Development

This aspect focuses on the highly skilled and experienced people which are vital requirements for the analytical transformation of the organization and for the need of industrialized individual analytical capabilities to move up on the ladder of analytical maturity scale.

- Firm Size

Size of a firm can be measured from different aspects such as annual revenue generated and number of employees employed, which would enhance the level of adoption of big data opening more avenues for organizations to enhance its position in terms of revenue. However, the existing literature indicates that there is no precise direction as some articles depicts a positive relationship [21], [22], [23], while others speak about a negative correlation [24], [25] & [26].

3) Environmental Context

- Market Pressure

Masrek, et al., [27] suggested that it is unavoidable for organizations to face competitive pressures and environment uncertainties, and it has persuaded them to engage in greater sensing and search activities to understand both the internal and the external environments of an organization through strategic initiatives such big data analytics.

- Big Data Pressure

Agrawal [28], has mentioned that big data adoption consists of two main constructs and they are the perceived competition intensity to implement big data adoption and also the risk of competitive disadvantage arises due to the inability in adopting big data analytics.

C. Technology Acceptance Model (TAM)

This study seeks the application of TAM model to observe the user attitude towards big data analytics as it satisfactorily determines how the user perceives the ease of using and as well as the usefulness of the new system which is going to be implemented [29].

1) Perceived Usefulness

A system which bears the status of high perceived usefulness, in turn is one for which user believes that there exists a positive relationship of performance in use [29]. According to Guriting and Ndubisi [30] and Eriksson et al., [31], usefulness is the subjective probability that how it will enhance the efficiency and effectiveness of the way the user supposes to complete a given task using the technology.

2) Perceived Ease of Use

This refers to the degree to which believes that a user has on using a certain system would be free from difficulties and great efforts [29]. Some researchers argued that perceived ease of use could be suggested as to the extent to which the user accepts that the use of an exacting method would be free from adding a single cost to that individual and the extent of reality of that view [29] & [32].

D. Attitude Towards using

Currently, the role performed by the affective attitudes remains as an open issue which needs to be attended [29]. At first, Lancaster [33] pointed that attitude can be introduced as the driver for utility or attributes while Triandis [34] described attitude as the perception of an individual where he/she is either positive or negative towards the innovation adoptions.

E. Intention to Adopt Big Data Analytics

Through this concept, a predefined future behavior is expected to achieve [35] and it is known to be a key predictor when comes to the assessment of an individual’s actual use of technology [36] & [37]. Muhammad, et al., [38] emphasized that the actual use of BDA (Big Data Analytics) in a given organization would depend on the number of individuals positive towards the intention of using BDA.

IV. Significance of the Study

A. Theoretical Significance

It is believed that this study would contribute with noteworthy research insights when comes to the assessment of the intention to adopt BDA in apparel sector, Sri Lanka and also expects that findings which have arrived with, would bridge the main gap in the literature concerning the empirical evidences for the intention to adopt BDA among Sri Lankan apparel sector organizations for the first time. Secondly, the majority of the existing literature supports the research insights of importance, challenges, and opportunities of BDA, as it has been the initial stage of BDA and it was comparatively a new concept [39]. The integration of two technology adoption frameworks (TAM model & TOE framework) would involve in providing the most relevant academic insights with an expanded research model in exploring the intention to adopt BDA and this could be treated as a remarkable point which would enhance the existing literature. Finally, once the investigation is completed, the research will give an idea regarding the factors mainly influence on the user intention to adopt BDA with the mediating effect of attitude towards using and the generated the results would provide the opportunity to reference the findings in the future research and to enhance the understanding of the adoption of BDA.

B. Practical Significance

This study is designed to contribute practically in several ways like theoretical contributions. It is expected that the findings will propose most important guidelines and implications for both practitioners and implementers of big data analytics systems which will ultimately affect to the successful adoption of big data analytics systems in the organizations. Bringing out the connectivity of system with the functions and required tasks along with facilitating the perceived usefulness and perceived ease of use of the system are vital to an organization [39]. This approach will further emphasize the importance of the results that are more fruitful for practitioners when implementing big data analytics systems in developing countries as the findings are based in a developing country which falls under the continent, Asia.

Finally, it is possible to consider that the results of the study will create an initial platform for adopting and promoting big data practices to obtain maximum advantages of innovation technologies in the context of developing countries.

V. Research Design and Methods

A. Research Approach

This research will be based on positivistic paradigm where new knowledge could be discovered through objective measures. In order to achieve the research objectives, quantitative research methodology is primarily used. This study mainly focuses on identifying the relationships between factors such as technological, organizational and environmental factors, perceived usefulness, perceived ease of use and the intention to adopt big data analytics with the mediating effect of attitude towards using. Through using the positivism approach, the study gives the possibility in identifying the sorts of relationships between the independent and dependent variables exist. The approach using, is justifiable as it tries to determine a causal relationship between the variables tested, to generalize to a larger group of individuals than those who are participating in the investigation and to understand behavioral patterns and the reasons behind that behavior. Therefore, the dominance strategy for this study could be identified as positivism.

B. Population and Study Sample

The sampling frame that is considered to test hypothesis are the 13 key players in the Sri Lankan Apparel sector as disclosed in the "*Industry Capability Report – Sri Lankan Apparel Sector*" by Export Development Board, Sri Lanka in January 2020. The study mainly focuses on assessing the level of acceptance of BDA by employees who belong to the category of executive and above, among the key players in Apparel sector. When considering the key 13 players, it approximately caters an employee base of 7,844 who are performing with in the category of executive and above, in the Sri Lankan territory. As the key 13 players are also having operations outside of Sri Lankan territory (overseas), those were not taken to the scope of Sri Lanka as those apparel exports will not get into the count of Export Development Board of Sri Lanka and as well as, the natives in those locations will mainly be used for the operations.

The approximate target population of 7,844 employee base spreads across the 13 key players as Mas Intimates (Pvt) Ltd - 25%, MAS Active Trading (Pvt) Ltd - 14% and Bodyline (Pvt) Ltd - 10%, EAM Maliban Textiles Mahiyanganaya (Pvt.) Ltd – 7%, Jay Jay Mills Lanka (Pvt) Ltd – 7%, Linea Aqua (Pvt) Ltd – 6%, Polytex Garments Ltd – 6%, Brandix Apparel Ltd – 5%, Smart Shirts Lanka Ltd – 5%, Orit Trading Lanka (Pvt) Ltd – 5%, Hirdaramani International Exports Ltd – 4%, Courtaulds Trading Co (Pvt.) Ltd – 4% and Omega Line Ltd – 2% respectively.

Out of the total population, randomly 365 individuals were selected on a pro-rata basis among the 13 key players to generate conclusions about the entire population.

C. Conceptual Framework

Based on the above information, this study mainly focuses on investigating whether attitude towards using, mediating the relationship between the independent variables; perceived usefulness, perceived ease of use, technological factors, organizational factors, environmental factors, and the dependent variable; intention to adopt BDA. This can be graphically presented as shown in Fig. 1 and from the diagram itself, two dependent variables can be identified as *"Attitude towards using"* and "*Intention to adopt big data analytics"*. Out of those two dependent variables, the dependent variable, *"Attitude towards using"* plays a role of a mediating variable which will be assessed in the next section.

D. Hypothesis

H1: Data-related infrastructure capabilities will positively influence attitude towards using big data analytics.

H2: Data management will positively influence attitude towards using big data analytics.

H3: Privacy & security will positively influence attitude towards using big data analytics.

H4: Vision and strategy will positively influence attitude towards using big data analytics.

H5: Sponsorship and governance will positively influence attitude towards using big data analytics.

H6: Well established organizational structure will positively influence attitude towards using big data analytics.

H7: Talent strategy & capability development will positively influence attitude towards using big data analytics.

H8: Firm size will positively influence attitude towards using big data analytics.

H9: Market pressure will positively influence attitude towards using big data analytics.

H10: Big data pressure will positively influence attitude towards using big data analytics.

H11: Perceived usefulness will positively influence attitude towards using big data analytics.

H12: Perceived ease of use will positively influence attitude towards using big data analytics.

H13: Attitude towards using big data analytics will positively influence the intention to adopt big data analytics.

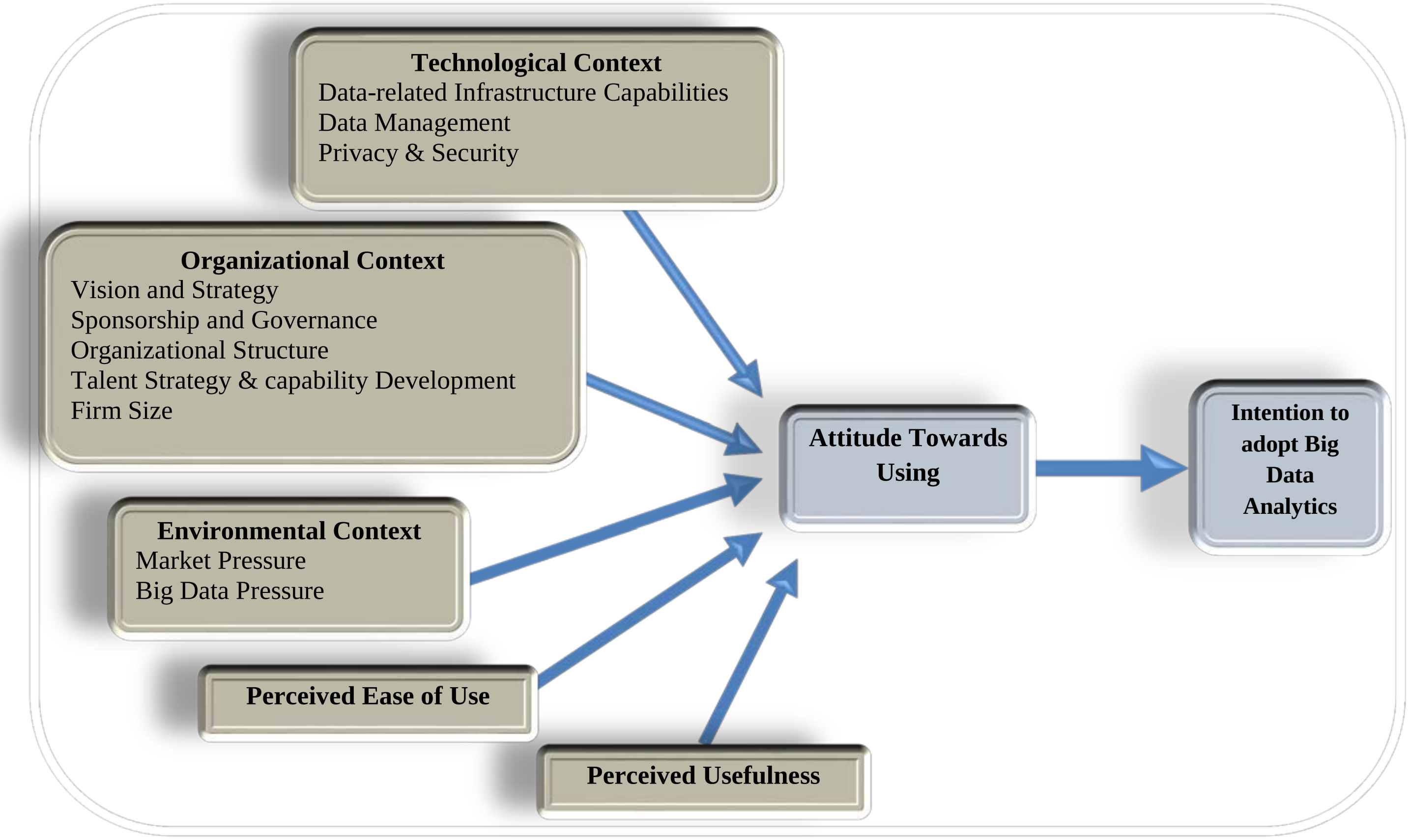


Fig 1. Research Model – (Author Constructed).

E. Source of Data

The only source of data collection considered under this study is the questionnaire developed using the current literature written about BDA. Therefore, the research has been carried out using primary source of data.

F. Instrumentation

In order to refine the survey tool; the questionnaire, expert opinion was obtained from both academic and industry experts and a pilot survey was initially done to check the applicability of the questions. The questionnaire consisted of two parts as mentioned below.

Part 1 consists of questions relating to demographics of the participants of the survey such as age, gender, employed function/department, employment tenure, level of employment, level of education and employed organization.

Part 2 comprises of questions which examine factors affecting the intention to adopt BDA and questions which assess the mediating effect of attitude towards using in the relationship between main determinants and intention to adopt BDA.

In the case of quantification of the content in the part 2 of the questionnaire, each sub variable is given a score within the range of 1 and 5, based on Likert scale (Table I) which was also used by Anke Schull and Natalia Maslan [40] in 2018.

TABLE I. LIKERT SCALE (1-5)

Scale	Criteria
1	Strongly Disagree
2	Disagree
3	Neutral
4	Agree
5	Strongly Agree

G. Collection of Data

Data with respect to the phenomenon; intention to adopt BDA in apparel sector of Sri Lanka, was collected in selecting individuals on random basis with the help of a survey questionnaire. Questionnaires were distributed among the sample selected using an online method. The current literature was gathered to support the information collected through the questionnaire.

H. Data Analysis Strategies with Justification

Data which was gathered through sharing a questionnaire among the selected individuals in 13 key players in the apparel sector was analyzed using IBM Statistical Package of Social Sciences (SPSS 23). In this case, techniques such as descriptive and inferential statistics were mainly used for describing and analyzing the collected data.

When comes to the further analysis of data, measures of central tendency such as mean, median and mode fall under descriptive statistics and measures such as regression and correlation analysis to assess the nature of relationships exist among the variables fall under inferential statistics were computed.

Following data analysis strategies were used when analyzing the data to assess whether the main objectives of the research achieved or not.

- To explore the determinants responsible for explaining the variation of attitude towards using big data analytics - Regression and Correlation Analysis.
- To identify the key influencing factors for the adoption of big data by organizations in apparel sector of Sri Lanka and to calculate the weight of each factor - Measures of Central Tendency.
- To examine the relationship among the variables; the attitude towards using big data analytics and the intention to adopt big data analytics - Regression and Correlation Analysis.
- To assess the mediation effect of attitude towards using on the direct relationship between critical factors and the intention to adopt of big data analytics – Sobel-Goodman Test.

VI. Analysis and Discussion

Initially, this section explains the sample overview. Then it depicts the results of the descriptive statistics in measuring the critical factors responsible for the behavior of the variable, attitude towards using BDA. Next, correlation analysis and multivariate regression analysis are performed to identify the relationships among the chosen variables/ determinants responsible for explaining the variation of attitude towards using BDA and the extent to which the attitude towards using BDA affecting the intention to adopt BDA. It will be followed by the multicollinearity testing to investigate whether there is a correlation among independent variables that would affect the regression results achieved. Further, to assess the mediating effect of attitude towards using on the direct relationship between critical factors and the intention to adopt of BDA, Sobel-Goodman test is used. Finally, the section is concluded with a brief summary of the results of the analysis performed.

A. Sample Overview

According to the Joint Apparel Association Forum (JAAF), Sri Lanka has already able to keep a record of 5.3 billion USD from apparel exports in 2019 while contributing 6% and 40% to Sri Lanka's GDP (Gross Domestic Production) and country's total exports respectively [41]. The behavior of the top 10 players in the apparel sector over a period of 7 years is graphically presented in the Fig. 2 and it highlights the apparel manufacturers; Brandix Apparel Limited, MAS Intimates (Pvt) Ltd, MAS Active Trading (Pvt) Ltd and Hirdaramani International Exports Ltd have operated above the annual revenue of USD 0.3 billion from 2014 onwards. In addition, MAS Holdings and Brandix Lanka have been nominated for ranking 1 and 2 with a score of 90.18 and 89.28 respectively in Export Corporate Brands 2020 [42].

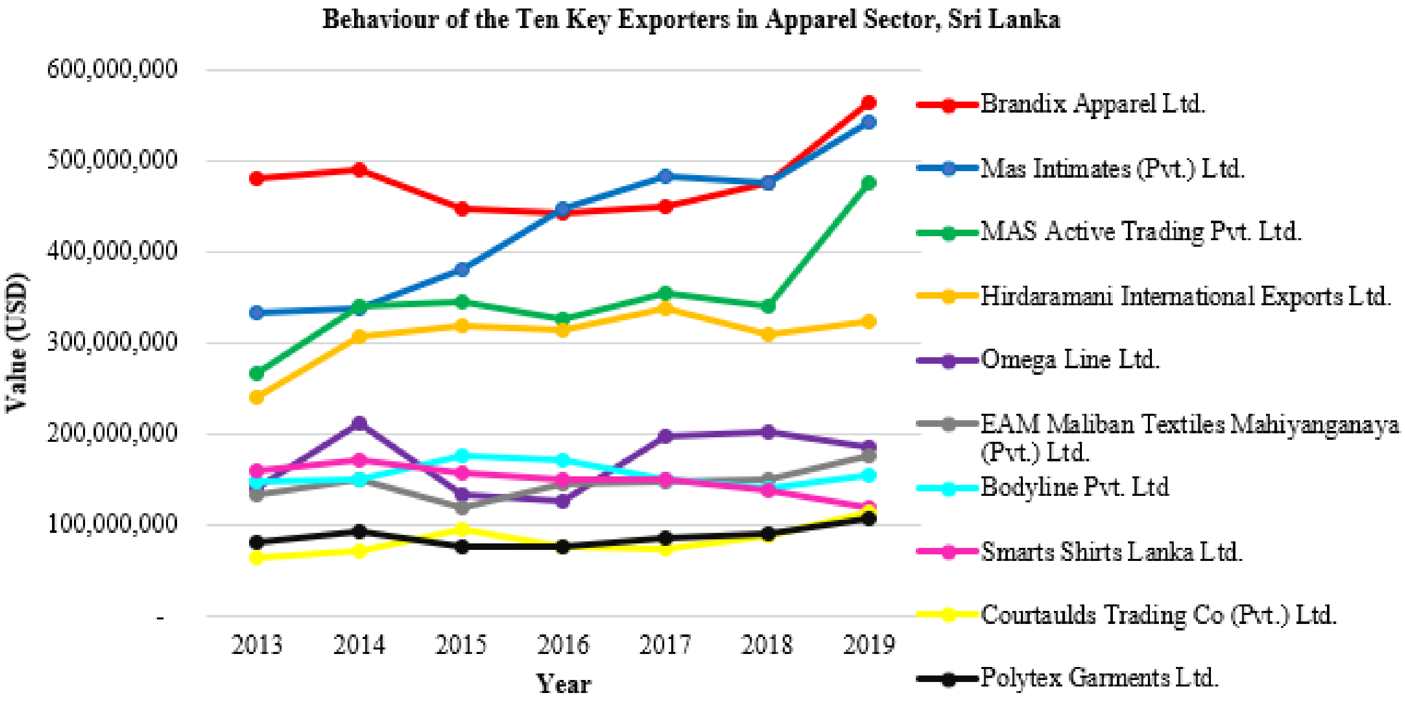


Fig 2. Sri Lanka Apparel Exports – (Board of Investments, Sri Lanka).

B. Demographics

The information reveals that majority of respondents represent the gender, male (60.8%) while female is 39.2%. When comes to the age, most of the participants are between 26 to 30 years (53.7%), with 18.9% from 31 to 35, 12.1% from 22 to 25, 9.3% from 36 to 40, 5.5% from 41 to 50 with a low minority (0.5%) above 50. Almost equal number of participants is shared among the functions finance (20.5%) and marketing (20%) while the rest are from operations (15.3%), human resources (9.6%), supply chain (8.8%), engineering (7.4%), information technology (6%), administration (2.2%), data analytics and other departments under general management falling into 10.1%. As per the work experience (tenure), majority represent 2 to 4 years (44.9%) while 22.7% was from more than 5 years, 14.2% from 4 to 5 years, 12.1% from 1 to 2 years and 6 % from less than 1 year. In terms of the roles perform in the given organizations, majority of the participants represents the executive level (45.5%) whereas the rest were from senior executives (21.4%), manager (14.5%), assistant manager (11%), general-manager (2.5%) and with a low minority from director, CXO & consultant (0.9%). Finally arriving at the last demographic variable; education, it denotes that the majority hold a bachelor’s degree (53.2%), while 36.2% consists with master’s degree, 6.3% has learnt up to high school and 4.4% holds a professional school degree/certificate.

C. Validity and Reliability

A reliability assessment was done using Cronbach Alpha. For all variables; technological (0.879), organizational (0.913), environmental (0.890), perceived ease of use (0.908), perceived usefulness (0.933), attitude towards using (0.921) and intention to adopt (0.929), Cronbach’s alpha coefficients showed a high level of reliability, ranging from 0.879 to 0.933, with a highest satisfactory value for the variable, intention to adopt BDA. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy for all variables was above 0.8 implied the adequacy of the sample and Bartlett's test of sphericity denoted the significance of each construct.

D. Descriptive Statistics

The generated descriptive statistics reveal that the mean value of the technological context amounts to a value of 4.136 with a median value of 4.2 while the standard deviation is 0.648 (15.6% of the mean value). The second independent variable, organizational context presents a mean and a median of 3.790 and 4.000 respectively while with a standard deviation of 0.623 (16.4% of the mean value). The third independent variable, environmental context denotes a mean value of 4.103 with a medium value of 4.2. It also consists with a standard deviation of 0.682 (16.6% of the mean value). Collected data also emphasizes a mean of 3.865 and a median of 4.2 along with the standard deviation of 0.793 which is 20.5% of mean value for the next independent variable, perceived ease of use. The final independent variable, perceived usefulness denotes a mean of value of 3.966 and a median value of 4.2 with a standard deviation of 0.792 (19.9% of the mean value). The collected data derives with a mean value of 4.037 and a median value of 4.166 along with a standard deviation of 0.706 (17.4% of the mean value) for the mediating variable, attitude towards using. Finally, the dependent variable, intention to adopt BDA denotes a mean value of 3.932 while the median value would be 4.166. The standard deviation of this dependent variable is 0.764 which is 19.4% of the mean value.

E. Regression Analysis

In this study, we examine the relationships among technological context, organizational context, environmental context, perceived ease of use, perceived usefulness, attitude towards using and intention to adopt BDA and R^2 and t-tests are used to identify the coefficients and relationships among variables.

1) Regression Analysis of Technological Context vs. Attitude Towards Using

In the testing of goodness of fit, technological context elucidates R^2 value of 0.533 on attitude towards using, which depicts an explanatory capability of 53.3%, along with a p-value of 0.000 less than 0.05, indicating that the attitude towards using is significantly and proportionally affected by the technological context. The coefficient is 0.730, which means that when increasing the technological context by one unit, the attitude towards using will increase by 0.730 units.

2) Regression Analysis of Organizational Context vs. Attitude Towards Using

Organizational context explains a R^2 value of 0.360 of attitude towards using which is having a considerable level of an explanatory power on the dependent variable as it explains 36%. The significance level which is below 0.05 reflects the variable; attitude towards of using will significantly be affected by the organizational context built in. The coefficient of 0.600 explains that if organizational context increases by 1 unit, then the attitude towards using will enhance by 0.600 units.

3) Regression Analysis of Environmental Context vs. Attitude Towards Using

R^2 value of 0.581 along with the p-value less than 0.05 reflect the significant explanatory power of the independent variable; environmental context over the dependent variable attitude towards using and this relationship further proved by the generated positive t-value. The coefficient 0.762 denotes if 1 unit of environmental context is enhanced, then it will lead to 0.762 units enhancement of attitude towards using.

4) Regression Analysis of Perceived Ease of Use vs. Attitude Towards Using

The relationship between perceived ease of use and attitude towards using generates a R^2 value of 0.543 which highlights that the 54.3% of attitude towards using is decided by perceived ease of use. The high explanatory power of this relationship is further confirmed by the p-value which is less than 0.05 and by the positive t-value. The coefficient value of 0.737 denotes if 1 unit of perceived ease of use is increased, it will lead to 0.737 units increase in attitude towards using.

5) Regression Analysis of Perceived Usefulness vs. Attitude Towards Using

The independent variable, perceived usefulness will decide the behavior of the Variable, attitude towards using by 62.2% as the regression statistic derives a R^2 value of 0.622. The power of explanation on the dependent variable, attitude towards using will be further strengthen by positive t-value generated along with a significance value of less than 0.05. The coefficient value of 0.789 reflects that if 1 unit of perceived usefulness is enhanced, it will lead to the enhancement of 0.789 units of attitude towards using.

6) Regression Analysis of Attitude towards using vs. Intention to Adopt Big Data Analytics

In this context, R^2 value of 0.639 denotes that 63.9% of the dependent variable is decided by the variable, attitude towards using. This relationship is further proved by positive t-value generated and the p-value which is below 0.05. The coefficient of 0.799 emphasizes that if 1 unit of attitude towards using is enhanced, it has the capability of increasing the dependent variable, intention to adopt BDA by 0.799 units.

7) Multiple Regression Analysis of Critical factors; Technological Context, Organizational Context, Environment Context, Perceived Ease of use and Perceived Usefulness vs. Attitude Towards Using

In the testing of goodness of fit, the technological context, organizational context, environmental context, perceived ease of use and perceived usefulness would be elucidating R^2 value of 0.735 of attitude towards using, which is having an explanatory power of 73.5%. The p-values of technological context, environmental context, perceived ease of use and perceived usefulness are lesser than 0.05 which have reached the significance level and the positive t values emphasize that attitude towards using would be significantly and proportionally affected by those independent variables.

However, the independent variable, organizational context shares a negative coefficient value of 0.091 along with a p value of 0.033. As the significance is below 0.05, it denotes that the variable, attitude towards using is known to be dependent on the independent variable, organizational context but the generated data doesn't have enough power to detect that dependence.

TABLE II. CRITICAL FACTORS VS. ATTITUDE TOWARDS USING

Model	Unstandardized Coefficients		Standard Coefficients	*t*	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	0.382	0.134		2.851	0.005		
Technological Context	0.204	0.052	0.187	3.935	0.000	0.327	3.061
Organizational Context	-0.103	0.048	-0.091	-2.135	0.033	0.409	2.443
Environmental Context	0.309	0.049	0.299	6.313	0.000	0.331	3.024
Perceived Ease of Use	0.186	0.043	0.209	4.290	0.000	0.312	3.203
Perceived Usefulness	0.306	0.043	0.343	7.095	0.000	0.316	3.166

Dependent Variable: Attitude towards using

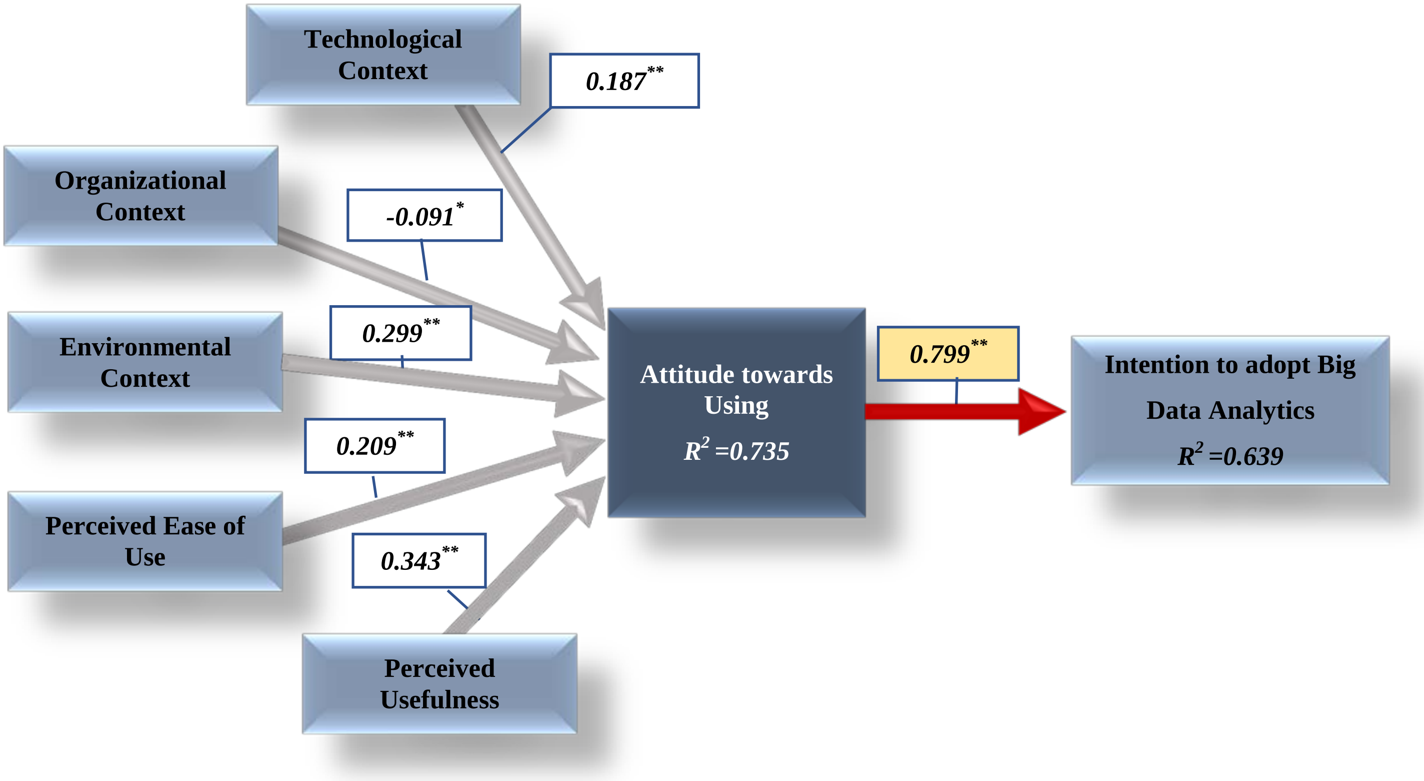


Fig 3. Research Results (*p<0.05, **p<0.01).

The coefficients of technological context (0.187), environmental context (0.299), perceived ease of use (0.209) and perceived usefulness (0.343) emphasize the stronger effect of those independent variables on the dependent variable, attitude towards using (Table II).

Summing up the above results, the regression results were graphically indicated in Fig. 3.

8) Multicollinearity among the Independent Variables

In this case to evaluate the multicollinearity effect of independent variables such as technological context, organizational context, environmental context, perceived ease of use and perceived usefulness, the statistical measurement; variance inflation factor (VIF) was used. Table II depicts the VIFs for all the independent variables in the model of this study which exist between 1 and 5 concluding an existence of a moderate correlation among the independent variables that emphasizes the strength of correlation between independent variables is not significant.

F. Hypothesis Testing

In this case, to test the proposed hypothesis, the results were interpreted using regression while taking into consideration the sample size. Table III will summarize the results.

1) Data-Related Infrastructure Capabilities

The results show a positive effect between data related infrastructure capabilities and the attitude of using BDA with a high significance level of lesser than 0.05 along with a positive t-value. This confirms enough evidences to reject the null hypothesis.

2) Data Management

The coefficient and p-value of H2 on the regression analysis are 0.711 and 0.000 (lesser than 0.05) which denote that there is a powerful positive relationship between data management and attitude towards using BDA. The hypothesis is further supported by positive t-value of 19.252. Researchers such as Malladi and Krishnan [43] and Kwon et al., [16] have also noted that how important it is to manage the challenges in data from both internal and external sources for BDA usage.

3) Privacy and Security

Privacy and security regarding data is having a coefficient of 0.542 which is strengthened by a p-value of less than 0.05 and a positive t-value. Therefore, the statistics support H3 and Gangwar [44] has also emphasized that the increase of privacy and security concerns will have a significant negative effect on big data adoption.

4) Vision and Strategy

The regression analysis for vision and strategy vs. attitude towards using BDA gave a p-value of 0.000 (lesser than 0.05) and a t-value of 13.935 which would the support the acceptance of H4. Therefore, the statistics results emphasize that a strategic business vision is most essential to make a change the way users think and to direct the BDA implementation effort.

5) Sponsorship and Governance

Sponsorship and governance are having a positive coefficient of 0.554 towards the attitude of using BDA. The positive relationship between these two variables is further confirmed by the p-value below 0.05 and the positive t-value of 12.687. Therefore, H5 hypothesis can be accepted and it highlights that user will never have a positive attitude unless it is backed by commitment of top management with a proper funding and governance mechanism.

6) Organizational Structure

Positive t-value of 11.981 and p-value of 0.000 which is lesser than 0.05 support the hypothesis which emphasizes that well established organizational structure will positively influence the attitude towards using BDA. These results indicate that organizational structure seems also affecting the user attitude towards BDA and Shiwei, et al., [4], has revealed that he has come across 22 frequencies through content analysis where it has been mentioned, "*organization has a well-organized structure that is well-suited to the adoption of big data*".

7) Talent Strategy & Capability Development

The regression analysis of H7 gave a p-value of 0.000 which is lesser than 0.05 along with a t-value of 13.562. It can thus be concluded that the null hypothesis for H7 should be rejected and that talent strategy & capability development have a positive effect on the users' attitude towards using BDA. This is further supported by Schüll & Maslan [40] mentioning a focus needs to be given on the skill development, when channeling the investments on BDA.

8) Firm Size

The regression analysis of H8 gives a p-value of 0.821 and therefore the null hypothesis is accepted, and H8 is not supported. It can be inferred that firm size does not influence the attitude towards using BDA. The findings of Gangwar [44], depicted that the firm size was playing a statistically significant role, but also emphasized that there wasn't exact size-fit relationship when comes to the organizational size and the adoption rate of big data.

9) Market Pressure

The effect of market influence on users' attitude of using BDA is positive and the hypothesis H9 is supported with a regression analysis giving a p-value of 0.000. Therefore, market pressure does have a positive as well a significant influence towards the users' attitudes of adopting BDA and this is further proven by, Lautenbach, et al., [45] agreeing with the view that BDA enables the organizations to gain competitive advantage over its rivals.

10)Big Data Pressure

The regression analysis of H10 gave a p-value of 0.000 and as a result, null hypothesis could be rejected. Therefore, it can be inferred that big data pressure does have a positive and a significant influence on the attitude towards using BDA.

Finally, hypothesis testing can be concluded with the hypotheses, H11, H12, H13 which have the highest positive influences which are confirmed by the positive r-values achieved as 0.789, 0.737 and 0.799 respectively. These positive influences can be nominated as significant relationships as the statistics derive positive t-values and p-values of 0.000 (less than 0.01) for all those hypotheses and all of them can be accepted while rejecting the null hypotheses (Table III).

TABLE III. HYPOTHESIS SUMMARY

Hypothesis	*r* value	*t* value	*p* value	Indicator
H1: Data-related infrastructure capabilities will positively influence attitude towards using big data analytics.	0.646	16.143	0.000	Supported
H2: Data management will positively influence attitude towards using big data analytics.	0.711	19.252	0.000	Supported
H3: Privacy & security will positively influence attitude towards using big data analytics.	0.542	12.297	0.000	Supported
H4: Vision and strategy will positively influence attitude towards using big data analytics.	0.590	13.935	0.000	Supported
H5: Sponsorship and governance will positively influence attitude towards using big data analytics.	0.554	12.687	0.000	Supported
H6: Well established organizational structure will positively influence attitude towards using big data analytics.	0.532	11.981	0.000	Supported
H7: Talent Strategy & capability Development will positively influence attitude towards using big data analytics.	0.580	13.562	0.000	Supported
H8: Firm size will positively influence attitude towards using big data analytics.	-0.012	-0.227	0.821	Not Supported
H9: Market pressure will positively influence attitude towards using big data analytics.	0.675	17.439	0.000	Supported
H10: Big data pressure will positively influence attitude towards using big data analytics.	0.755	21.924	0.000	Supported
H11: Perceived usefulness will positively influence attitude towards using big data analytics.	0.789	24.464	0.000	Supported
H12: Perceived ease of use will positively influence attitude towards using big data analytics.	0.737	20.776	0.000	Supported
H13: Attitude towards using big data analytics will positively influence the intention to adopt big data analytics.	0.799	25.344	0.000	Supported

G. Mediating Effect of Attitude towards Using on the Direct Relationship between Critical Factors and the Intention to Adopt Big Data Analytics

To identify the mediating effect of attitude towards using on the direct relationship between critical factors (technological context, organizational context, environmental context, perceived ease of use and perceived usefulness) and the intention to adopt BDA, Sobel-Goodman test was performed.

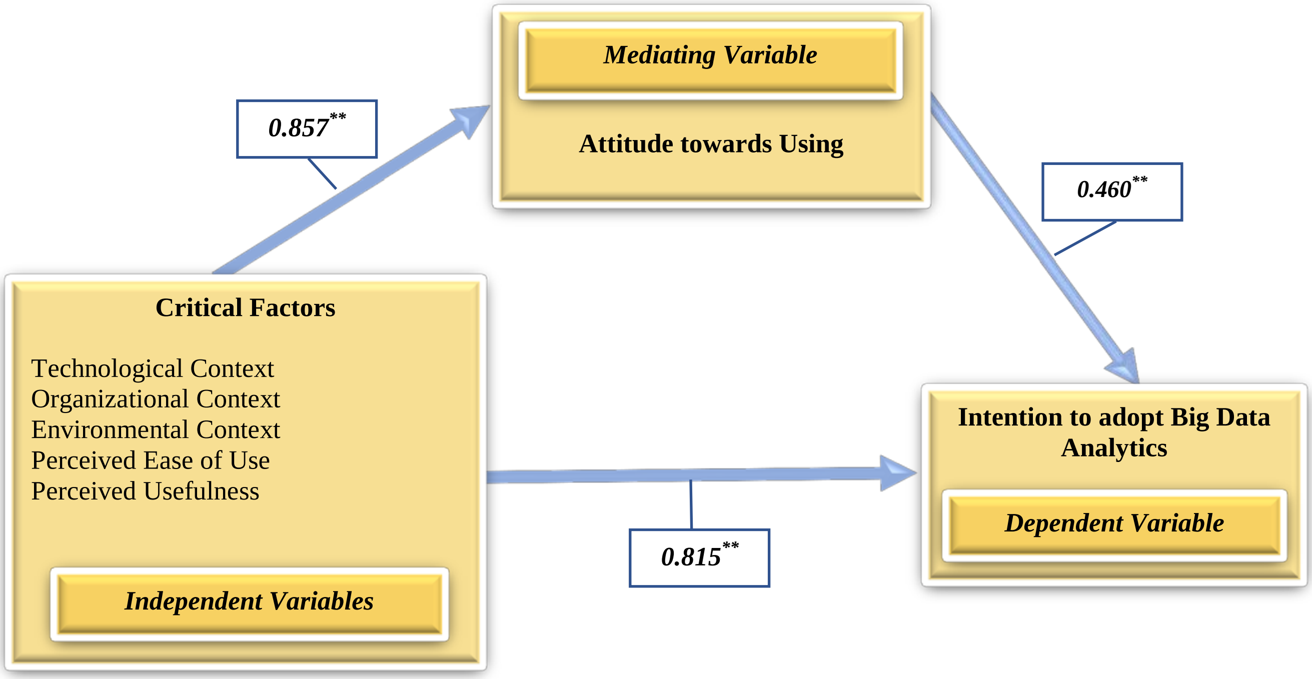


Fig 4. *R*esearch Results (Mediating Effect) (*p<0.05, **p<0.01).

TABLE IV. MEDIATOR ROLE OF ATTITUDE TOWARDS USING

Test Type	**Test Statistic:**	**Std. Error**	***p*-value**
Sobel-Goodman test	2.252	0.189	0.024

As per the Fig. 4, the variable, *"Attitude towards using"* could be considered a mediator to the extent to which it carries the influence of the given independent variables, *"Critical factors"* to the given dependent variable, *"Intention to adopt BDA"*. Overall, fulfillment of four criteria will assure the existence of a mediating effect and those will be (1) when there is a significant influence on the variable, attitude towards using from the independent variables; critical factors, (2) in the absence of the variable, attitude towards using, if the critical factors cause a significant impact on the behavior of intention to adopt BDA, (3) the mediator, itself having a significant and as well as a unique effect on the dependent variable, intention to adopt BDA and (4) the final criteria that the significance of the influencing power of critical factors on the dependent variable, intention to adopt BDA will shrink upon the addition of the variable, attitude towards using as a mediator to the model [46]. The results depict the successful achievement of the specified four criteria which emphasizes the mediating role performs by the variable, attitude towards using BDA. However, MacKinnon & Dwyer [47] have popularized statistically based methods by which mediation may be formally assessed. Table IV depicts the statistics derived from Sobel-Goodman test and it shows test statistic of 2.252 followed by a standard error worth of 0.189 with a p-value of 0.024 which is below the significance level 0.05. Therefore, possible to conclude that the variable, attitude towards using acts a mediator in between the independent variables - critical factors; technological context, organizational context, environmental context, environmental context, perceived ease of use and perceived usefulness and dependent variable - intention to adopt BDA.

H. Discussion

There are many challenges and barriers to the success of BDA adoption in apparel sector in the introductory phase. Previous studies have dealt with critical success factors for big data adoption. But there are limited studies which have analyzed how the both internal and external factors would affect for the users' attitudes and ultimately influence the intention to adopt BDA in apparel sector. This study is mainly woven around three main objectives and let's see how they have been achieved.

The first objective is to explore the key determinants responsible for explaining the variation of attitude towards using BDA and to calculate the weight of each factor. Based on the methodology and results presented, it emphasizes that the variables such as technological, and environmental factors as well as the behavioral factors such as perceived ease of use and perceived usefulness positively affects for the user' attitudes towards BDA adoption. All four factors derive a relationship with the variable, attitude towards using which performs statistically significant at a 0.01 level. However, the organizational context denotes a negatively correlated relationship with the variable, attitude towards using at a statistical significance level of 0.05. This can be mainly due to data generated are not supportive enough to explain the effect of the organizational context on the attitude towards using BDA. Lautenbach, et al., [45] have also confirmed that factors from each of the T (Data-related Infrastructure Capabilities, Data management challenges, Privacy & Security), O (Vision and strategy, Sponsorship and governance, Organizational structure, Talent Strategy & capability Development) and E (Market Pressure, Big Data Pressure) contexts of the TOE framework were significantly indicating that this framework has been appropriate for gaining insights into BDA adoption and usage at an organizational level. In addition, Brock V & Khan, [48] explained how the factors, perceived ease of use and perceived usefulness brought up by TAM framework significantly affected to the variation of the technology acceptance of BDA. When comes to the weight calculation of each of the determinants, statistical mean was used and high weight denoting independent variable could be introduced as technological context (4.136) while rest would be environmental context (4.103), perceived usefulness (3.966), perceived ease of use (3.865) and organizational context (3.790). Therefore, can summarize the first objective that the technological context, environmental context, perceived ease of use and perceived usefulness as key determinants which are responsible for explaining the variation of attitude towards using BDA while the explanatory power of the variable, organizational context remains at a minimum level.

Second objective was to examine the relationship among the variables; the attitude towards using BDA and the intention to adopt BDA. The results collected from a sample of 365 respondents confirm that there is a strong positive relationship between the variables attitude towards using and the intention to adopt BDA and it performs at a statistical significance of 0.01 which further confirms the explanatory power of the two variables. However, there is a dearth of studies which have analyzed the explanatory power of these two variables rather moving directly to the intention of adopting BDA from the critical factors. When comes to the real-world scenario, it is obvious that there is no way that critical factors could flow directly to intention to adopt BDA. As there is a human involvement in the adoption and implementation of BDA, it is important to give a focus on users' attitudes generated due to different external and internal factors which would ultimately decide the extent to which the user having the intention to adopt BDA. Jahangir & Begum [49], have revealed that in their study on *"The role of perceived usefulness, perceived ease of use, security and privacy, and customer attitude to engender customer adaptation in the context of electronic banking"*, that it is of paramount importance to ensure that people will actually use e-banking systems, as considerable amount of investment has been done in developing the system. Their study also suggests that in order to attract more users towards electronic banking, it is not going to be enough to merely introduce an e-banking system but also essential to develop the belief of usefulness of the system. Therefore, the findings emphasize the importance of assessing the relationship between attitude towards using and intention to adopt BDA as both BDA and e-banking systems are results of technology developments.

The third and the final objective of this study is to assess the mediation effect of attitude towards using on the direct relationship between critical factors and the intention to adopt BDA. The findings of this study, which has been measured using Sobel-Goodman test denotes that the variable, attitude towards using performs a role as a mediator on the direct relationship between critical factors and the intention to adopt BDA. However, there is a dearth of studies which assess the mediating role of attitude towards using between the direct relationship of critical factors and intention to adopt BDA. But Jahangir & Begum [49] in their study of customer adaptation in the context of electronic banking have contributed by providing support for the contention that customer attitude performs a mediating role in the link between perceived usefulness, ease of use, security and privacy, and customer adaptation. As both BDA and e-banking systems are outputs of technological innovations, it further proves that attitudes can perform a significant role as a mediator.

VII. Conclusion

This study examined how certain factors would influence the extent to which BDA could be adopted in Sri Lankan context. The assessment of extent to which organizations' willing to adopt BDA was done based on the factors in the TOE framework and TAM model which were proven as influential in prior studies.

Data & Analytics Report for the year 2017, released by MIT Sloan Management Review found that the percentage of organizations deriving competitive advantage from analytics is rising for the first time in four years [50]. It is found that the ability to innovate using analytics is driving resurgence of strategic benefits across all industries and this can be denoted as a finding arrived as a result of incorporating survey results and interviews with practitioners and scholars. The primary and as well as the main source of data for the MIT SMR's seventh annual analytics global survey has been the 2,602 responses received from business executives, managers, and analytics professionals from many organizations located all over the world [50]. In 2017, Lautenbach, et al., [45] confirmed that organization wished to be a differentiator among its rivals, needed to practice data-driven decision making which was supported by effective BDA usage in contrast to decision making based on intuition or gut feeling.

Technological context (Data-related infrastructure capabilities, Data management challenges & Privacy & Security), environmental context (Market Pressure & Big Data Pressure), perceived ease of use and perceived usefulness can be presented as the most essential ingredients which would intensively affect towards enhancement of positive attitudes of users which would ultimately lead towards the intention of adopting BDA among the Sri Lankan, Apparel Sector. However, when comes to organizational context, it doesn't reflect a positive correlation towards the attitude towards using BDA. But the components in the organization context such as vision & strategy, sponsorship and governance, organizational structure, talent strategy and development except firm size denoted a significant positive influence towards the attitude of using BDA when they were analyzed separately. Firm size didn't reflect a significant influence towards the attitude towards using and it could be due to generated data not enough to explain the dependence as well could be due to, there was no exact size-fit relationship as the adoption would mainly depend on the business requirements. Besides, the study presents evidences for a positive link between the variables; attitude towards using and intention to adopt BDA.

In addition to the presented perspectives of the critical factors, the study presented a new perspective of the mediator role performed by the variable, attitude towards using BDA on the direct relationship between critical factors and the intention to adopt BDA. This can be presented as a vital finding for the BDA practitioners and researchers. Finally, can conclude the study highlighting that the organizations which operate in apparel sector and which wish to promote data-driven decision making through greater use of BDA are specifically encouraged to focus on data-related infrastructure capabilities, vision and strategy, dynamic changes in consumer demands, transparency and understandability along with enhancement of effectiveness of the job roles of users. In return, it is expected that the implementation of BDA will lead to increase in organizational performance.

VIII. Limitations and Future Research Directions

This study inherits several limitations which are not able to fulfill due to many resource restrictions. First, the focus of this study is on key apparel sector organizations in Sri Lanka regarding to the intention to adopt BDA. In this case, impact of organizational culture is ignored which may influence the level of attitude towards using along with the intention to adopt BDA. Future researchers has the opportunity to test the same research model in other organizations considering different cultural setups which they are exposed to, because the organizational setups and cultures vary from industry to industry; therefore, the findings of this study likely to vary when applied to different sector organizations [39]. In addition, the study was performed among limited number of respondents which might affect to the broader generalization. Furthermore, the focus of the study is on the intention to adopt BDA in a developing country; Sri Lanka. Thus, testing this model in developed countries grants the opportunity in enhancing the generalization of the study, as the severity of resistance to change from employees is greater in developing countries than in developed countries [51]. Also, this study is focusing to investigate the user intention to adopt BDA which totally neglects the system implementation. Therefore, future researchers could identify the developers/architects' intentions in developing and implementing BDA systems. Another limitation is that the study doesn't provide an implementation road map with respect to business applications. But high-level references regarding individuals' perceptions and factors affecting the intention to adopt BDA are given. Finally, the study is based on cross-sectional settings which restricts the measurement of the consistency in respondent behavior and to remove this gap and to significantly contribute to the knowledge, the study needs to be performed in a longitudinal setup.

References

[1] G. George , M. Haas and A. Pentland, "Big data and management.," Acad Manage J., vol. 57, no. 2, p. 321–326, 2014.

[2] E. Brynjolfsson, L. M. Hitt and H. H. kim, "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?," 2011. [Online]. Available: [online] http://ssrn.com/abstract=1819486..

[3] L. Klie, "Getting closer to customers tops big data agenda.," CRM Mag., vol. 17, no. 1, pp. 15-15, 2013.

[4] S. Shiwei , G. Casey , Cegielskia, J. Lin and J. H. Dianne , "Understanding the Factors Affecting the Organizational Adoption of Big Data," JOURNAL OF COMPUTER INFORMATION SYSTEMS, 2016.

[5] L. G. Tornatzky and M. Fleischer, "The processes of technological innovation.," in Lexington, Mass: Lexington Books., 1990.

[6] "The future of Sri Lanka Apparel," 2018. [Online]. Available: http://www.ft.lk/apparel__fashion__design/The-future-of-Sri-Lanka-Apparel/10404-665483. [Accessed 26 October 2018].

[7] R. U. Kuruppu, "'New policy for textile and clothing must emerge'," 12 January 2020. [Online]. Available: http://www.sundayobserver.lk/2020/01/12/%E2%80%98new-policy-textile-and-clothing-must-emerge.

[8] "May will be a "tough month" – Brandix CEO," 2020. [Online]. Available: http://bizenglish.adaderana.lk/may-will-be-a-tough-month-brandix-ceo/. [Accessed 21 April 2020].

[9] S. Premaratne, "Analysing Sri Lankan lifestyles with data mining: two case studies of education and health," 2017.

[10] Gartner, "IT Glossary – Big Data," 2012. [Online]. Available: http://www.gartner.com/it-glossary/big-data.

[11] T. Lukoianova and V. Rubin, "Veracity Roadmap: Is Big Data Objective, Truthful and Credible?," Advances In Classification Research Online,, vol. 24, no. 1, pp. 4-15, 2014.

[12] M. Z. Elbashir, P. A. Collier and S. G. Sutton, "The role of organizational absorptive capacity in strategic use of business intelligence to support integrated management control systems," The Accounting Review, vol. 86, no. 1, pp. 155-184, 2011.

[13] D. Mungree, A. Rudra and D. Morien, "A framework for understanding the critical success factors of enterprise business intelligence implementation.," Chicago, 2013.

[14] K. R. Ramamurthy, A. Sen and A. P. Sinha, "An empirical investigation of the key determinants of data warehouse adoption," Decision Support Systems, vol. 44, no. 4, pp. 817-841, 2008.

[15] W. Yeoh and A. Koronios, "Critical success factors for business intelligence systems," Journal of Computer Information Systems, vol. 50, no. 3, pp. 23-32, 2010.

[16] O. Kwon, N. Lee and B. Shin, "Data quality management, data usage experience and acquisition intention of big data analytics," International Journal of Information Management, vol. 34, no. 3, pp. 387-394, 2014.

[17] N. A. M. Rafhi and S. A. Rahm, "Factors of Big Data Analytics in Enabling the Knowledge Management Practice," International Journal of Academic Research in Business and Social Sciences, vol. 7, no. 11, 2017.

[18] B. B. Jayasingh, M. R. Patra and D. B. Mahesh, "Security issues and challenges of big data analytics," pp. 204-208, 2016.

[19] W. Yeoh, A. Koronios and J. Gao, "Managing the Implementation of Business Intelligence Systems: A Critical Success Factors Framework," International Journal of Enterprise Information Systems, vol. 4, no. 3, 2008.

[20] H. Watson, D. Annino, B. Wixom, K. Avery and M. Rutherford, "Current practices in data warehousing.," Information Systems Management, vol. 18, no. 1, pp. 1-9, 2001.

[21] R. Bose and X. Luo, "Green It Adoption: A Process Management Approach.," International Journal Of Accounting and Information Management, vol. 20, no. 1, pp. 63-77, 2012.

[22] Y. Wang, Y. Wang and Y. Yang, "Understanding the determinants Of RFID adoption in the manufacturing industry.," Technological Forecasting and Social Change, vol. 77, no. 5, pp. 803-815, 2010.

[23] P. F. Hsu, K. L. Kraemer and D. Dunkle, "Determinants Of E-Business Use In Us Firms.," International Journal of Electronic Commerce, vol. 10, no. 4, pp. 9-45, 2006.

[24] K. Zhu, K. L. Kraemer and S. Xu, "The process of innovation assimilation by firms in different countries: A technology diffusion perspective on e-business," Management Science, vol. 52, no. 10, pp. 1557-1576, 2006.

[25] K. Zhu, S. T. Dong, S. X. Xu and K. L. Kraemer, "Innovation diffusion in global contexts: determinants of post-adoption digital transformation of European companies.," European Journal of Information Systems, vol. 15, no. 9, pp. 601-616, 2006b.

[26] K. Zhu and K. L. Kraemer, "Post-Adoption Variations In Usage and Value Of E-Business By Organizations: Cross-Country Evidence From The Retail Industry," Information Systems Research, vol. 16, no. 1, pp. 61-84, 2005.

[27] M. N. Masrek, A. Jamaludin and D. M. Hashim, "Determinants of strategic utilization of information systems: A conceptual framework," Journal of Software, vol. 4, no. 6, pp. 591-598, 2009.

[28] K. Agrawal, "Investigating the determinants of Big Data Analytics (BDA) adoption in Asian emerging economies.," 2015.

[29] F. D. Davis , "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS Q., vol. 13, no. 3, 1989.

[30] P. Guriting and N. O. Ndubisi , "Borneo online banking: evaluating customer perceptions and behavioural intention.," Manage. Res. News, vol. 29, no. (1/2), pp. 6-15, 2006.

[31] K. Eriksson, K. Kerem and D. Nilsson , "Customer acceptance of internet banking in Estonia," Int. J. Bank Mark, vol. 23, no. 2, pp. 200-216, 2005.

[32] S. A. Gahtani , "The applicability of TAM outside North America: an empirical test in the United Kingdom.," Info. Resour. Manage. J., pp. 37-46, 2001.

[33] K. Lancaster , "A new approach to consumer theory.," J. Polit. Econ., vol. 74, no. 2, pp. 132-157, 1966.

[34] H. C. Triandis , "Values, attitudes and interpersonal behaviour.," Unpublished paper, University of Nebraska Press, Lincoln, NE., 1979.

[35] J. Esteves and J. Curto , "A risk and benefits behavioral model to assess intentions to adopt big data.," J Intell Stud Bus., vol. 3, pp. 37-46, 2013.

[36] J. A. Castañeda , F. Muñoz-Leiva and T. Luque , "Web acceptance model (WAM): moderating effects of user experience.," Inf Manag., 2007.

[37] C. Brock , M. Blut , M. Linzmajer and B. Zimmer , "F-commerce and the crucial role of trust.," 2011.

[38] S. Muhammad , G. Changyuan , Z. LiLi , Fakhar Shahzad and H. Yanling , "Investigating the adoption of big data analytics in healthcare: the moderating role of resistance to change," Journal of Big Data, 2019.

[39] M. Shahbaz, C. Gao, L. Zhai, F. Shahzad and Y. Hu, "Investigating the adoption of big data analytics in healthcare: the moderating role of resistance to change," Journal of Big Data, 2019.

[40] A. Schüll and N. Maslan, "On the Adoption of Big Data Analytics: Interdependencies of Contextual Factors," 2018.

[41] Xinhua, "Asia&Pacific," 2020. [Online]. Available: http://www.xinhuanet.com/english/2020-01/22/c_138726334.htm. [Accessed 22 1 2020].

[42] "LMD The Voice of Business," 2020. [Online]. Available: https://lmd.lk/export-corporate-brands-2/.

[43] S. Malladi and M. Krishnan, "Determinants of Usage Variations of Business Intelligence & Analytics in Organizations – An Empirical Analysis.," Milan, 2013.

[44] H. Gangwar, "Understanding the Determinants of Big Data Adoption in India: An Analysis of the Manufacturing and Services Sectors," Information Resources Management Journal, vol. 31, no. 4, 2018.

[45] P. Lautenbach, K. Johnston and T. Adeniran-Ogundipe, "Factors influencing business intelligence and analytics usage extent in South African organisations," S.Afr.J.Bus.Manage, vol. 48, no. 3, 2017.

[46] K. J. Preacher and G. J. Leonardelli, "Calculation for the Sobel Test," 2001. [Online]. Available: http://quantpsy.org/sobel/sobel.htm.

[47] B. Azvine, Z. Cui and D. D. Nauck, "Towards real-time business intelligence," BT Technology Journal, vol. 23, no. 3, pp. 214-225, 2005.

[48] V. Brock V and H. U. Khan, "Big data analytics: does organizational factor matters impact technology acceptance?," J Big Data, 2017.

[49] N. Jahangir and N. Begum, "The role of perceived usefulness, perceived ease of use, security and privacy, and customer attitude to engender

customer adaptation in the context of electronic banking," African Journal of Business Management, vol. 2, no. 1, pp. 032-040, 2008.

[50] S. Ransbotham and D. Kiron, "Analytics as a Source of Business Innovation," MIT Sloan Management Review, 2017.

[51] M. Nejati, S. Rabiei and C. J. Chiappetta Jabbour , "Envisioning the invisible: understanding the synergy between green human resource management and green supply chain management in manufacturing firms in Iran in light of the moderating effect of employees' resistance to change.," 2017.

Distributed Denial of Service Attacks in Cloud Computing

Hesham Abusaimeh
Associate Professor in Computer Science
Middle East University, Amman, 11831 Jordan

***Abstract*—The Cloud Computing attacks have been increased since the expanded use of the cloud computing. One of the famous attacks that targets the cloud computing is the distributed denial of service (DDoS) attack. The common features and component of the cloud structure make it more reachable from this kind of attack. The DDOS is targeting the large number of devices connected in any cloud service provider based on its scalability and reliability features that make the cloud available from anywhere and anytime. This attack mainly generate a large number of malicious packets to make the targeted server busy dealing with these huge number of packets. There many techniques to defend the DDOS attack in the regular networks, while in the cloud computing this task is more complicated regarding the various characteristics of the cloud that make the defending process not an easy task. This paper will investigate most of the method used in detecting and preventing and then recover from the DDoS in the cloud computing environment.**

Keywords—Cloud; cloud computing; DoS attacks; DDoS attacks; DDoS prevention; DDoS mitigation

I. Introduction

Cloud computing features include availability at any time, network access, resource pooling, flexibility, and measured service. Availability means that cloud users can access and can manage their computing resources anytime, anywhere. Pooled resources mean Cloud users can use them from a range of computing resources if they need more resources to add to their existing cloud. Flexibility means that services can increase in size more or less. Moreover, the cloud user will only pay for their cloud resources.

Cybersecurity researchers are considering the attacks performed on the cloud as these attacks affect the budget, resource management, and quality of service they are providing. We provide a comprehensive classification of solutions to classify DDoS attack solutions, and to provide a comprehensive discussion of important measures to evaluate different solutions.

Many companies have adopted cloud computing due to its various features like on-demand service, wide network access, resource pooling, fast flexibility, and measured services. These features allow companies to look after their business operations, while the Cloud Service Provider (CSP) is managing the computing resources. Cloud model contoured to reduce business costs by making the installation of hardware and software updates and ensuring computing resources at the cloud service providers' side.

This paper is looking to provide information about DDoS attacks over the cloud environment. We also will try to distinguish between the types of various DDoS attacks, exploring and classifying the various contributions in this field. For this purpose, we prepared a detailed classification of these studies to assist to understand this survey.

A. DOS vs DDoS Attacks

When developing services on the cloud, safety must be taken into account critically. Some of the aspects that pose a challenge for cloud computing are:

Identity Authentication Authorization Confidentiality Integrity Isolation Availability

In a DDoS attack, hosts, such as robots or zombies, maybe a virtual machine, PC, or laptop. They have a remote-control feature. The use of a large number of hosts in an attack is called DDoS. More annoying DDoS than DoS. A group, hundreds of thousands of robots known as a botnet. The DDoS attack targets connection bandwidth and resources such as buffers, network protocols, or application processing logic.

II. Background

In this section, we will try to highlight the purpose or motivations behind some of the common DDoS attacks. However, many different categories can be identified to characterize the motivations behind DDoS attacks; the following is a summary of the purpose or motivations behind DDoS attacks.

- Financial or criminal benefit: This is classified as a motivation, and considers the most dangerous attack as the attackers try to get financial benefit by performing their attacks.
- Revenge: This type is classified as a motivation, as some frustrated individuals perform some of the attacks as payment of some injustice perceived.
- Ideological belief: Attackers performing this attack are motivated by their ideological beliefs.
- Intellectual challenge: Attackers perform DDoS attacks as a way to show off the capabilities and what they can harm as self-arrogant.
- Cyberwarfare: some well-trained military people or some of what is called terrorist individuals or organizations make this type of attack.

- Script Kiddies: New enthusiastic attackers who are trying some of the new tools on the Internet. They could use a bot-master that manages a network of bots or becomes part of Botnet that is involved in the attack.

- Hacktivists: They are recruited through social networking sites. They fight for the cause gives them a sense of purpose.

- Labor issue: The competitors may play a dirty game. You can launch DDoS attacks easily to the point of stop web services for competitors, by making use of the services of the attack Botnet provider. Botnet provider has fields to fill the target, and then a single click can be thousands of nodes begin to flood server's competition.

- Thrill attacks: Where the attacker so only to feel a sense of pride in this achievement.

Extortion or Ransom: Criminals who are willing to do anything malicious in exchange for money [1].

III. Types of DDoS Attacks

In this section, an analysis of the basic functioning of the DDoS attacks will be covered, and list the main types of DDoS attacks and provide a brief description of each type of attacks as the following.

- Direct Flood Attacks: Indirect flood attacks, an attack transfers a single package directly from its computer to the victim's site. It is the simplest type of DDoS attack.

- Remote Controlled Network Attacks: In these attacks instead of individual attackers like direct flood attacks, an attacker breaches a series of computers and places an application or proxy on computers.

- Flooding Attacks: These attacks generate the source of the IP packet source address with the victim's IP address and send it to an intermediate host whenever there is a response from the intermediate host; it is sent to the victim's destination address, and the victim is dumped [2].

- Worms: We can distinguish between a worm and a virus in the fact that the virus needs human intervention to inject a computer that the worm does not need. Worms can greatly disrupt the normal operation of the Internet.

- Viruses: Viruses have had a major impact on network providers. To structure a large zombie network, viruses are oftentimes used. In 1983 and 1984, serious Internet viruses included Melessia (1999), Love letter (2000), Nimda (2001 - a bunch of worms and viruses)[3].

- Fragmentation Attacks: Fragmentation Attacks have occurred on the firewalls of Cisco checkpoints and routers from Cisco and Windows PCs.

- Network infrastructure: Attacks targeting the network infrastructure can affect all Internet operations. Mostly, these types of attacks can create regional or global networks outside or slow down. A warning signal was sent to the root name server operators to reinforce the robustness of their infrastructure.

- Protocol violation attacks: In protocol violation attacks, the attacker originally sends packets. Attacks that generally use invalid or reserved IP protocols are protocol breach attacks. Protocol (255) is reserved, and protocols (135-254) are not allocated according to the specified online powers [4].

- Buffer Overflow Attack: where in this attack a large data is sent to the targeted buffer in a certain machine, and the size of this data is larger than the buffer size, which cause to save the data on a different buffer and remove the needed data, exist on that buffer.

- Email Bombing: where the inbox of a certain victim has been attacked with lots of emails.

- Ping of Death: the ping command is used to send a huge amount of data in the same packet while the received computer cannot accept and process this size of the data, which will slow the processing on this machine and reduce the connection between that computer and any other server.

- Smurf Attack: the ICMP protocol is used in this attack to obtain the same IP of the targeted machine and send back all the responses to the source machine with a larger bandwidth than the network bandwidth which is originally used.

- Synchronisation Flood: the attacker takes the advantage of the TCP protocol of starting the synchronisation process, which reserve a server for further data that should be sent after the synchronisation packets. While the attacker aim is to keep the server, busy with many Synchronisation packet and do not send any actual data after that.

- GET Flooding: the attacker in this attack generate many request packets using the HTTP protocol to a certain server that becomes busy with many GET messages from that client, and the server will also wait for the confirmation of these request, which never respond [5].

- Reflection Attacks: the attacker here use the UDP protocol to send many requests after spoofing the victim IP address [5].

- Amplification Attack: In this attack generating a large number of packets to target a victim website and use the DNS request after spoofing the source IP, address [5].

IV. Methodology

We conducted a set of literature by conducting a comprehensive search on previous papers and surveys and collecting a large number of papers related to the topic. The study results from the last papers we used. We believe that the contributions contained in this survey are comprehensive and include a list of all-important contributions in the field to date. In this paper, cloud security problems and some security mechanisms focus on eliminating and emphasizing them. Despite the need to reinforce existing security measures to provide more security in the cloud environment.

V. Related Work

Andrew Carlina (2017) a number of low load systems have been specifically proposed for WBANS. Based on a review of these modern methods, it is clear that the cloud model presents new security vulnerabilities. However, he began to monitor new cloud-based systems themselves to defend against widespread DDoS attacks. This enables systems to adopt scalability features to enhance the cloud for all parties. It is also necessary to think about safety models in terms of protecting individual clients and their services as well as the cloud as a whole, he said. To develop an effective defence system, aspects of these research systems must be combined to protect from a wide range of attacks. A number of these devices use VMs as system administration units. This allows these systems to take advantage of the flexibility and scalability of the cloud model to provide a more effective attack response and help reduce system bottlenecks [6].

It was concluded that there are two main research methods that must be followed. First, the hack tries to hack VMs to launch The DDoS attacks against a target outside the cloud. Although this may seem like a simplified solution for outbound systems, it is not a widely adopted solution by CSPs as it adds to their overheads, although indirectly jam their infrastructure. Second, developing more traditional cloud infiltration defence mechanisms target of the attack is the cloud or any part in the cloud itself.

Tasnuva Mahjabin (2017) in his survey, he provided a comprehensive and systematic analysis of DDoS attacks. It summarizes different types of attacks, filtering techniques, and methods for detecting attacks. However, his survey was an easy way to get the idea of DDoS attacks in which to systematically understand and analyse these attacks. Since his survey included recent attacks and recent researches against these attacks, it shows the current state of the attacks. It provides some discussions about DDoS attacks on unconventional systems such as clouds, smart grids, smart homes, CPS systems, and Internet of Things systems. As his study extracts an understanding of the search for DDoS attacks, it is important to understand the mechanisms for categorizing DDoS attacks in this survey. The author looks to analyse all the sorting of those attacks and provide an easy-to-understand classification mechanism.

Gaurav Somani (2017) in his survey provides a detailed survey on the DDoS attacks and its defence mechanisms on the cloud-computing environment. It has been demonstrated through discussion that a DDoS attack is the primary form of a DDoS attack in the cloud [7].

There is a number of solutions to DDOS attack such as attack detection and attacks mitigation. Among these solutions, a few contributions target specific cloud features such as allocating resources to demand resources reconfiguration employ SDNs. We also provide a broad list of performance metrics for these solution categories for assessment and comparison.

Seth Djane Kotey et al. (2016) concluded that with the increase in size, sophistication, and scale of modern DDoS attacks, further research is important to come up with very strong defences to fight these attacks.

Some DDoS defence types are discussed in this survey. They classified defence mechanisms according to their main functions: detection, tracking, and mitigation. They also discussed their strengths and weaknesses. It has been discovered that most solutions conflict with scalability and may not be able to perform well in the real world, due to the increasing size of attack and traffic robots involved in recent attacks. Most of the current solutions also added some additional computing and additional expenses to the network, which will have an impact on the network, and some of them may slow down, in a real scenario with large amounts of attack traffic [8].

A comparison was made between the different mechanisms; however, not all solutions had results for the criteria they have used in the comparison. For such mechanisms, it will be tested to determine their actual performance based on the metrics chosen by them. Overall, most of the defensive solutions reviewed were performing reasonably well.

Zargar et al. (2014) also provided an evaluation of the DDoS and OSI layer deployment mechanisms and defence system-based mechanisms: source-based, networked, and hybrid (hybrid) mechanisms. They also discussed the advantages, advantages, and disadvantages of defence mechanisms based on on-site deployment. In addition, the authors classified defence systems according to the time they start the process (before the attack or during the attack or after the attack). Then compared the performance of the defence mechanisms in accordance with the classifications they used. Following is a summary of the features, advantages, and disadvantages of defence mechanisms against DDoS flood attacks at the network/transport-level based on their deployment location.

For the source-based, the detection and the response is done at the source hosts directly and the pros of this is that it aims to detect and to respond to the attack traffic at the source before to wastes lots of resources, and the cons is that the sources are distributed at different domains, hence it is hard for each source or detect bad filter each attack in an accurate way, besides it is difficult to differentiate DDoS attacks at the source, since the traffic volume is low, as it is not clear who would pay for these services [9].

For the destination-based, the detection and response tools are installed on the victims hosts, they are easy to set up and cheaper compared to other tools in detecting DDoS attacks as they can access to aggregated traffic near the destination sources, but they cannot accurately detect and respond to attacks before it reaches to the victim and can waste resources while the attack is on its wait to the victim.

For the Network-based, the detection in response tools are deployed ate the network itself, one of its advantages is that it detects and responds to the attacks at the network and try to be closer to the attack source as it can, although some of its disadvantages is it needs high storage and the overhead that happens on the routers is difficult to detect because of lack of aggregated traffic destined for the victims.

In the hybrid model, the detection and the response tools are deployed at various locations; detections usually occurs at the victim side and the network, and the response usually takes place at the source and upstream routers near the source, the advantage of this Hybrid approach is that it is more robust against DDoS attacks and many resources can be used to stop the attacks, the disadvantages of this approach is the complexity and the overhead because of the communication that happens between many distributed components all over the internet.

VI. DDoS Protective Controls

In the survey conducted by Ahmed Bakr et al. (2019) prevention techniques are proactive, unlike detection and recovery that are reactive. Preventive controls must contain or eliminate the effects of a DDoS attack, and some techniques used to achieve this, following we will list some of these control and their techniques [10].

- Moving Target Defence (MTD): The Idea is rather than using layered defence by building static walls around your IT assets; it is working on making the attack surface dynamic.
- Completely Automated Public Turing Test (CAPTCHA): CAPTCHA is considered the most widely used prevention control by web applications. The shield can be used to protect web applications from malicious programs like Bots [11].
- EDoS-Shield Mitigation: This approach relies on implementing a front-end virtual firewall that maintains white and blacklists for IP addresses.
- sPoW (self-verifying Proof of Work): This approach relies on implementing a front-end virtual firewall that maintains white and blacklists for IP addresses.
- DNS based techniques: By blocking these malicious namespaces through ISP or web filtering agents, it will help with preventing launching bots to perform DDoS attacks [12].

As discussed previously, detection accuracy might be higher on the victim side but not powerful; victims cannot stand the large volume of DDoS traffic. Stopping attacks on the source can be the best option to respond, but it is very difficult since the amount of traffic in the sources is not important to distinguish the project from the harmful traffic. Moreover, side damage is high in midway networks due to insufficient memory and CPU cycles to determine traffic. Therefore, the central mechanisms in which all defence components (i.e. prevention, detection, and response) are deployed on the same site, are not practical against DDoS flood attacks.

VII. Defense against DDoS Attacks

Cloud Computing is now more targeted from the attackers by the DoS attacks. Solutions are being figured to deal with such risky attacks. Generally, guarding against DDoS attacks can be classified into three main categories: preventing attacks, detecting attacks, and responding to attacks as shown in Fig. 1 [13].

The cloud pool of resources will be blocked from the user access when a DoS attack is detected. In addition, even without detecting DoS attack the prevention scenario can also stop the user from accessing the cloud resources. The prevention techniques may install different protection components on all the cloud sites such as the user system, the network controllers, the internet routers, and track the attacker site [14].

Furthermore, the security prevention and detection methods can be placed on the VM of the cloud services, which includes all the hosted operating system installed. While this process will have some characteristics impact on the cloud service such as limited the processing capacity, reduce the network access, reduce the outsourcing dependency, limit the enhancement of the available protocols used in the cloud, reduce the network bandwidth with extra overhead protocols, and increase the power consumption of these units attached to the cloud services [17].

Since mitigating DDoS attacks is challenging, efforts should be increased to prevent these attacks from happening. Most of the cloud service providers establish a set of procedures to treat all the rules and actions that would be used by the customers as a policy requirement. These procedures are mandatory to follow the participation in the cloud service activities used. Table I shows an example of how the procedures are linked to the DDoS attack.

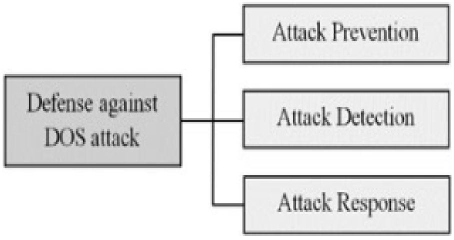


Fig. 1. Defense Type Against DoS Attack.

TABLE I. Procedures to Deal with the DDoS Attack [13]

DDoS attack	Procedure
Before the DDoS attack	To prevent DDoS attack the customer should use a firewall and prevent any unauthorized access and use proxies to connect to the hosts on the cloud pool.
During the DDoS attack	Stop most of the administrators' access to the network services and reduce the traffic on the cloud hosts.
After the DDoS attack	A team of administrator should act on recovering all the services and track the down services. In addition to document the type of the attack and the reason behind the attack and update the policy to prevent it in the future.

There are also other classifications of the DDoS attacks on the cloud and the detection or prevention techniques used for these types of attacks as summarized below [14].

1) Virtual machines level attacks: this kind of attack targeted the hypervisor layer in the virtual machines, and it needs an advanced cloud protection system is used that track the hosted virtual machines inside the hypervisor.

2) Resources Attack: this attack consumes all of the targeted system's resources by many pieces of data packet send and received from the attacker Screen OS.

3) BGP Prefix Hijacking: This kind of attack happens by flooding many announcement about fake IP addresses related to fake systems to attract certain users.

4) Port Scanning: this attack target the default and the unprotected ports like HTTP that are always open to provide web services; this can be prevented by, securing ports with encryption and using firewalls.

VIII. DDoS Attacks on SDN Overview

DDoS is increasing in the cloud computing environments due to the features of the cloud. With recent developments in Software Defined Networking (SDN), the SDN-based cloud provides is now a new opportunity to defeat DDoS attacks in cloud computing environments [15]. However, there is a contradictory relationship between SDN and DDoS attacks as shown in the different types in Table II. On the one hand SDN capabilities including existing traffic analysis software on the central control vision of the World Wide Web and the dynamic update for re-routing, make it easy to detect and respond to DDoS attacks. On the other hand, the SDN security itself remains to be addressed, and potential vulnerabilities in the DDoS system exist across SDN platforms.

In the following table Qiao Yan, F. Richard Yu published in their survey called "Software-Defined Networking (SDN) and Distributed Denial of Service (DDoS) Attacks in Cloud Computing Environments: A Survey, Some Research Issues, and Challenges" lists a comparison of DDOS attacks defence mechanisms using SDN [16].

TABLE II. Types of DDoS Attack on SDN

Types	SDN capabilities exploited	Description of the solution
Source-based Mechanisms Using SDN	Programmability	The programmable home network using the routing switches compatible with Open Flow and NOX as a controller detects security issues in SOHO.
	Traffic analysis	Discover the access point, which is a converter that supports Open Flow is whereby the console Open Flow controller, malware by analysing traffic in real-time.
	Traffic analysis, dynamic rules updating and global views.	Suggests VALVE that uses Open Flow protocol to resolve the problem of validating the source address with a view to improving the global SAVI solution.
Network-based Mechanisms Using SDN	Traffic analysis and centralized control	Use statistical information in the flow table to classify traffic as normal or harmful to self-organizing maps.
	Traffic analysis and dynamic rules updating	A novel content-oriented networking architecture (CONA) can react to DDoS attacks by the use of accountability and content-aware supervision.
	Programmability	And displays FRESCO, which is in the same application Open Flow limit. It can provide programming inspired by clicking frame.
	Abstraction ability	An agent-based framework, Agnos, has been introduced to build collaborative SDNs that extend beyond enterprise networks and built on the abstraction provided by SDN.
	Programmability and dynamic rules updating	An efficient memory system is proposed for distributed and collaborative monitoring of each stream called DCM. DCM uses Bloom filters to represent the monitoring rules and to install a custom and dynamic monitoring tool at the switch data level.
	Global views and centralized control	Abstraction suggests full of resources and the provision of anti-DDoS and alignment with the network operations to provide, manage and control the protection of DDoS as a service within the system of environmental Open Flow.
Destination-based Mechanisms Using SDN	Dynamic rules updating	It analyses the theoretical quantitative relationship between the probability that flow is successfully tracking number again jump level, and the probability of sampling independently, and the package number that includes the flow of the attack.
	Global views and centralized control	It was built Net Sight, which is an extendable platform that captures the history of packages and applications can retrieve from the date of packets are concise and flexible packages of interest.

IX. Conclusion

Attacks on the Cloud Computing components and software become a daily issue especially after the wide use of it in various applications. The demand to have a stable cloud services with high availability to be offered for all the kind of device PCs, Laptops, and mobile is an urgent issue. The DDoS attack is one of the simplest and high redundant attack in the Cloud Computing environment where it has different types that attack different cloud computing resources. The Distributed resources and the multi virtual platforms inside these distributed resources are the main vulnerability in the cloud computing services. This paper discussed the most kinds of the DDoS attacks that targeted the pool of resources in the cloud computing and give the most defending procedures that is used to prevent, detect and recover the tracks of the DDoS attack and its damage. This damage may cause to stop the cloud service and may consume losing the data stored in the cloud without any backup or replica. The main way to protect the cloud is to define a policy for using the cloud resources and make rules based on the statistics threshold of the previous use of that service.

Acknowledgment

The author is grateful to the Middle East University, Amman, Jordan for the financial support granted to cover the publication fee of this research article.

References

[1] Ryan K.L. Ko, Kim-Kwang Raymond Choo, "Cloud Security Ecosystem", Elsevier,pp.1-14, 2015

[2] Taghavi Zargar, S., "Towards Coordinated, Network-Wide Traffic Monitoring for Early Detection of DDoS Flooding Attacks", University of Pittsburgh, June 2014.

[3] Somani, G., Singh Gaur, M., Snaghi, D., and Conti, M., "DDoS attacks in cloud computing", the international Journal of Computer and Telecommunications Networking, Vol. 109, November 2016.

[4] Ludiora, S., Abiona, O., Oluwatope, A., et al., "A user identity management protocol for cloud computing paradigm", International Journal of Communication, Network, and System Science, Vol. 4, Issue 3, pp. 152–163, 2015.

[5] Han, J., &Kamber, M. (2018), "Data mining: Concepts and techniques (2ed Ed.)", Beijing: China Machine Press. DDoS Attacks and Impacts on Various Cloud Computing Component, 2018.

[6] Hammoudeh, M., Aldabbas, O., "Intrusion Detection and Countermeasure of Virtual Cloud Systems - State of the Art and Current Challenges", International Journal of Advanced Computer Science and Applications, Vol. 6, No. 6, 2015.

[7] Kalkan, K., Gur, G., and Alagoz, F., (2016), "Filtering-Based Defense Mechanisms Against DDoS Attacks: A Survey", IEEE Systems Journal PP(99):1-13 · September 2016.

[8] Masdari, M., Jalali, M. (2016), "A survey and taxonomy of DoS attacks in cloud computing", Security and Communication Networks, 13 July 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/sec.1539.

[9] Malomo, O., Rawat, D., and Garuba, M. (2017), "A Survey on Recent Advances in Cloud Computing Security", Journal of Next Generation Information Technology, Vol. 9, No. 1, March 2018.

[10] Bakr, A., Abd El-Aziz, A., and Hefny, H. (2019), "A Survey on Mitigation Techniques against DDoS Attacks on Cloud Computing Architecture", International Journal of Advanced Science and Technology, Vol.28, No. 12, 2019.

[11] Aldaej, A., "Information Security and Distributed Denial of Service Attacks : A Survey", the 2017 International Conference on Electrical and Computing Technologies and Application (ICECTA), UAE, November 2017.

[12] Zargar S., Joshi, J., and Tipper D., (2014), "A Survey of Defense Mechanisms against Distributed Denial of Service (DDoS) Flooding Attacks", IEEE communications surveys & tutorials, Vol. 15, No. 4, 2013.

[13] Alsowail S., Sqalli, M., Abu Amara, M., Baig, Z., and Salah, K., "An Experimental Evluation of the EDoS-Shield Mitigation Technique for Securing the Cloud". Arabian Journal for Science and Engineering, Vol. 41, No. 12, pp.5037– 5047, May 2016.

[14] Ubale, T., and Jain, A. K., "Survey on DDoS Attack Techniques and Solutions in Software-Defined Network", Springer International Publishing, 2020.

[15] Yan, Q., Yu, F.R., Gong, Q., Li, J., "Software-Defined Networking (SDN) and Distributed Denial of Service (DDoS) Attacks in Cloud Computing Environments: A Survey, Some Research Issues, and Challenges", IEEE Communications Surveys and Tutorials, Vol. 18, Issue 1, 2016.

[16] El-Sofany, H., El-Seoud, S., and Taj-Eddin, I., "Case Study of the Impact of Denial of Service Attacks in Cloud Applications", Journal of Communications, Vol. 14, No. 2, February 2019.

[17] Abusaimeh, H. and Yang, S.H., "Reducing the transmission and reception powers in the AODV", Proceedings of the 2009 IEEE International Conference on Networking, Sensing and Control, ICNSC 2009, pp. 60-65, 2009.

Analyzing the Performance of Web-services during Traffic Anomalies

Avneet Dhingra[1]
Research Scholar, Department of Computer Science and Engineering, I.K. Gujral Punjab Technical University
Kapurthala-144603, India

Monika Sachdeva[2]
Associate Professor, Department of Computer Science and Engineering, I.K. Gujral Punjab Technical University
Kapurthala-144603, India

***Abstract*—Intentional or unintentional, service denial leads to substantial economic and reputation losses to the users and the web-service provider. However, it is possible to take proper measures only if we understand and quantify the impact of such anomalies on the victim. In this paper, essential performance metrics distinguishing both transmission issues and application issues have been discussed and evaluated. The legitimate and attack traffic has been synthetically generated in hybrid testbed using open-source software tools. The experiment covers two scenarios, representing DDoS attacks and Flash Events, with varying attack strengths to analyze the impact of anomalies on the server and the network. It has been demonstrated that as the traffic surges, response time increases, and the performance of the target web-server degrades. The performance of the server and the network is measured using various network level, application level, and aggregate level metrics, including throughput, average response time, number of legitimate active connections and percentage of failed transactions.**

Keywords—Denial of service; DDoS attack; flash event; performance metrics; throughput; response time

I. Introduction

In the event of network traffic anomaly, the users get to grips with either a drastic slowdown of the service or a complete outage. Recent years have witnessed a rise in the frequency and strength of some illegitimate anomalies known as DDoS attacks. These attacks compromise the availability of the web-services of the victim server. The motive behind such activity varies from being personal to political. Whatever the cause, these attacks can be very troublesome and costly for a target. For instance, the online encyclopedia, Wikipedia, suffered a DDoS attack on September 6, 2019, that lasted for about three days [1]. The intermittent outages and performance degradation were faced by users in the Middle East, Europe, the United Kingdom, and the United States. Many such instances are confronted daily, across the globe, due to an exponential increase in the use of Internet-based applications. As per the Kaspersky report, the number of attacks has increased by 80% in the first three months of 2020 against the attacks observed in 2019 [2].

The need hence arises to generate realistic techniques to evaluate the performance and measure the impact of anomalies (legitimate or illegitimate) on the services of the web-server. Measuring the performance of the server under such anomalies can help understand the preventive techniques required to be installed along with the type of potential defenses. The importance of performance testing is also realized in the situation when multiple users generate concurrent traffic creating a heavy load on the network similar to that created during a DDoS attack. The network responding to anomalies needs to be tested repetitively with short-duration attack traffic to evaluate the overall performance of the server and the cost involved for installing the required security measures. The metric, thus calculated, provides the information related to the network traffic in case of saturation.

The literature reviewed [3,4,5] for the impact analysis highlights the use of a simulator for generating traffic and analyzing the performance of the network. However, the experiment presented in this paper makes use of emulation to generate synthetic traffic. Emulation has the advantage of using real-time OS and apps, along with the simulated elements such as virtual nodes and soft network links. Exploits. The paper also presents the exhaustive review undertaken to comprehend the concept of performance and quantifying the impact of anomalies on the web-services. Various application-level, as well as network-level and server-level performance metrics, have been identified and evaluated using the synthetically generated traffic in DDoSTB hybrid testbed [6]. The results of the study have been presented as graphs showing the effect of traffic surges and realize their impact on performance. The background traffic is mainly composed of TCP protocol. The attack traffic is composed of UDP, with varying packets per second. The HTTP traffic is generated with varying percentage of requests per second and represents the flash event (a legitimate anomaly). The paper defines performance metrics quantifying the quality of service (QoS) of the web server during normal conditions and under the increased traffic load. The experimental set-up and procedure to evaluate performance metrics of the designed network have been discussed.

The paper has been organized as follows. The related literature has been reviewed in Section 2. Section 3 gives an overview of what performance metrics are and its importance in the detection of anomalies. Section 4 describes the model of an experimental network using realistic topology and software tools used to generate legitimate and attack traffic. Section 5 discusses the metrics selected for analysis, and the results obtained are presented as graphs for better understanding. The paper concludes the observations of the experiment in Section 6. The scope for future work in the same field has been mentioned in section 7.

II. Literature Review

Researchers have studied the damage caused due to the poor performance of the victim server and suggested specific damage models. The models give an insight into the situation, the risks involved, and highlight the importance of analyzing and evaluating the performance of the system. One such model proposed by [7], divides the financial damage into characteristics like disaster recovery, loss due to downtime, liabilities involved, and losing the customers. The model can be generalized to any traffic anomaly affecting the availability of the server (like FE) and hence degrading the performance.

Vasudevan et al. [8] recommended metrics based on the cost of losing customers and the cost of SLA violations, which is from the point of view of the network users. It draws attention to the possible financial impact a DDoS attack can have on the network-provider of the affected network. The MIDAS2007NET factor has been defined based on the fact that allocated bandwidth is proportional to the volumes of traffic on the network and which in turn are proportional to the related profits.

Gade et al. [9] studied the impact of the DoS Land (Local Area Network Denial) attack to compare the memory and processor utilization during the attack. Chertov et al. [10] concluded that simulated networks produced different results from the emulated ones. The difference is because of the assumptions taken for simulation. The authors suggest the use of testbeds for accurate results. The metrics computed to measure performance are the average goodput (computed using transfer size and time for completion of transfer), size of the congestion window (calculated by dividing the average of weighted congestion window by average of segment size), percentage of CPU utilization and packets sent (and received) per second.

Mirkovic et al. [11,12,13,14] defined the various legacy metrics along with the applications where these metrics give accurate results. The metrics discussed are packet loss, throughput, request/response delay, transaction duration, and allocation of resources. Packets lost in the transit after an anomaly hits the network is termed as packet loss. In this case, a network could be either a direct hit network or nearby networks experiencing collateral damage. It measures the network congestion caused by flooding attacks. Throughput is the number of bytes transmitted or re-transmitted per time-interval. Goodput is the same as throughput with the difference that re-transmitted bytes are not counted. This metric, however, does not give accurate results for connections with few packets as the throughput, in this case, is already low.

Request/response delay metric cannot be applied to non-interactive applications and one-way traffic. Transaction duration captures the time required for exchanging a set of messages between a source and destination. The metric efficiently measures the services for the interactive applications (example-browsing internet) but fails to give results for one-way traffic. The allocation of resources is in proportion to a critical shared resource, like bandwidth. The ratio of resources allocated to legitimate traffic versus the attack traffic captures the service quality as viewed by the user. It measures the server load but fails to determine the collateral damage caused to the network. The authors categorized the applications into five groups: interactive applications (web-based), media applications (audio-video streaming), online games, chat applications, and non-interactive applications(email). Metrics vary across these applications and cannot be generalized. They proposed a few impact metrics keeping in mind the QoS requirements of different applications. One of them being a percentage of failed transactions (pft), which quantifies the quality of services experienced by the users. Another metric, DoS-hist, shows the resilience of an application to an attack with the help of histograms for pft.

Singh et al. [3], has applied application layer metrics and network layer metrics on the trace generated using the NS-2 simulator. The author speculates that the network level and transport-level parameters are not sufficient by themselves to detect application-layer attacks. They have checked the performance of the network for various GET-HTTP DDoS attacks.

Sachdeva et al. [4] have defined the ratio of average serve rate and average request rate to check the performance. This ratio is 1 for normal traffic and decreases as the strength of attack increases. These are evaluated using the NS-2 simulator. Sachdeva et al [15] evaluate the DDoS and FE impact on web services using DETER testbed. The metric throughput was evaluated as goodput and badput. Badput is defined as the attack traffic over bottleneck link during a given time window. Authors have experimented with traffic for response time, percentage of request packets lost, legitimate packet survival ratio, percentage of failed transactions, and bottleneck bandwidth utilization (for goodput).

Bhatia et al. [16] recommended the performance evaluation of network using server-level metrics viz, system-level CPU utilization, user-level CPU utilization, CPU load, and real memory utilization. The authors have proposed a framework for generating realistic traffic for the anomalies under study, using minimal hardware. Apart from the legacy metrics, Behal et al. [6] have used sever load metrics – CPU utilization, memory utilization, and CPU load to test the performance of the server. The researcher has developed DDoSTB testbed to generate the required synthetic traffic for experimentation.

Bhandari et al. [17] consolidate the metrics used for evaluating the performance of the defense framework and classifies them at a packet level, aggregate level, and application-level based on the level of the network they are used for. The author has also discussed the system parameters affected in the case of DDoS attacks and the different tools to measure the same.

Performance metrics have been classified as external metrics and internal metrics [6, 11, 17, 18]. The metrics that do not need privileged access to the system or its network are termed are external—for instance, attack strength or defense parameters. The internal metrics measure CPU load, CPU utilization, memory utilization, internal algorithms, and data structures. Another valid criterion for classification is the OSI layer involved in measuring the metric [6, 18]. The metric may measure the aggregate performance of networks such as throughput, or it may work on application level like transaction duration or at a packet level, such as a number of re-transmissions [4, 17].

III. Performance Metrics

As each network has a different topology and varies in purpose, different factors define the performance for each environment. Performance metrics quantify the QoS provided by the server, during normal conditions, and under the increased traffic load. The performance of a network depends on the following factors:

- The rate at which the information can be or is transferred,
- The delay between the request sent, and the response received,
- Error in transmission.

These factors are quantified with generated metrics, keeping in view the user requirements for a particular application. Mirkovic et al. [12] have explicated the response times of various kinds of applications and concluded that there is no particular set of metrics that can be considered as the benchmark for measurement of performance. The interactive applications are expected to respond in minimum time. For such applications, including VoIP, video streaming, chatting, and gaming applications, the delay is expected to be the minimum. The applications, such as email transfer, which are non-interactive, do not have any defined or tolerable limit for the delay. Also, there is no specific threshold or baseline to define the best performance. To efficiently examine the impact of anomalies in network traffic, metrics need to be accurate, quantitative, and versatile [11]. Ideally, this can be achieved using realistic networks, giving a holistic view of the network parameters. However, it is practically not possible to have such an expansive view due to specific economic, legal, and privacy issues. Thus, techniques like simulation or emulation are used to create a virtual network that can generate realistic synthetic traffic.

Table I outlines various network level, application level, and aggregate level metrics evaluated in this paper. As per [3], network level, transport level, and application-level metrics need to be assessed together to evaluate the overall performance of the network.

TABLE I. Performance Metrics

Level	Metric
Aggregate	Throughput (Goodput, Badput)
Application	Response time Number of legitimate requests dropped Number of legitimate active connections Percentage of failed transactions Average serve rate/ Average request rate Legitimate packet drop probability
Network	Percentage of link utilization
Server	CPU-utilization

IV. Experimental Set-up

The set-up of an experiment and its procedure includes three components - network topology, legitimate traffic, and attack traffic [14]. The best option is to set up an experimental environment in a real-time operational network that handles live traffic. However, the drawback involved in a real-time environment is that it cannot be reconfigured or scaled as per the experiments' needs. Thus, other options of simulated environment or emulated environment can be taken into consideration. Out of the two, simulation and emulation, the former lacks realism, and traffic replay is slow. Thus, the latter is the preferred method for testing by the researchers [3, 6, 13, 15, 19]. Emulator, for instance - DETER, is the combination of real hardware plus simulator to achieve the desired topology of a network. It provides a more stable environment as compared to a theoretical model or simulation alone. Soft routers and soft network links are used with the real systems real applications. However, the emulator is scalable only to a certain extent. Identifying the compatible tool along with the data source and its deployment on the systems can, somehow, be a challenge.

In this section, the performance of the testbed network has been carefully examined using the three components discussed in [14]. Flooding attack is generated using the realistic topology on the available testbed with the help of available and compatible software tools.

A. Network Topology

For the experimentation, authors used the hybrid DDoSTB Testbed, developed by Behal et al. [6,20], consisting of real as well as emulated systems. 45 physical systems have been deployed and grouped into 3 clusters of 15 computers each. Out of the 3 clusters, 2 clusters are specified for background traffic and 1 cluster for attack traffic. The systems are run on Windows XP and Ubuntu instances. To increase the nodes, the v-core emulator [6] is used for attack cluster that is, cluster 3. One v-core node implements 30 virtual nodes. This scales the network to (15 x 30) 450 attack nodes. Each node in cluster 1 and cluster 2 is deployed with 30 instances of Apache JMeter to scale the number of legitimate nodes to a total of 900 nodes. For making identification of source easier, different pools of IP-address are assigned to the users in each cluster. The victim web server gives access to the resources to the users from both legitimate and attack clusters. It records every request attempt as a log file. The topology deployed on DDoSTB is shown in Fig. 1, and the associated parameters have been summed up in Table II. Generally, the typical link speeds are used to assign link bandwidths. In Fig. 1, the link between R1 and server acts as the bottleneck link. The bandwidth of this link is 100 Mbps, with a delay of 5ms. The rest of the links in the topology have a bandwidth of 1 Gbps. Topology for the experiment is finalized considering the principles of benchmarking, as suggested in [21]. An attempt has been made to keep it realistic and comparable to the topology of the Internet.

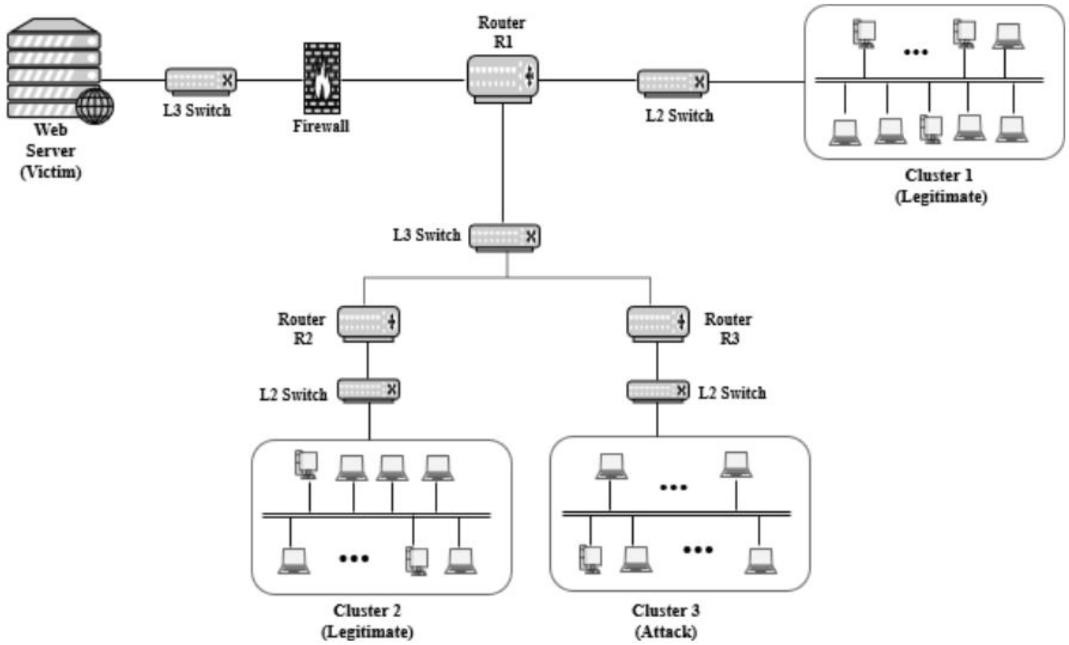


Fig. 1. Network Topology for Experimentation.

TABLE II. TOPOLOGY PARAMETERS

Parameter	Value
Number of legitimate nodes	2 clusters times 15 nodes times 30 VMs = 900
Number of attack nodes	1 cluster times 15 nodes times 30 VMs = 450
Number of routers	3 (R1, R2, R3)
Number of switches	5 (2 L3 Switch, 3 L2 Switch)
Bandwidth from R1 to web Server (Bottleneck Link)	100 Mbps
Delay of a link from R1 to a web server	5 ms

B. *Generating Traffic Traces*

The network traffic is generally a blend of two protocols: Transmission Control Protocol (TCP) and User Datagram Protocol (UDP); working at layer 4 of the OSI model. HTTP is an underlying generic protocol used by the World Wide Web to transfer data to and from the client and server. It operates in the application layer and relies on TCP in the transport layer for establishing a connection between the client and server. For the successful connection, TCP requires a valid IP-address. Each command is executed independently without the knowledge of the commands before and after it. GET and POST are two types of HTTP requests used for applications for which transmission time is not very critical but at the same time need reliability. These commands are not dependent on the commands that precede them or follow them. UDP, on the other hand, is a connectionless protocol as it does not require any virtual connection to be established before any data transfer. It is used for fast, efficient transmission like games and VoIP. It enables process-to-process communication by sending messages, called datagrams. It is efficiently used for applications that are loss tolerating and require low latency.

In the absence of real-time datasets, the traffic is generated synthetically to evaluate the system [6,19]. To achieve realistic trace, the client machines (nodes) interact with applications on a victim server with a non-spoofed and broad spectrum of source IP-addresses. Various open-source traffic generators are considered and studied to obtain a manageable mix of normal traffic and attack traffic for the experiment performed. Finally, Apache JMeter and D-ITG were chosen for generating background traffic and attack traffic.

Apache JMeter [1] is a pure Java open-source software designed to measure the performance of web services. It is a multi-platform Java desktop application. It has a friendly and easy to use Graphical User Interface (GUI). The results can be visualized as a log file and using a chart, table, or tree. Multiple virtual users can be created to generate heavy load against the victim server. JMeter supports all basic protocols, such as HTTP and FTP. Therefore, JMeter has been the choice of researchers of late [20, 22]. For the experiment described in this paper, the same version of JMeter is loaded on all the systems in cluster 1 and cluster 2. The distributed testing in JMeter requires one master, number of slaves, and one target machine, as shown in Fig. 2. One of the nodes in cluster 1 is configured as a master. The GUI runs on master and controls the rest of the nodes in cluster 1 and 2. Slaves run JMeter-server and take commands from the GUI and send requests to the target system. In short, each cluster has 15 nodes, and 30 virtual machines (VMs are defined as users in JMeter) are added hierarchically in the topology. Thus, the number of users generating traffic is scaled up to 900. The ramp-up time is set to 10 seconds so that each VM sends 3 requests per second. Similarly, the master-slave model is configured in cluster 3 with JMeter to generate HTTP traffic as attack traffic. Users in each cluster are assigned IP-addresses from different pools to distinguish between legitimate and illegitimate users [23].

[1] https://jmeter.apache.org/

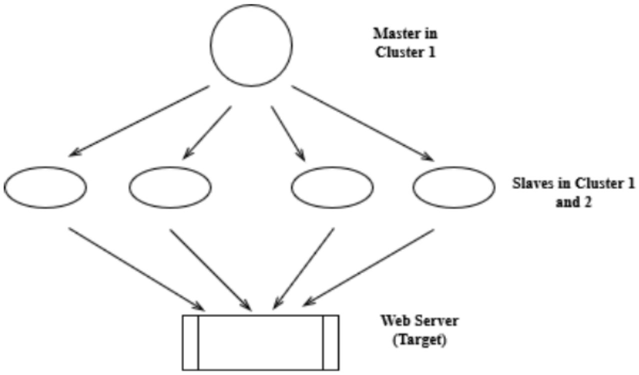


Fig. 2. Master- Slave Model.

Distributed Internet Traffic Generator (D-ITG) is a software platform capable of producing traffic that accurately adheres to patterns defined by the inter-departure time (IDT) between packets and packet size stochastic processes [24]. It supports both Linux and windows-based OS. It generates traffic having layer 3, layer 4, as well as layer 7 features. It emulates the sources of protocols like TCP, UDP, ICMP, DNS, Telnet, and VoIP [6, 24, 25]. The header fields like packet size, source, and destination IP-address, source, and destination port numbers can be customized as per the requirements. Due to its wide-ranging features, it has been widely used in research to generate synthetic attack traffic [8, 25] and hence, is the choice for generating attack traffic for the experiment mentioned in this paper. Cluster 3, as shown in Fig. 1, contains 15 nodes and 30 VMs on each node. Thus, attack traffic is generated from 450 users in cluster 3.

V. Results and Discussions

The performance of the victim server has been analyzed under two experimental set-ups. In the first set-up, the UDP traffic (representing DDoS attack) is generated as attack traffic in cluster 3 using the D-ITG tool and mixed with the normal traffic from 60th second to 120th second of the experiment targeting the victim server. Traffic is studied with IDT of 50 pps, 100 pps, and 200 pps.

In the second set-up, the HTTP traffic (representing FE) is generated in cluster 3 using Apache JMeter from 60th second to 120th second of the experiment. The attack traffic generated is mixed uniformly with the legitimate traffic generated by cluster 1 and 2 and targeted towards the victim server. Performance is observed for three instances- when attack traffic is 10%(approx.), 30%(approx.) and 75%(approx.) of total traffic. Parameters of experimental set-ups 1 and 2 have been outlined in Table III. The performance metrics used for analyzing the impact of an attack and FE are discussed:

A. Throughput

It is the amount of data transferred (in packets, in bits or bytes) in the unit time interval. It gives insight to the congestion of the network. Goodput is computed as the number of bits per second of legitimate traffic received at the server. Higher the goodput, the higher is the efficiency of the system being tested. These can be expressed as:

Throughput=Total traffic reaching server /Time-interval

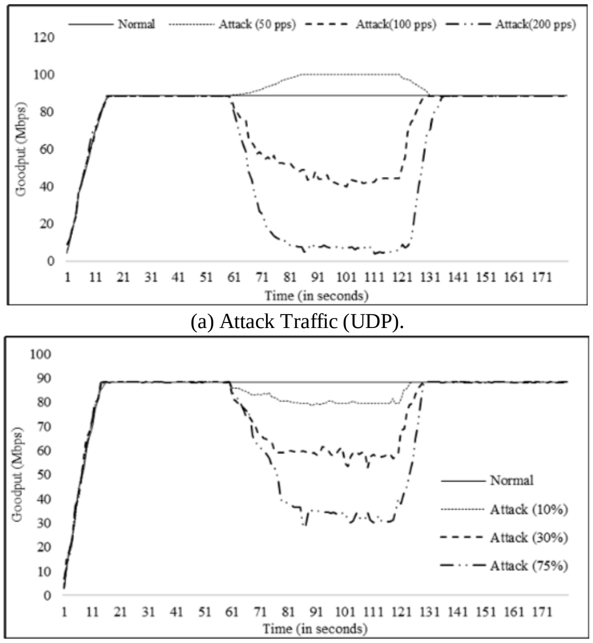
Goodput=Legitimate traffic reaching server/ Time interval

Fig. 3(a) and Fig. 3(b) show that for normal traffic, goodput increases slowly in the beginning until it either reaches the maximum traffic or the bandwidth limit before sit stabilizes. The slow start can be attributed to the congestion control strategy used by TCP. The transmission rate is increased by the slow-start algorithm until either a loss is detected or the receiver server's bottleneck link bandwidth is reached. In case loss occurs, the algorithm assumes that the network is congested and takes measures to reduce load. Otherwise, goodput increases exponentially until it reaches the bandwidth limit.

The traffic through the bottleneck link increases suddenly during the 60th second. As the attack strength increases, the value of goodput decreases. This is because as TCP senses packet loss, it decreases the transmission rate and hence decreasing the goodput. Hence, it is concluded that as soon as the traffic increases, whether UDP or HTTP, the attack traffic reaching the server increases. This rise plummets the goodput of the server to almost 7 Mbps.

B. Average Response Time

It is defined as the time between the request being sent from the client and receiving the first response [4]. It is also known as average latency [3]. This attribute is directly proportional to the amount of congestion in the network—lower the value of response time, less the congestion, and vice versa. In [15], the authors suggest that in the case of HTTP transactions, it is essential that the response time is less than 10 seconds to involve the users in the service and make the transaction successful. If the request crosses that limit, it is considered to be failed.



(a) Attack Traffic (UDP).

(b) Flash Event (HTTP).

Fig. 3. Goodput.

TABLE III. PARAMETERS OF EXPERIMENTAL SETUP-1 AND SETUP-2

Parameter	**Tool**	**Value**
Background/Legitimate traffic	JMeter	2 clusters x15 nodes x30 VMs = 900
Experiment time		180 seconds
Attack duration		60 seconds (from 60th sec to 120th sec)
Legitimate requests generated		Cluster 1: 15 x 30 x 3 = 1350 requests/sec Cluster 2: 15 x 30 x 3 = 1350 requests/sec Total = 2700 requests/sec
Packet Size		4096 bytes
Set-up 1:		
Attack traffic	D-ITG	UDP
Attack type		Constant rate
Packet size		512 bytes
Attack generated		Cluster 3: At 50pps per user Attack traffic = 50 x 450 = 22500 pps = 92.16 Mbps At 100 pps per user Attack traffic = 100 x 450 = 45000 pps = 184.32 Mbps At 200 pps per user Attack traffic = 200 x 450 = 90000 pps = 368.64 Mbps
Set-up 2:		
FE traffic	JMeter	HTTP
Traffic type		Uniformly distributed
Packet Size		500 bytes- 1000 bytes
Traffic generated		Cluster 3: For 10% FE traffic: VMs =10, 2 request/sec for each VM Total requests/sec = 15 x10 x 2 = 300 10% (approx.) of total traffic reaching server For 30% attack traffic: VMs =30, 6 request/sec for each VM Total requests/sec = 15 x 30 x 6 = 1200 30% (approx.) of total traffic reaching server For 75% attack traffic: VMs =50, 10 request/sec for each VM Total requests/sec = 15 x 50 x10 = 7500 75% (approx.) of total traffic reaching server

Average response time can be computed as

$$Response\ time = \frac{\sum(Tc + Td + Ts)}{N}$$

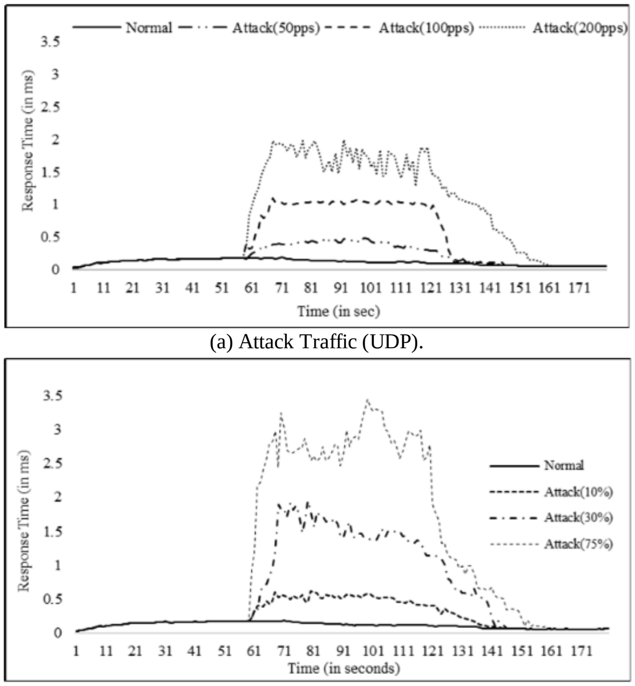
Where T_c is time for request sent from client to server, T_s is time for response sent from server to client, T_d is the time taken by the server to process the request, and N the number of time-intervals. Fig. 4(a) and Fig. 4(b) show an increase in response time with an increase in strength of attack in case of UDP traffic and HTTP traffic, respectively. HTTP traffic has a higher response time as compared to UDP traffic. The reason for the same is that in UDP, a connection is not formed between user and server, a datagram is just sent. Thus, UDP is faster than TCP, where the next packet is sent only once the acknowledgment is received for the previous one leading to the wait time and hence, an increase in the response time.

C. Number of Legitimate Requests Dropped

The number of requests dropped, due to an attack, measures the amount of congestion in the bottleneck link. For experiment 1, as UDP traffic increases, the congestion of the bottleneck link leads to the dropping of a large number of requests to the server. Fig. 5 shows the scenario of experiment setup-1, where the number of legitimate requests dropped increases with an increase in the attack strength.

D. Number of Legitimate Active Connections

Clients who have successfully connected themselves to the server and started sending the data are considered active connections. In the case of TCP connections, the active connections complete the 3-way handshake. When the attack packets of type HTTP surge the traffic, several packets endure time-out and hence reduce the window-size to almost one. According to the slow-start algorithm, the network reduces the load on the server by dropping the requests. Thus, the number of active connections decreases with an increase in the traffic beyond the capacity of the bottleneck link. The number of connections can reach as low as 90%, as suggested in Fig. 6. It can be observed that a large number of legitimate clients are denied services as the strength of attack increases. Thus, with a decrease in the number of active connections, the percentage of failed transactions increases.



(a) Attack Traffic (UDP).

(b) Flash Event (HTTP).

Fig. 4. Response Time.

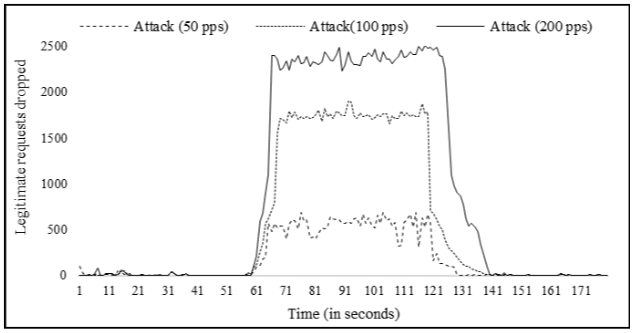


Fig. 5. Number of Legitimate Requests Dropped.

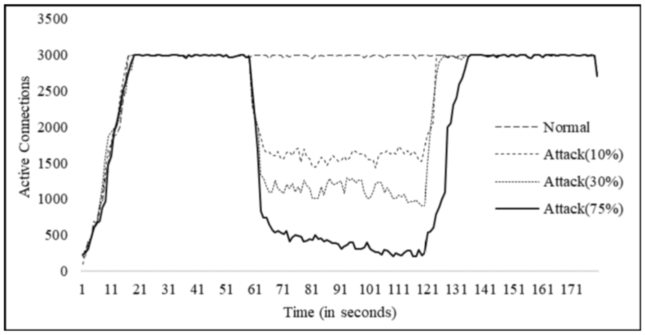


Fig. 6. Number of Legitimate Active Connection.

E. *Percentage of Failed Transactions*

It gives an insight to a number of requests timed-out or not being served by the server for whatever reason. In case the user sends requests using HTTP, the transaction is complete only if the response is received within the defined time (3-way handshake). When traffic increases during the 60th second, the large number of transactions fail due to congested bottleneck link. This decreases the throughput and also increases the response time, increasing the percentage of transactions that fail. It is directly proportional to the attack strength and can be represented in the equation as

$$\%age\ of\ failed\ transactions = \frac{T_{sent} - R_{recvd}}{T_{sent}} \times 100$$

where R_{recvd} is the number of responses received, T_{sent} is the total transactions sent.

F. *Average Serve Rate/ Average Request Rate*

It is the ratio of the rate at which the server serves the requests to rate at which the request is generated. The value of 1 indicates that all the requests generated are being served. As the strength of the attack increases, the ratio decreases. Higher the attack strength, the lesser the ratio. Fig. 7(a) and Fig. 7(b) show the pattern the ratio follows when under load in case of UDP traffic and HTTP traffic, respectively.

G. *Percentage of Link Utilization*

It is the percentage of bandwidth link used by legitimate requests. It can be computed as

$$\%age\ of\ link\ utilization = \frac{BW_{used}}{BW_{Total}} \times 100$$

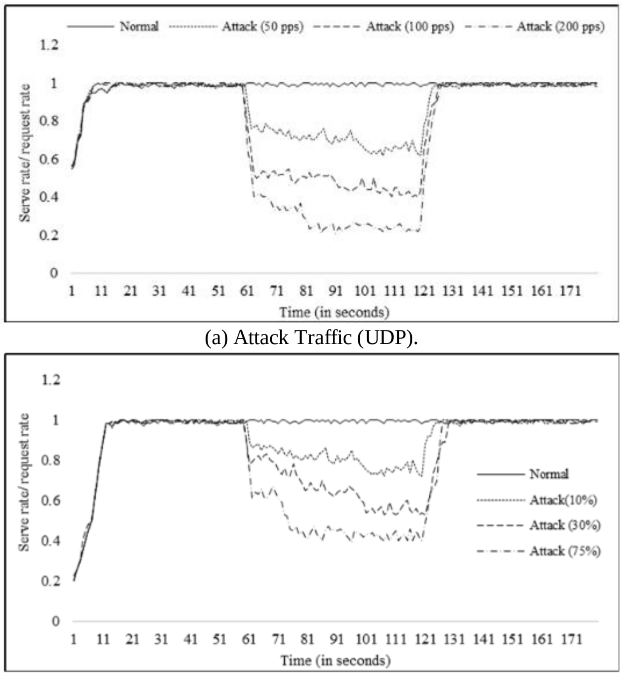
where BW is the link bandwidth. Fig. 8(a) and Fig. 8(b) show the link utilization (LU) in the case of UDP traffic and HTTP traffic, respectively. LU is 100% in case of normal conditions, but as soon as the traffic increases, LU reduces as a lesser number of legitimate requests reach the server. Legitimate traffic follows a congestion control protocol, so when the bandwidth gets clogged, the legitimate packets are dropped to decongest the bandwidth in the case of FE (HTTP traffic).

H. *Legitimate Packet Drop Probability*

A packet-level metric that compares the number of legitimate packets dropped with the total number of legitimate packets in the network. During normal traffic conditions, many dropped legitimate packets is negligible, making the ratio value to be zero. With an increase in traffic, the number of dropped legitimate packets increases, thus, increasing the ratio. Fig. 9(a) and Fig. 9(b) show the response of the server for packet drop in the case of DDoS and FE, respectively. As can be seen, the performance of the server degrades with an increase in traffic, whether UDP or HTTP.

I. *CPU Utilization*

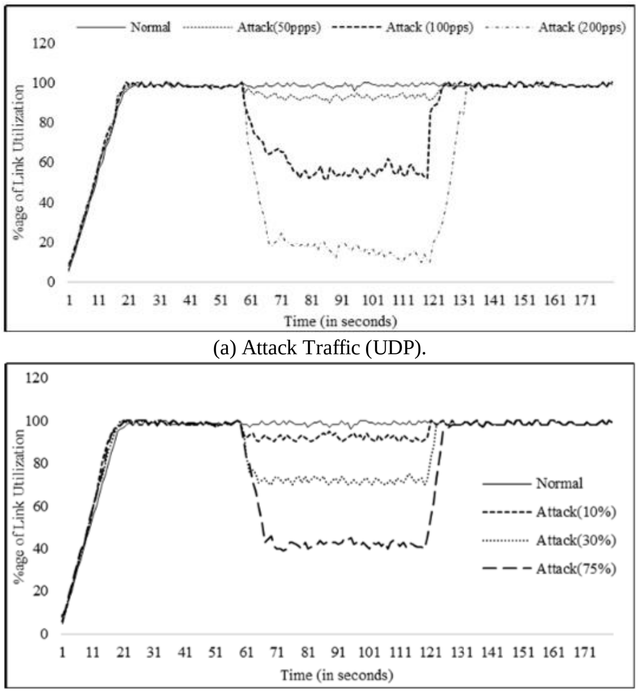
It is the server level metric that quantifies the utilization of CPU of the victim server. As the traffic increases in case of an anomaly, the total utilization of CPU increases though variations are seen for different applications. In the case of UDP attack traffic as in experiment 1, CPU utilization increases with an increase in strength of the attack, as shown in Fig. 10(a). If the attack is layer 7 attack (HTTP), the level of CPU utilization is more as compared to scenario-1, as shown in Fig. 10(b). This is because the request made to some running application has to be processed by a victim server. This leads to higher CPU utilization as compared to the UDP attack, where the requests are just dropped.



(a) Attack Traffic (UDP).

(b) Flash Event (HTTP).

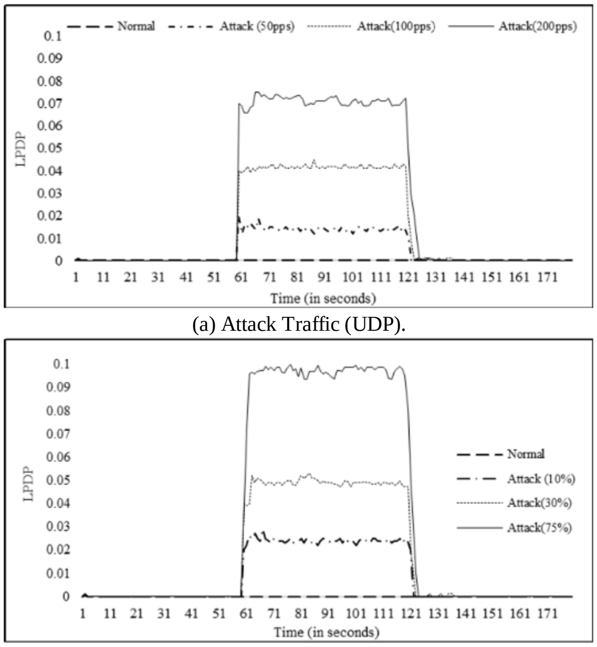
Fig. 7. Average Serve Rate/ Average Request Rate.



(a) Attack Traffic (UDP).

(b) Flash Event (HTTP).

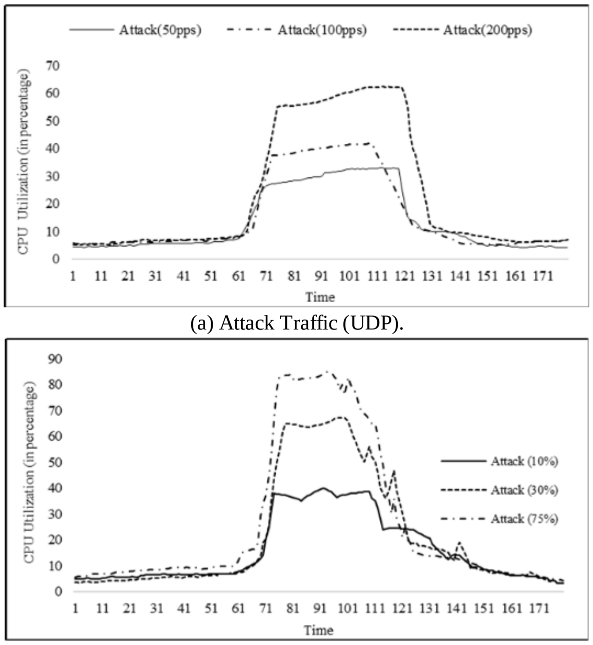
Fig. 8. Link Utilization.



(a) Attack Traffic (UDP).

(b) Flash Event (HTTP).

Fig. 9. Legitimate Packet Drop Probability.



(a) Attack Traffic (UDP).

(b) Flash Event (HTTP).

Fig. 10. CPU Utilization.

VI. Conclusions

The paper attempts to define and evaluate the performance metrics for the network, assuming that the traffic consists mainly of TCP, HTTP, and UDP protocols. The impact metrics have been quantified using throughput, response time, number of active connections, percentage of failed transactions, percentage of link utilization, serve rate/request rate, and legitimate packet drop probability. The experiment was done under two set-ups built within a hybrid testbed. The first set-up creates the scenario of a DDoS attack, and the other creates the

FE effect. Each traffic is generated with varying strengths and has been analyzed for the system with a realistic topology using the testbed. The analysis of the impact indicated that the performance of the system degraded with the surge in traffic due to anomalies. As the number of requests increased, the link bandwidth choked, CPU utilization increased, the number of legitimate active connections decreased, legitimate requests reaching the server decreased, thus, increasing the response time. An increase in response time degraded the services. However, it is worth mentioning that the performance is also affected by the hardware (server speed, HDD, the available RAM) used at the server. The response time gets affected by the increase in throughput as well as the number of intermediate points through which the user and server are connected. The user's experience is affected by a change in response time, especially the commercial websites where performance degrades with an increase in response time.

VII. Future Scope

Each network environment uses different metrics for evaluating performance. Also, the metric chosen reveals the users' scenario when service is denied due to the attack. Such a metric compares the values with the baseline values defined during normal conditions of a particular network. There is a need to form a shared repository of the metrics being used for various applications and to generalize the performance measures. Future work is focused on explicating these metrics with the baseline model and also defining the detection metrics.

Acknowledgment

The authors would like to acknowledge I. K. Gujral Punjab Technical University, Kapurthala, India, for the great support and assistance rendered to carry out this research work.

References

[1] S. Newman, "DDoS attack on Wikipedia site," Available at: https://www.corero.com/blog/934-ddos-attack-on-wikipedia-site-smacks-of-hacktivism.html.

[2] O. Kupreev, E. Badovskya, and A. Gutnikov, "DDoS attacks in Q1 2020," Available online: https://securelist.com/ddos-attacks-in-q1-2020/96837/.

[3] K. Singh, K. K. Saluja, and P. Singh, "Impact analysis of application layer DDoS attacks on web services: a simulation study," Intl. J. of Intl. Engg. Informatics, vol. 5, no. 1, 2017, pp. 80-100, https://doi.org/10.1504/IJIEI.2017.10003432.

[4] M. Sachdeva, K. Kumar, G. Singh, and K. Singh, "Performance analysis of web service under DDoS attacks," Proc. IEEE Intl. Advance Comput. Conf., IEEE, March 2009, pp. 1002–1007, https://doi.org/10.1109/IADCC.2009.4809152

[5] M. Kumar, and A. Bhandari, "Performance evaluation of web server's request queue against AL-DDoS attacks in NS-2." International Journal of Information Security and Privacy, vol. 11, no.4, 2017, pp. 29-46. https://doi.org/10.4018/IJISP.2017100103.

[6] S. Behal, and K. Kumar, "Measuring the impact of DDoS attacks on web services - A Real-time experimentation," Intl. J. of Comp. Sci. Info. Security (IJCSIS), vol. 14, no. 9, 2016.

[7] T. Dubendorfer, A. Wagner, and B. Plattner, "An economic damage model for large-scale Internet attacks," Proc. 13th IEEE Intl. Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, IEEE Comput. Soc, 2004, pp. 223–228, https://doi.org/10.1109/ENABL.2004.11

[8] R. Vasudevan, Z. M. Mao, O. Spatscheck, and J. Van der Merwe, "MIDAS: An impact scale for DDoS attacks," Proc. 15th IEEE Workshop on Local and Metropolitan Area Networks, Princeton, NJ, 2007, pp. 200-205. https://doi.org/10.1109/LANMAN.2007.4295999

[9] R. S. R. Gade, H. Vellalacheruvu, and S. Kumar, "Performance of windows XP, windows vista and Apple's leopard computers under a Denial of Service Attack," Proc. Fourth International Conference on Digital Society, IEEE, February 2010, pp. 188–191, https://doi.org/10.1109/ICDS.2010.39

[10] R. Chertov, S. Fahmy, and N. B. Shroff, "Emulation versus simulation: A case study TCP-targeted denial of service attacks," Proc. 2nd International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities, 2006, TRIDENTCOM 2006, pp. 316–325, https://doi.org/10.1109/TRIDNT.2006.1649164.

[11] J. Mirkovic, A. Hussain, S. Fahmy, P. Reiher, and R. K. Thomas, "Accurately measuring the denial of service in simulation and testbed experiments," IEEE Trans. on Dependable and Secure Comput. vol. 6, no. 2, 2009, pp. 81-95. https://doi.org/10.1109/TDSC.2008.73.

[12] J. Mirkovic, A. Hussain, B. Wilson, S. Fahmy, P. Reiher et al, "Towards user-centric metrics for denial-of-service measurement," Proc. 2007 workshop on experimental computer science, San Diego, California, 2007. https://doi.org/10.1145/1281700.1281708.

[13] J. Mirkovic, S. Fahmy, P. Reiher, and R. K. Thomas, "How to test DoS defenses," Proc. 2009 Cybersecurity Applications and Technology Conference for Homeland Security, March 2009, pp. 103–117. https://doi.org/10.1109/CATCH.2009.23.

[14] J. Mirkovic, S. Wei, A. Hussain, B. Wilson, R. Thomas et al, "DDoS benchmarks and experimenter's workbench for the DETER Testbed," Proc. 3rd International Conference on Testbeds and Research Infrastructure for the Development of Networks and Communities, Lake Buena Vista, FL, 2007, pp.1-7, https://doi.org/10.1109/TRIDENTCOM.2007.4444680.

[15] M. Sachdeva, G. Singh, and K. Kumar, "An emulation based impact analysis of DDoS attacks on web services during flash events," Proc. 2nd International Conference on Computer and Communication Technology (ICCCT-2011), pp 479–484, 2011. https://doi.org/10.1109/ICCCT.2011.6075134.

[16] S. Bhatia, "Ensemble-based model for DDoS attack detection and flash event separation," Proc. Future Technologies Conference 2016, San Francisco, US, 2016, pp. 958-967, https://doi.org/10.1109/FTC.2016.7821720.

[17] A. Bhandari, A.L. Sangal, and K. Kumar, "Performance metrics for defense framework against distributed denial of service attacks," Intl. J. of Netw. Security, vol. 6, April 2014.

[18] C. Bannwart, "Predicting the impact of denial of service attacks," Master thesis submitted to ETH, Zurich, Semantics scholar, 2012.

[19] S. Bhatia, D. Schmidt, G. Mohay, and A. Tickle, "A framework for generating realistic traffic for Distributed Denial-of-Service attacks and flash events," Comp. and Sec. vol. 40, 2014, pp. 95-107. https://doi.org/10.1016/j.cose.2013.11.005.

[20] K. Singh, P. Singh, and K. Kumar, "User behavior analytics-based classification of application layer HTTP-GET flood attacks. J. Netw. Comput. Appl. vol. 112, C (June 2018), 97–114. https://doi.org/10.1016/j.jnca.2018.03.030K.

[21] J.T.J. Midgley, "The linux HTTP benchmarking HOWTO," Available online: http://www.xenoclast.org/doc/benchmark/ HTTPbenchmarking-HOWTO/node3.html, 2001.

[22] M. A. Saleh, and A. A. Manaf, "A novel protective framework for defeating HTTP-based denial of service and distributed denial of service attacks," Sci. World Journal, 2015, https://doi.org/10.1155/2015/238230.

[23] Cisco Prime Network User Guide, 4.2.3, https://www.cisco.com/c/en/us/td/docs/net_mgmt/prime/network/4-2-/user/guide/CiscoPrimeNetwork423UserGuide/ip-pool.pdf.

[24] A. Avallone, S. Guadagno, D. Emma, A. Pescape, and G. Ventre, "D-ITG- distributed internet traffic generator," Proc. First Intl. Conf. on the Quantitative Evaluation of Systems (QEST '04), IEEE, Computer Society, 2004. https://doi.org/10.1109/QEST.2004.1348045.

[25] A. Botta, A. Dainotti, and A. Pescapè, "A tool for the generation of realistic network workload for emerging networking scenarios," Comp. Netw. vol. 56, no. 15, 2012, pp.3531-3547. https://doi.org/10.1016/j.comnet.2012.02.019.

4GL Code Generation: A Systematic Review

Abdullah A H Alzahrani[1]
Department of Computer Sciences, Computing College -Alqunfuda
Umm Al Qura University, Makkah, Saudi Arabia

***Abstract*—Code generation is longstanding goal in software engineering. It allows more productivity of computer programming as it aims to provide automation of transformation of models into actual source code. This process has been covered adequately in many programming languages. However, this topic has not been covered sufficiently with regards to Fourth Generation Languages (4GL) which have a high specialized nature. The goal of this paper is to represent a systematic literature review of 4GL Code generation. The paper focuses on reviewing systemically the studies published in the past 20 years on the topic. This is to investigate the trends in the topic and the approaches introduced in order to identify potential new research lines.**

Keywords—Software engineering; code transformation; 4GL; code generation; Model Driven Development (MDD); Extraction Transform Load (ETL); Model Driven Engineering (MDE); Rapid Application Development (RAD)

I. Introduction

Fourth-generation languages (4GL) are a class of high-level programming languages. The idea of 4GL is to make programming as simple as possible by the mean of programming in natural human language [2], [3]. 4GLs aim to make programming easier, more efficient and more effective for users with less programming skills [4].

However, most of 4GLs are proprietary or designed and developed for a very specific scope and purpose. This has led to a hinder in progressing research and development in the area of 4GL. These difficulties are derived from highly-specialized nature of 4GL. In addition, the limited target users pose accessibility issues. Furthermore, insufficient tool support causes a problem as 4GL needs in-house development of tools support. Consequently, this introduces high costs for 4GL evolvements in all areas of research [5]. The followings summarize the reasons behind the recessions in progressing research and development: proprietary developments, complexity, modularity, and integrations with other interfaces (web browsers) issues [5], [6].

A number of 4GL exist and are used in many areas from research and industry, for example: Query languages SQL, Oracle, Report Generator, Magic, Informix, Advanced Business Application Programming (ABAP), MathWorks, MATLAB, SPSS, etc. [5]. Some of these languages are popular as they are open to use and some are not as they are proprietary.

Code generation is longstanding goal in software engineering [1]. However, it is not a straightforward process nor an easy to achieve optimal solution for it [1]. Code generation is the process of transforming a model of high-level representation to a source code that can be read and understood by a computer. Usually, this is done via a use of Computer-Assisted Software Engineering (CASE) tools [5], [7], [8]. A CASE tool can generate initial software or database code directly from system models. Examples of 4GL CASE tools are: Oracle2Java, Evo, Jheadstart, Pitss, Ormit [9]. Code generation is an actual practice of forward engineering.

Several problems are often attached to the CASE code generations. These issues are related mainly the complexity of the models or the capability of the target language. In addition, in the case that models gained by reverse engineering legacy systems, code generations CASE tools are often regarded as less useful especially if they are independent from the reverse engineering CASE tools.

In general, the existing approaches and tools offered for 4GL code generation suffer from several limitations. First, the majority focuses on auto-transformation of Oracle forms, whereas many 4GL languages are not considered. Second, majority of these approaches are semi-automated approaches which often need experts to be involved. Finally, immaturity is a nature for these approaches as only proof-of-concept models and tools are presented. However, mature approaches might be developed but for proprietary purposed and not shared for researchers.

The systematic review is a process of addressing a research question then finding and evaluating all available research done with relation to it [10]. It helps highlighting the major work conducted in the area of the research question. From such a review, research gaps can be found in a way that assure a satisfactory coverage of area of research.

A number of researchers [11]–[13] have considered reviewing the topic of transformation between 4G languages from different points of view. In addition, they have concluded that it is non-trivial process and needs a considerable manual effort and knowledge from all of the people involved. However, sufficiency when authors are reviewing related work of others is still an issue. In addition, up to the date of writing this paper, there is no systematic review on the topic of 4GL code generation. This leads to the importance of investigating objectively the current state of arts in the topic of 4GL code generation. This paper aims to provide an objective and systematic review of the topic of 4GL code generation. Introduction of such a review allows evolving the research in this area and highlighting the current research gaps that are not resolved.

This paper has been structured as follows. Section 1 introduced the topic of 4GL and code generation. In addition, it highlights the importance of systematic review on the considered topic. Section 2 illustrates the methodology used in this paper and formulate the research question. Section 3 discusses the trends in the area of the 4GL code generations. Section 4 categories the main works conducted in the 4GL code generations and discussed the main findings. Finally, Section 5 concludes the review with the identification of trends and new research areas.

II. Methodology

This paper employs a systematic review methodology demonstrated in [10], [14]. The methodology allows a structural and objective review of a topic under consideration. In addition, it allows providing a broad view on works which are primary and related to the topic under consideration.

A. *Research Question*

What are the initiatives undertaken in relation to 4GL code generations in the last 20 years? This research question can be answered by identifying the approaches, models, and CASE tools introduced, Thereafter, highlighting the major difficulties and issues faced in order to achieve the goal of code generations. Therefore, keywords leading the search have been listed. These keywords are: code generation, fourth generation languages, and 4GL. It is important to mention here that 4GL is often referred to fourth generation languages [5].

B. *Sources Selection*

Having the aforementioned keywords, a search string has been made from a combination of these keywords and been used in the search by the search engines of the selected digital libraries. Table I shows the formulated search string with an OR logical operator which is one of useful operators offered by the search engines in the digital libraries.

Four well known digital libraries were selected in order to perform the search for related studies in. These libraries are: Springer Link, IEEE Digital Library, ACM Digital Library, and ScienceDirect. These libraries offer variety ways of searching for journal articles, conference papers, books, and other publication types. In addition, these libraries offer the use of logical operators in the searching.

C. *Inclusion and Exclusion Criteria*

Inclusion criteria, for studies to be considered relevant, are the outcome of analyzing the title, research keywords, abstract, and the conclusion of a paper. In addition, as this research aims to investigate the work done towards 4GL code generation in the past 20 years, an exclusion criterion of year of publication has been applied to retrieve only the work published between the years 2000 and 2020. Another exclusion criterion was a non-English publication items which have been found in the retrieved list of items from the digital libraries.

As can be seen in Table II, in the first iteration of applying the search string in all digital libraries, the search resulted in a total of 1925 items. Repetition of items in some between libraries was noticed, so, it was inevitable to eliminate repetitions. After removing repetitions, the total was 1770 items. The total become 593 when applying exclusion criterion of publication before the year of 2000. Finally, 187 of the publication items were relevant to this research topic by inclusion criterion.

TABLE I. SEARCH STRING

Search string	"4G languages" OR "4G languages code generation" OR “4GL code generation” OR “4GL” OR “fourth-generation-languages” OR “fourth-generation-languages code generation” OR “fourth generation languages code generation” OR “fourth generation languages”

TABLE II. SOURCES AND STUDIES FOUND

Sources	Studies					
	Found	Not repeated	Since 2000	Relevant	Primary	%
Springer Link	793	694	216	64	6	21.43%
IEEE Digital Library	35	35	15	12	7	25.00%
ACM Digital Library	302	274	99	41	6	21.43%
ScienceDirect	795	767	263	70	9	32.14%
Total	1925	1770	593	187	28	100%

Primary studies were, as shown Table II, 28 studies. Complete list of primary studies is shown in (Appendix A). The primary studies were nominated from the relevant studies after in-depth reading and analysis of the entire list of the relevant studies. The primary studies are only the studies which mainly consider the 4GL code generation by introduction new approaches, model development, evaluation of current approaches, and branches of code generation with relation 4GL such as code transformation. It is important to mention that studies that are related to the 4GL code generation have been considered to be relevant studies. Examples of which are code quality, effort estimation, optimization, refactoring, and code maintenance.

III. Publication in 4GL Code Generation in the Past 20 Years

This section shows an analysis of primary and relevant studies in general and the main aspects and remarks identified during the reviewing process. Obviously, from the number of relevant studies, it can be concluded that the topic of 4GL code generation has not been considered sufficiently in the past 20 years. Fig. 1 illustrates publications trends of primaries and relevant studies of 4GL code generation over the past 20 years in the four digital libraries namely Springer, IEEE, ACM, and ScienceDirect.

From Fig. 1, it can be noticed that the years of 2003, 2005, 2011 were the years were most studies were published with a number of around 14 publications in all digital libraries investigated. Then, the years of 2008 and 2019 come in almost close number of the aforementioned. In addition, it can be obvious that the average of publications related to the topic of 4GL code generation is 9 studies each year over the past 20 years. Most of these primaries and relevant studies have been published in Springer and ScienceDirect. This highlights a lack of studies in this topic.

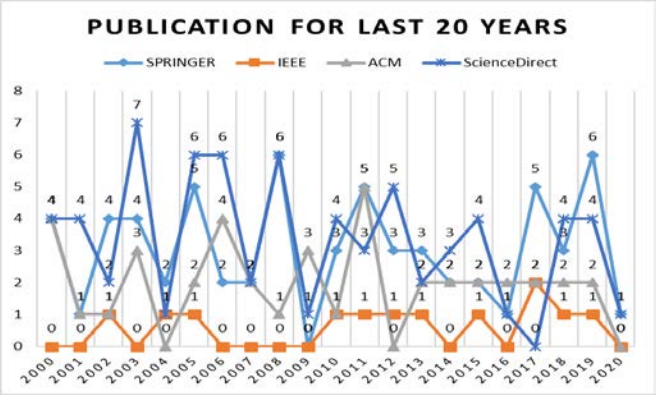


Fig. 1. Run graph of publications (Relevant Studies Published in the Past 20 Years based on Number of Publications).

Table III shows the categories of the publications based on the types of work. The main types are conference papers, journal articles, and books and technical reports. The majority of primaries and relevant studies are conference papers and journal articles with the total of 153 studies. In addition, as illustrated in Fig. 2, 59% of journal articles which are primaries and relevant studies published in ScienceDirect library. Furthermore, 74% of books and technical reports are published in same library. However, 45% of conference papers which are primaries and relevant studies published in Springer library with 0% published in ScienceDirect library of this category of the publications.

TABLE III. TYPES OF PUBLICATIONS OF PRIMARIES AND RELEVANT STUDIES

Type / Library	Conference Paper	Journal Article	Book and technical reports	Total
SPRINGER	35	20	9	64
IEEE	12	0	0	12
ACM	30	11	0	41
ScienceDirect	0	44	26	70
Total	77	75	35	187

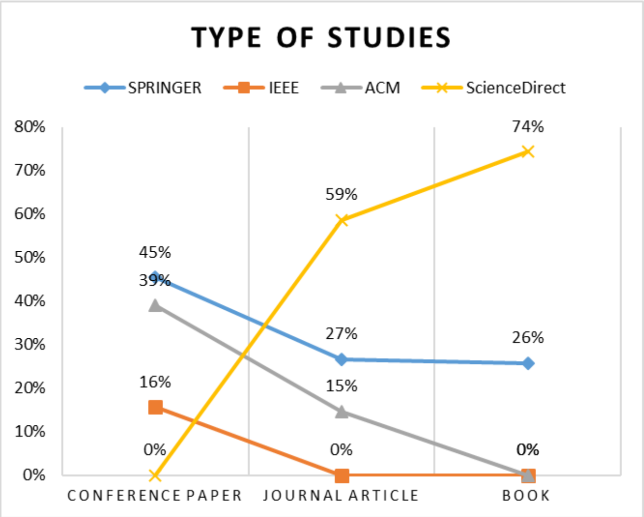


Fig. 2. Studies Types (Only Relevant Studies Published in the Past 20 Years.

IV. PRIMARY STUDIES AND DISCUSSIONS

In this section the main findings of this systematic review regarding 4GL code generation are presented and discussed. The findings have been categorised into three main sections: 1) Transformation between 4GL languages; 2) End user computing; 3) other related studies. In this review, the related studies are the studies which indirectly consider the code generation in 4GL, such as studies considering effort estimation in producing, manually re-engineering, refactoring, and maintaining 4GL software systems.

A. Transformation between 4GL Languages

Kicsi et al. [15] have introduced a semi-automatic approach which extracts features from a 4GL language namely Magic language. They have tested their approach on 2000 programs written in Magic. However, the completed stage up-to-date is extraction stage and the project is still ongoing. In addition, experts are needed to provide design decisions.

In addition, Kicsi et al. [16] have introduced an approach that extracted the structural and conceptual feature of legacy systems built in Magic language. The approach aims to provide two level of views on the legacy systems. The first level views are for expert which show the conceptual features. The second level views are for developer which shows structural feature. Although, this work is a promising in the reengineering of 4GL systems as it helps is the stage of design discovery, the work currently provides information for different level of stakeholders to make decisions in the re-designing of the legacy systems when the adopting Software product line (SPL) architecture.

Mendivelso et al. [9] have introduced an approach that relies on Model-Driven Reverse Engineering (MDRE) in order to reverse engineer programs in 4GL languages such as Oracle Forms, Visual Basic and Delphi. The resulting outputs are two different levels of views on Sirius graphical editor. The first level is for the end user who are the developers, architects, and testers. The second level is for MDRE experts. The approach seems promising, however, experts are needed to validate and verify the output model of a given source code. In addition, no forward engineering (complete code generation) is completed by the approach.

Newcomb et al. [17] have reported on a project called Pilot Project which aims to transform 4GL software systems into more standardized and modernized platforms namely Java and JavaScript. The work has a feature of considering conversion of non-functional requirements as well as security requirements. Moreover, tests were done on small scale programs to prove the concept. However, an important point is that 4GL software are often large scale ones. So, in order to generalize the findings this could be a point of weakness. In addition, further manual tuning was needed as performance issues occurred.

Sneed et al. [18] have introduced an approach that consider re-implementing legacy systems built by 4GL languages into object-oriented 3GL languages. This approach aim is to avoid the risks of automated conversion of such systems and taking into account the preserving functionalities. The approach was tested on two 4GL languages programs, namely, VisualAge/PL/I-DB2 and COBOL-IMS applications. However, they have reported a number of side effects including comprising design and possibility of re-developing the whole architecture. This stands against the idea of preserving the original architecture.

Garcés et al. [19] have introduced a new semi-automated approach that transformed 4GL program to modernized platforms namely from Oracle Forms to Java programs. The approach has been tested on 5 medium scale 4GL programs. Although, the approach has a tool supports and seems promising, it has some backwards which are worth to mention here. Firstly, the approach is semi-automated and needs a human intervene. This might, as author reported, time consuming and error-prone. In addition, migration is a manual process. Finally, although, the results showed a reduction in time with comparison with other transformation processes, time overhead and code defects are still a considerable issue.

Salvatierra et al. [20] have introduced an indirect and semi-automatic approach for migration of legacy systems in COBOL into Service-Oriented Architecture (SOA). The approach is called Assisted Migration and has been tested on a legacy system of an Argentinean government agency. The aim of the approach is to enhance the quality of direct migrated version of the legacy system. However, the approach introduces a need for human experiences to perform as intended. In addition, accuracy is still an issue. Furthermore, legacy systems are not transformed or replaced which means adding more layers to use these systems.

Sánchez Ramón et al. [21] have introduced a new approach based on Model Driven Engineering (MDE). The approach aims to automated re-engineering process of the interfaces of programs built in 4GL namely Oracle Forms or Borland Delphi. Currently, the approach allows detecting the main elements in GUI and generate a tree to represents the arrangement of element on the GUI window. The resulting outputs is a model in Concrete User Interface (CUI) which can be used for further forward engineering.

Nagy et al. [22] in 2011 have investigated the lack of work on 4GL same language version transformation. In addition, the authors have offer a new approach to automatically transform code from Magic older to Magic version 5. However, the focus of the work was only on version 5 of Magic. In addition, performance issues have been raised. In addition later, Nagy [23] in 2013 have introduced an automated approach to recovering architecture of data-intensive applications developed in Magic 4GL. However, the approach only support static reverse engineering from SQL no forward engineering or round-trip engineering.

Martin et al. [20] introduced a new approach to transform source code between two different 4GL platforms. The approach aims to overcome the incompatibilities between 4GL. Therefore, a tool, called OctaveToR, was developed to automate the transformation from a source code written in Octave to a target code in R language. The approach employed TXL transformation language and was tested in a medium size source code. Although, the approach provides almost an instant code transformation, a number of issues are reported. These issues are performance, readability, and information loss. In addition, a use of TXL cannot be an ideal solution as it does not offer a feature abstraction representation.

Yafi et al. [24] introduced a new method that allow overcoming a problem that occurs when parsing Uniface 4GL languages source code embedded in XML format. Although the tool provides an automated way for reverse-engineer a 4GL code, the work was only to improve the readability and the work is in progress.

Nandivada et al. [25] introduced a framework that translate 4GL program in ARAP to java equivalent. This is for the purpose of debugging fault in 4GL programs. However, the work still incomplete and suffer from incorrectness of transformation with some statements in selected 4GL syntax as well as some overhead issues.

Bimonte et al. [26] introduced a new Model Driven Development (MDD) method which combines the use of ETL (Extraction, transform, load) and their Business Process Modeling Notation (BPMN) approaches to transform source from ETL to Oracle MetaBase (OMB) scripting language code [27]. However, efficiency in resources and time is an issue.

Reus et al. [28] have introduced an approach which aids in reverse engineer legacy systems to model-driven architecture (MDA). In specific, reverse engineer a 4GL program to language-independent models in UML, namely, Class, State-chart, Collaboration diagrams. Code generation then is offered to transform the models to Java classes. The approach has been tested to an Oracle's PL/SQL program of an insurance company in Netherlands. Although, the authors have introduced a promising approach which aims to automate completely the re-engineering trip from a 4GL source code to another platform, a number of pitfalls are there such as scalability issues. Another open challenge is the representation of business logic. In addition, the code generation is an uncompleted goal as the approach only generate code stubs (no functionally).

Cleve et al. [29] introduced an approach for transformation of date structure from 4GL language in legacy system to a modular structure for other platforms. However, the main concern was the data migration and data structure re-factoring. In addition, it took 10 days for re-engineering which is an obvious overhead. This can be linked to the previous work by Canfora et al. [30] where they in the same manner introduced their approach and noted the re-engineering risks increase beside the costs and performance issues.

Andrade et al. [12] have introduced a tool called Forms2Net which aims to transform the Oracle Forma and PL/SQL code to .NET C# program bearing in mind semantics and similarity and differences. Authors have investigated semantics and functionality in such transformations and offer the tool based on this. Forms2Net is a promising tools to facilitate an automatic code transformation, however, a number of shortcomings are impotent to mention here. First, complex transformations decisions are not made and left to the developer to re-engineer which introduced the human intervention. Second, only one output architecture is allowed which is the Model View Controller (MVC) architecture. Third, migration process needs to be simplified as currently

Forms2Net is attached with a number of guides for explaining it. Finally, runtime calls in the Oracle Forms are not sufficiently represented in C# outputs program.

B. End user Computing

Waszkowski [31] have introduced Aurea BPM low-code platform that allows users to draw in BPMN diagrams which will be transformed into working web pages with XML and supplementary files. The main goal is to automate the generation of application for business processes. However, authors stated that low-code platform is hard in manufacturing and it raises the risk of verifications.

Related to this topic, a number of authors [4], [32] have published a book in which they describe a number of examples of 4GL languages for End-user programing. The books show an exploring view on the available tools for such an approach. Others [33] have looked at it from different viewpoints such as the risks of privacy and errors which might be posed. Furthermore, bridging the knowledge gap between engineers and business users [34].

C. Other Related Topics

A number of related topics have been a focus for 4GL community, for example, Effort Estimation, quality assurance, testing, and distributed programming. However, it is important to mention here that Effort Estimation in producing, manually re-engineering, refactoring, and maintaining 4GL software systems has gained a considerable amount of attention form 4GL community.

Although, many researchers [35]–[45] have considered Effort Estimation, other have considered useful topics as well. For example, Shasharina et al. [46] have considered a model that offer ability of automating the linking the Grid Technology and Web services for 4GL legacy systems in Interactive Data Language (IDL). Furthermore, many researchers [47]–[49] have introduced their models and methods on measuring quality and have offers a number of matrices for this. However, quality assurance for 4GL suffer from a lack of work on it. Other have considered different related topics to 4GL. For instance, Zaytsev [50] has reported onto a new tool which generates test codes for a 4GL programs in specific C# programs. It has been tested on a large scale code of a company where the author is working. However, this is an ongoing project of generating code of testing for 4GL languages compiler with focus to C#. Furthermore, Albizuri-Romero [51] discussed different factors influenced organizations when choosing CASE tools.

V. Conclusions

In conclusion of this systematic review, four well-known digital libraries have been used to search for research studies that are related to the topic of 4GL code generation over the past 20 years. These libraries are Springer, IEEE, ACM, and ScienceDirect. Total of 593 studies were found. After applying the criteria of inclusion and exclusion employed in this paper, a total of 187 studies were found relevant studies. Out of these relevant studies only 28 were found primaries studies.

The following summarizes the main findings of the primary studies published over the past 20 years:

1) In general, there is a lack of studies in 4GL code generation.

2) The topic of 4GL languages code generation is often focused on transforming 4GL source codes to different 4GL or 3GL languages.

3) Lack of studies is variety of 4GL languages as the majority focuses on auto-transformation of Oracle forms.

4) The majority of the studies introduce semi-automated approaches which often need experts to be involved.

5) The majority of studies are Immature studies which might be due to the specialized nature of 4GL and that most of 4GL are proprietary.

Reaching this point of this paper, the previously mentioned research question of this paper can be answered. 4GL code generation is a topic that has not been considered sufficiently over the past 20 years. The offered studies and approaches are inadequate, and more work is needed in this topic. Furthermore, the previous section shows a detailed answer of the research question.

For future work, this paper can be basis for researchers interested in 4GL code generation and code transformation as the paper aimed to cover the work done on the topic for the past 20 years. As for authors for this paper, the future work direction is to fill the research gap of transforming Uniface 4GL to C#.net as the limitation of coverage for this direction is clear. In addition, a data set that allow constructing the transformation model has been shared from one of institute interested in the direction.

Acknowledgment

I would like to express my gratitude to my University (Umm Al Qura University) which gave me the opportunity to do this research. Secondly I would also like to thank family for their unlimited support. In addition, I would like to thank Saudi Digital Library (SDL) for giving me the membership to allow accessing digital libraries used in this paper.

References

[1] J. Krogstie, A. L. Opdahl, and S. Brinkkemper, Eds., Conceptual Modelling in Information Systems Engineering. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[2] W. H. Inmon, D. Strauss, and G. Neushloss, 'Chapter 1 - A brief history of data warehousing and first-generation data warehouses', in DW 2.0, W. H. Inmon, D. Strauss, and G. Neushloss, Eds. Burlington: Morgan Kaufmann, 2008, pp. 1–22.

[3] C. Combe, 'Chapter 2 - E-business technology', in Introduction to e-Business, C. Combe, Ed. Oxford: Butterworth-Heinemann, 2006, pp. 21–52.

[4] C. Shipley and S. Jodis, 'Programming Languages Classification', in Encyclopedia of Information Systems, H. Bidgoli, Ed. New York: Elsevier, 2003, pp. 545–552.

[5] B. Selic, 'Personal reflections on automation, programming culture, and model-based software engineering', Autom. Softw. Eng., vol. 15, no. 3–4, pp. 379–391, 2008.

[6] J. Jong, 'History: How Did We Get Here?', in Vertically Integrated Architectures: Versioned Data Models, Implicit Services, and Persistence-Aware Programming, J. Jong, Ed. Berkeley, CA: Apress, 2019, pp. 15–28.

[7] J. Whitten and L. Bentley, Systems Analysis and Design Methods, 7th edition. Boston: McGraw-Hill/Irwin, 2005.

[8] F. J. Budinsky, M. A. Finnie, J. M. Vlissides, and P. S. Yu, ‘Automatic code generation from design patterns’, IBM Syst. J., vol. 35, no. 2, pp. 151–171, 1996.

[9] L. F. Mendivelso, K. Garcés, and R. Casallas, ‘Metric-centered and technology-independent architectural views for software comprehension’, J. Softw. Eng. Res. Dev., vol. 6, no. 1, p. 16, Dec. 2018, doi: 10.1186/s40411-018-0060-6.

[10] D. Budgen and P. Brereton, ‘Performing systematic literature reviews in software engineering’, in Proceedings of the 28th international conference on Software engineering, 2006, pp. 1051–1052.

[11] S. M. F. Ali and R. Wrembel, ‘From conceptual design to performance optimization of ETL workflows: current state of research and open problems’, VLDB J., vol. 26, pp. 777–801, 2017.

[12] L. Andrade, J. Gouveia, M. Antunes, M. El-Ramly, and G. Koutsoukos, ‘Forms2Net–migrating oracle forms to microsoft. NET’, in International Summer School on Generative and Transformational Techniques in Software Engineering, 2005, pp. 261–277.

[13] A. Kicsi, V. Csuvik, L. Vidács, Á. Beszédes, and T. Gyimóthy, ‘Feature level complexity and coupling analysis in 4GL systems’, in International Conference on Computational Science and Its Applications, 2018, pp. 438–453.

[14] B. Kitchenham, ‘Procedure for undertaking systematic reviews’, Comput. Sci. Depart-Ment Keele Univ. TRISE-0401 Natl. ICT Aust. Ltd 0400011T 1 Jt. Tech. Rep., 2004.

[15] A. Kicsi et al., ‘Feature analysis using information retrieval, community detection and structural analysis methods in product line adoption’, J. Syst. Softw., vol. 155, pp. 70–90, 2019, doi: https://doi.org/10.1016/j.jss.2019.05.001.

[16] A. Kicsi, L. Vidács, V. Csuvik, F. Horváth, A. Beszédes, and F. Kocsis, ‘Supporting product line adoption by combining syntactic and textual feature extraction’, in International Conference on Software Reuse, 2018, pp. 148–163.

[17] P. H. Newcomb, D. Henke, J. LoVerde, W. Ulrich, L. Nguyen, and R. Couch, ‘Chapter 6 - PowerBuilder/4GL Generator Modernization Pilot**© 2010. The Software Revolution, Inc. All rights reserved.’, in Information Systems Transformation, W. M. Ulrich and P. H. Newcomb, Eds. Boston: Morgan Kaufmann, 2010, pp. 133–170.

[18] H. Sneed and C. Verhoef, ‘Re-implementing a legacy system’, J. Syst. Softw., vol. 155, pp. 162–184, Sep. 2019, doi: 10.1016/j.jss.2019.05.012.

[19] K. Garcés et al., ‘White-box modernization of legacy applications: The oracle forms case study’, Comput. Stand. Interfaces, vol. 57, pp. 110–122, 2018, doi: https://doi.org/10.1016/j.csi.2017.10.004.

[20] G. Salvatierra, C. Mateos, M. Crasso, and A. Zunino, ‘Towards a computer assisted approach for migrating legacy systems to SOA’, in International Conference on Computational Science and Its Applications, 2012, pp. 484–497.

[21] Ó. Sánchez Ramón, J. Sánchez Cuadrado, and J. García Molina, ‘Model-driven reverse engineering of legacy graphical user interfaces’, in Proceedings of the IEEE/ACM international conference on Automated software engineering, 2010, pp. 147–150.

[22] C. Nagy, L. Vidács, R. Ferenc, T. Gyimóthy, F. Kocsis, and I. Kovács, ‘Solutions for Reverse Engineering 4GL Applications, Recovering the Design of a Logistical Wholesale System’, in 2011 15th European Conference on Software Maintenance and Reengineering, Mar. 2011, pp. 343–346, doi: 10.1109/CSMR.2011.66.

[23] C. Nagy, ‘Static Analysis of Data-Intensive Applications’, in 2013 17th European Conference on Software Maintenance and Reengineering, Mar. 2013, pp. 435–438, doi: 10.1109/CSMR.2013.66.

[24] M. Z. Yafi and A. Fatima, ‘Syntax Recovery for Uniface as a Domain Specific Language’, in 2018 UKSim-AMSS 20th International Conference on Computer Modelling and Simulation (UKSim), Mar. 2018, pp. 61–66, doi: 10.1109/UKSim.2018.00023.

[25] V. K. Nandivada, M. G. Nanda, P. Dhoolia, D. Saha, A. Nandy, and A. Ghosh, ‘A framework for analyzing programs written in proprietary languages’, in Proceedings of the ACM international conference companion on Object oriented programming systems languages and applications companion, 2011, pp. 289–300.

[26] S. Bimonte, É. Edoh-Alove, H. Nazih, M.-A. Kang, and S. Rizzi, ‘ProtOLAP: rapid OLAP prototyping with on-demand data supply’, in Proceedings of the sixteenth international workshop on Data warehousing and OLAP, 2013, pp. 61–66.

[27] Z. El Akkaoui, E. Zimànyi, J.-N. Mazón, and J. Trujillo, ‘A model-driven framework for ETL process development’, in Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP, 2011, pp. 45–52.

[28] T. Reus, H. Geers, and A. Van Deursen, ‘Harvesting software systems for MDA-based reengineering’, in European Conference on Model Driven Architecture-Foundations and Applications, 2006, pp. 213–225.

[29] A. Cleve, J. Henrard, and J.-L. Hainaut, ‘Co-transformations in Information System Reengineering’, Electron. Notes Theor. Comput. Sci., vol. 137, no. 3, pp. 5–15, Sep. 2005, doi: 10.1016/j.entcs.2005.07.001.

[30] G. Canfora, A. Cimitile, A. De Lucia, and G. A. Di Lucca, ‘Decomposing legacy programs: a first step towards migrating to client–server platforms’, J. Syst. Softw., vol. 54, no. 2, pp. 99–110, Oct. 2000, doi: 10.1016/S0164-1212(00)00030-3.

[31] R. Waszkowski, ‘Low-code platform for automating business processes in manufacturing’, IFAC-Pap., vol. 52, no. 10, pp. 376–381, 2019, doi: https://doi.org/10.1016/j.ifacol.2019.10.060.

[32] J. Stigliano and M. Bruni, ‘End-User Computing Tools’, in Encyclopedia of Information Systems, H. Bidgoli, Ed. New York: Elsevier, 2003, pp. 127–139.

[33] R. R. Panko and D. N. Port, ‘End User Computing: The Dark Matter (and Dark Energy) of Corporate IT’, in 2012 45th Hawaii International Conference on System Sciences, Jan. 2012, pp. 4603–4612, doi: 10.1109/HICSS.2012.244.

[34] G. Baster, P. Konana, and J. E. Scott, ‘Business components: a case study of bankers trust Australia limited’, Commun. ACM, vol. 44, no. 5, pp. 92–98, 2001.

[35] P. A. Whigham, C. A. Owen, and S. G. Macdonell, ‘A baseline model for software effort estimation’, ACM Trans. Softw. Eng. Methodol. TOSEM, vol. 24, no. 3, pp. 1–11, 2015.

[36] L. Song, L. L. Minku, and X. Yao, ‘A novel automated approach for software effort estimation based on data augmentation’, in Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2018, pp. 468–479.

[37] S. Amasaki and C. Lokan, ‘A replication study on the effects of weighted moving windows for software effort estimation’, in Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering, 2016, pp. 1–9.

[38] L. L. Minku and X. Yao, ‘An analysis of multi-objective evolutionary algorithms for training ensemble models based on different performance measures in software effort estimation’, in Proceedings of the 9th international conference on predictive models in software engineering, 2013, pp. 1–10.

[39] P. Pospieszny, B. Czarnacka-Chrobot, and A. Kobylinski, ‘An effective approach for software project effort and duration estimation with machine learning algorithms’, J. Syst. Softw., vol. 137, pp. 184–196, 2018.

[40] T. Tran, V. Nguyen, T. Truong, C. Tran, and P. Le, ‘An Evaluation of Parameter Pruning Approaches for Software Estimation’, in Proceedings of the Fifteenth International Conference on Predictive Models and Data Analytics in Software Engineering, 2019, pp. 26–35.

[41] M. Tanrıverdi and Ö. Ö. Tanrıöver, ‘An experimental comparison of software effort estimation methods of ORM based 4GL software applications’, in 2017 International Conference on Computer Science and Engineering (UBMK), 2017, pp. 239–243.

[42] P. Pospieszny, B. Czarnacka-Chrobot, and A. Kobyliński, ‘Application of function points and data mining techniques for software estimation-a combined approach’, in Software Measurement, Springer, 2015, pp. 96–113.

[43] P. Rijwani and S. Jain, ‘Enhanced Software Effort Estimation Using Multi Layered Feed Forward Artificial Neural Network Technique’, Procedia Comput. Sci., vol. 89, pp. 307–312, 2016, doi: https://doi.org/10.1016/j.procs.2016.06.073.

[44] F. Ferrucci, C. Gravino, and F. Sarro, ‘Exploiting prior-phase effort data to estimate the effort for the subsequent phases: a further assessment’, in Proceedings of the 10th International Conference on Predictive Models in Software Engineering, 2014, pp. 42–51.

[45] C. Nagy, L. Vidács, R. Ferenc, T. Gyimóthy, F. Kocsis, and I. Kovács, ‘MAGISTER: Quality assurance of Magic applications for software developers and end users’, in 2010 IEEE International Conference on Software Maintenance, Sep. 2010, pp. 1–6, doi: 10.1109/ICSM.2010.5609550.

[46] S. G. Shasharina, O. Volberg, P. Stoltz, and S. Veitzer, ‘GRIDL: high-performance and distributed interactive data language’, in HPDC-14. Proceedings. 14th IEEE International Symposium on High Performance Distributed Computing, 2005., Jul. 2005, pp. 291–292, doi: 10.1109/HPDC.2005.1520980.

[47] Z. Tóth, L. Vidács, and R. Ferenc, ‘Comparison of static analysis tools for quality measurement of rpg programs’, in International Conference on Computational Science and Its Applications, 2015, pp. 177–192.

[48] M.-A. Côté, W. Suryn, C. Y. Laporte, and R. A. Martin, ‘The evolution path for industrial software quality evaluation methods applying ISO/IEC 9126: 2001 quality model: example of MITRE’s SQAE method’, Softw. Qual. J., vol. 13, no. 1, pp. 17–30, 2005.

[49] G. C. Green, A. R. Hevner, and R. W. Collins, ‘The impacts of quality and productivity perceptions on the use of software process improvement innovations’, Inf. Softw. Technol., vol. 47, no. 8, pp. 543–553, 2005.

[50] V. Zaytsev, ‘An industrial case study in compiler testing (tool demo)’, in Proceedings of the 11th ACM SIGPLAN International Conference on Software Language Engineering, 2018, pp. 97–102.

[51] M. B. Albizuri-Romero, ‘A retrospective view of CASE tools adoption’, ACM SIGSOFT Softw. Eng. Notes, vol. 25, no. 2, pp. 46–50, 2000..

Appendix A

Primary Studies

Item	Bibliography	Source
PS1	H. Sneed and C. Verhoef, ‘Re-implementing a legacy system’, Journal of Systems and Software, vol. 155, pp. 162–184, Sep. 2019, doi: 10.1016/j.jss.2019.05.012.	ScienceDirect
PS2	R. Waszkowski, ‘Low-code platform for automating business processes in manufacturing’, IFAC-PapersOnLine, vol. 52, no. 10, pp. 376–381, 2019, doi: https://doi.org/10.1016/j.ifacol.2019.10.060.	ScienceDirect
PS3	A. Kicsi et al., ‘Feature analysis using information retrieval, community detection and structural analysis methods in product line adoption’, Journal of Systems and Software, vol. 155, pp. 70–90, 2019, doi: https://doi.org/10.1016/j.jss.2019.05.001.	ScienceDirect
PS4	L. F. Mendivelso, K. Garcés, and R. Casallas, ‘Metric-centered and technology-independent architectural views for software comprehension’, J Softw Eng Res Dev, vol. 6, no. 1, p. 16, Dec. 2018, doi: 10.1186/s40411-018-0060-6.	SPRINGER
PS5	M. Z. Yafi and A. Fatima, ‘Syntax Recovery for Uniface as a Domain Specific Language’, in 2018 UKSim-AMSS 20th International Conference on Computer Modelling and Simulation (UKSim), Mar. 2018, pp. 61–66, doi: 10.1109/UKSim.2018.00023.	IEEE
PS6	V. Zaytsev, ‘An industrial case study in compiler testing (tool demo)’, in Proceedings of the 11th ACM SIGPLAN International Conference on Software Language Engineering, 2018, pp. 97–102.	ACM
PS7	A. Kicsi, L. Vidács, V. Csuvik, F. Horváth, A. Beszédes, and F. Kocsis, ‘Supporting product line adoption by combining syntactic and textual feature extraction’, in International Conference on Software Reuse, 2018, pp. 148–163.	SPRINGER
PS8	K. Garcés et al., ‘White-box modernization of legacy applications: The oracle forms case study’, Computer Standards & Interfaces, vol. 57, pp. 110–122, 2018, doi: https://doi.org/10.1016/j.csi.2017.10.004.	ScienceDirect
PS9	R. R. Panko and D. N. Port, ‘End User Computing: The Dark Matter (and Dark Energy) of Corporate IT’, in 2012 45th Hawaii International Conference on System Sciences, Jan. 2012, pp. 4603–4612, doi: 10.1109/HICSS.2012.244.	IEEE
PS10	G. Salvatierra, C. Mateos, M. Crasso, and A. Zunino, ‘Towards a computer assisted approach for migrating legacy systems to SOA’, in International Conference on Computational Science and Its Applications, 2012, pp. 484–497.	SPRINGER
PS11	C. Nagy, L. Vidács, R. Ferenc, T. Gyimóthy, F. Kocsis, and I. Kovács, ‘Solutions for Reverse Engineering 4GL Applications, Recovering the Design of a Logistical Wholesale System’, in 2011 15th European Conference on Software Maintenance and Reengineering, Mar. 2011, pp. 343–346, doi: 10.1109/CSMR.2011.66.	IEEE
PS12	V. K. Nandivada, M. G. Nanda, P. Dhoolia, D. Saha, A. Nandy, and A. Ghosh, ‘A framework for analyzing programs written in proprietary languages’, in Proceedings of the ACM international conference companion on Object oriented programming systems languages and applications companion, 2011, pp. 289–300.	ACM
PS13	Z. El Akkaoui, E. Zimànyi, J.-N. Mazón, and J. Trujillo, ‘A model-driven framework for ETL process development’, in Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP, 2011, pp. 45–52.	ACM
PS14	Ó. Sánchez Ramón, J. Sánchez Cuadrado, and J. García Molina, ‘Model-driven reverse engineering of legacy graphical user interfaces’, in Proceedings of the IEEE/ACM international conference on Automated software engineering, 2010, pp. 147–150.	SPRINGER
PS15	P. H. Newcomb, D. Henke, J. LoVerde, W. Ulrich, L. Nguyen, and R. Couch, ‘Chapter 6 - PowerBuilder/4GL Generator Modernization Pilot**© 2010. The Software Revolution, Inc. All rights reserved.’, in Information Systems Transformation, W. M. Ulrich and P. H. Newcomb, Eds. Boston: Morgan Kaufmann, 2010, pp. 133–170.	ScienceDirect
PS16	T. Reus, H. Geers, and A. Van Deursen, ‘Harvesting software systems for MDA-based reengineering’, in European Conference on Model Driven Architecture-Foundations and Applications, 2006, pp. 213–225.	SPRINGER
PS17	A. Cleve, J. Henrard, and J.-L. Hainaut, ‘Co-transformations in Information System Reengineering’, Electronic Notes in Theoretical Computer Science, vol. 137, no. 3, pp. 5–15, Sep. 2005, doi: 10.1016/j.entcs.2005.07.001.	ScienceDirect
PS18	S. G. Shasharina, O. Volberg, P. Stoltz, and S. Veitzer, ‘GRIDL: high-performance and distributed interactive data language’, in HPDC-14. Proceedings. 14th IEEE International Symposium on High Performance Distributed Computing, 2005., Jul. 2005, pp. 291–292, doi: 10.1109/HPDC.2005.1520980.	IEEE

PS19	C. Nagy, ‘Static Analysis of Data-Intensive Applications’, in 2013 17th European Conference on Software Maintenance and Reengineering, Mar. 2013, pp. 435–438, doi: 10.1109/CSMR.2013.66.	IEEE
PS20	L. Andrade, J. Gouveia, M. Antunes, M. El-Ramly, and G. Koutsoukos, ‘Forms2Net–migrating oracle forms to microsoft. NET’, in International Summer School on Generative and Transformational Techniques in Software Engineering, 2005, pp. 261–277.	SPRINGER
PS21	J. Martin and J. Gutenberg, ‘Automated source code transformations on fourth generation languages’, in Eighth European Conference on Software Maintenance and Reengineering, 2004. CSMR 2004. Proceedings., Mar. 2004, pp. 214–220, doi: 10.1109/CSMR.2004.1281422.	IEEE
PS22	J. Stigliano and M. Bruni, ‘End-User Computing Tools’, in Encyclopedia of Information Systems, H. Bidgoli, Ed. New York: Elsevier, 2003, pp. 127–139.	ScienceDirect
PS23	C. Shipley and S. Jodis, ‘Programming Languages Classification’, in Encyclopedia of Information Systems, H. Bidgoli, Ed. New York: Elsevier, 2003, pp. 545–552.	ScienceDirect
PS24	A. Arkusinski and E. Green, ‘A software port from a standalone communications management unit to an integrated platform’, in Proceedings. The 21st Digital Avionics Systems Conference, Oct. 2002, vol. 1, pp. 6B3-6B3, doi: 10.1109/DASC.2002.1067987.	IEEE
PS25	G. Baster, P. Konana, and J. E. Scott, ‘Business components: a case study of bankers trust Australia limited’, Communications of the ACM, vol. 44, no. 5, pp. 92–98, 2001.	ACM
PS26	G. Canfora, A. Cimitile, A. De Lucia, and G. A. Di Lucca, ‘Decomposing legacy programs: a first step towards migrating to client–server platforms’, Journal of Systems and Software, vol. 54, no. 2, pp. 99–110, Oct. 2000, doi: 10.1016/S0164-1212(00)00030-3.	ScienceDirect
PS27	S. Barker and A. Monday, ‘Business students in information systems: wizards or apprentices?’, in Proceedings of the Australasian conference on Computing education, 2000, pp. 6–11.	ACM
PS28	M. B. Albizuri-Romero, ‘A retrospective view of CASE tools adoption’, ACM SIGSOFT Software Engineering Notes, vol. 25, no. 2, pp. 46–50, 2000.	ACM

Measuring the Performance of Inventory Management System using Arena Simulator

Fawaz J. Alsolami
Computer Science Department
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

Abstract—**The demonstration of inventory management systems is presented in this study to deal with a situation, where organizations are facing challenges due to uncertain behavior of demand. The implementation conducted using simulation technique to generate sampling experiment through computing and statistical methods. The important aspect of the simulation is to understand level of satisfaction regarding proposed system and its attributes. To assist the organization in controlling and managing the inventory system, the research provides the solution using Arena simulation tool. Firstly, the study explained the use of simulation and then reviews the common approaches of simulation. Secondly, the research proposed a framework for inventory control system. For practical implementation Arena tool used for model implementation. The main purpose of this experiment is to measure and analyze the applicability of potential system using a simulation tool. The results indicated the successful implementation of the proposed framework for inventory system. The model used multiple inventory system's variables such as demand, inventory stock and realized cost. For demonstrating the real world system's behavior, we used stochastic model for demand. The model executed using single server queuing (M/M/1) approach, but replicated several times. The results highlighted the high performances of machine located on multiples places and processed demand requests on each replication run. The study presented in this research demonstrates the association between demand and inventory. The proposed model can support the manufacturing organizations to control and manage inventory system.**

Keywords—Inventory management system; simulation; arena simulation tool; demand and inventory

I. Introduction

Simulation is a method of imitating the real world scenario on a computing system using different technology and tools. Commonly, it is used for experiment and get sampling data using a computing techniques [1]. The main purpose of this research is to apply an efficient implementation of inventory management and control system using simulation strategy. The simulation analysis performed using time period factor, which is one of the important factors to realize the performance prior to implement the actual system. In addition, the simulation helps to create assumptions related to system and user's behavior, operational analysis, system effectiveness, and evaluation strategies [2].

Model plays an important role in performing simulation. As the variables defined inside the model provide real description of the system. Those variable are also known as measuring factors, which used for collecting sampling data from the selected population using a computer program [3]. With the help of the models the experiment can perform to dig out the probability of system implementation and behavior using different statistical and computing strategies [4]. Therefore, efficient model designing with all possible variables is the critical phase before simulation. Whereas, the testing and validation of models executed with the help of simulation. For this reason, most of the time the terminologies modeling and simulation are often used together [5]–[7], for the same purpose, to do the sampling experiment.

The general steps of applying simulation start from conceptual model → use of tool to verify the system → and finally validate the system as final output [8]. There are different factors and sub-tasks need to be performed on every step as shown in Fig. 1. The figure proposed earlier in the research showing the association between different phases during whole experiment. The first step in this figure is initiated by developing a conceptual model of an already implemented system that need to be enhanced or a new system to be developed. The second phase is to convert conceptual model into statistical, mathematical, or graphical framework. Third step is to implement the model using computing tool, which can generate statistical report for measuring the performance of the factors. Finally, the last phase is to validate the simulation performance using generated output. If the output is in the favor of the system it can be finalized, else the process will restart after proper modification in the conceptual model again. In this way, the trial and implement procedure applied by repeating the steps of implementation, which is the core of simulation.

Simulation is applied for many disciplines such as science, technology, health industry, manufacturing, education and others [2]. A quantitative analysis performed by implementing the supply chain management system using simulation strategies [9], [10]. The idea for controlling the inventory for agricultural product using simulation is presented by [11] using multiple attributes such as transportation cost, goods lost cost, and shortage cost. The research article suggested that the simulation tool can help in measuring the performance of inventory system and can assist the organization to develop the monitoring policy. The monitoring policy is used to create positive relationship between demand and supply variables. Overall, the process will further help the organization in building efficient inventory control system.

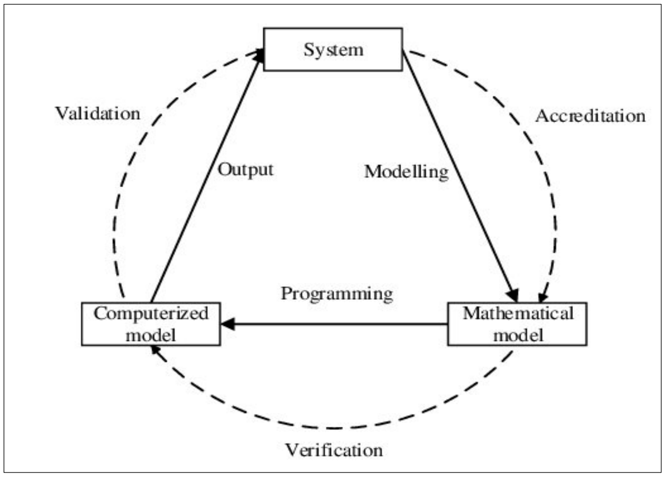


Fig. 1. Steps of Simulation [8].

The main problem has taken in this study is, “How an organization can improve and manage their inventory management system?” Therefore, the research proposes and validates the model for improving the inventory system using demand, supply and shortage variables. For this, the model validation applied using simulation tool, and to measure the performance of the model, replication applied several times. The research will help the organization to understand the behavior and impact of different variables of inventory system identified and analyzed in this research.

The rest of the paper is organized as follows. The next section discussed the research context of inventory control and management system. Section III described the differences between spreadsheet and Arena simulation. Afterwards, the proposed framework of inventory management system discussed in Section IV. Whereas, Section V illustrates the implementation of proposed model and results analysis generated by the simulation tool. Finally, the research ends by mentioning about final words and future work to enhance the research idea provided in this study.

II. Inventory Control Management System - Overview and Research Context

Generally, Inventory Control and Management System (ICMS) has an important role in the organization, which integrated with different other system such as customer management system, marketing, and manufacturing. This kind of system is mainly responsible to deal with multiple operational activities, and it is one of the major requirements of the organization as the system contains detailed information of every product [12]. There are different attributes highlighted in previous research that handled by the ICMS. Modeling and simulation for ICMS applied and reviewed several time. There are different applications used for validating the model. Number of available products, demand and supply ratio, order and shortage cost, transportation and raw material cost are example of attributes used for ICMS model simulation [11].

Inventory management system can be a part of any organization dealing with customer and offering services. It usage in healthcare industry presented by [13] to improvise the services provided by the hospitals. This is one the critical organizations need to be operated their business processes without any delay and mistakes as dealing with different types of patients and lifesaving medicines. Inventory control of medicine and other surgical instruments are very essential to be managed properly. The study suggested major characteristics of inventory system operating in healthcare department. Some of the characteristics mentioned are, rapid and continuous changes in orders, use of medical accessories, availability of beds and surgeons, no definite time of patients to check out the hospital, and the dependency of medical accessories on each other.

Another approach of ICMS presented in the form of multilevel approach used by [11] to deal the problem of agricultural products. The purpose of that system was to control the request in different layers to improve the overall performance of ICMS. The simulation results showed that the multilevel hierarchy helped them to deal with technical issues and to develop inventory control policy. The hierarchy presented in Fig. 2, showing the association between suppliers, manufacturers, retailers, and customers in ICMS. The hierarchy is in generic format, which can be useful for any industry. But the important thing in this approach is the development of communication channels to control the inventory system. The journal of simulation published an article that proposed the hybrid approach improving the performance of inventory system [14]. The article suggested number of important factor during simulation of ICMS that needs to be taken under consideration during ICMS development. Those factors are; uncertainty in demand and supply, the ratio of returned items, and the association between manufacturing and remanufacturing. The algorithm used in that study was Mixed-Integer Linear Programming.

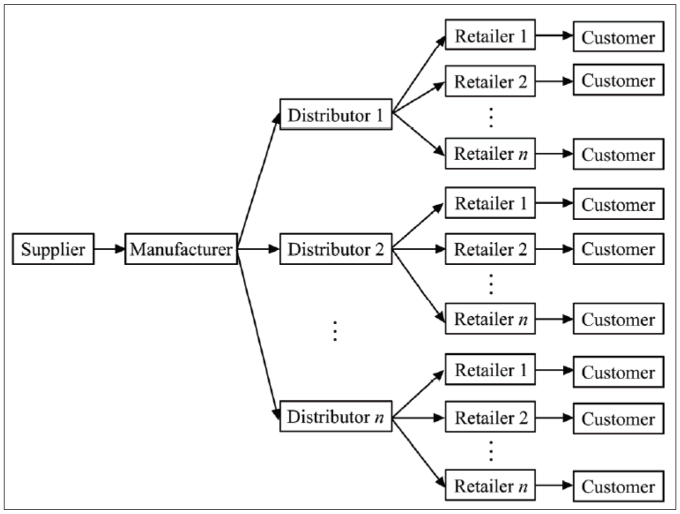


Fig. 2. The Working Scenario of ICMS [11].

Furthermore, the same kind of idea presented that described the level of uncertainty in producing the quantity and maintaining the quality of the products [15]. The proposed model was developed for the concept of hybrid manufacturing system. As previous literature suggested that inventory models are using two different types of approaches; Deterministic and Stochastic. The first one is used for the model where all variables are known with certainty. Whereas the second one is representing where demand and products involved uncertainty [16]. In order to cover up the uncertainty and matching with demand and supply, the research presented the idea for adding a layer of multiple supplier [17]. The simulation strategy was applied to validate the propose model, which also highlights the use of simulation technique for validating the ICMS model.

The latest technology has already changed the working environment in every organization. Internet of Things (IoT) [12], cloud computing [18], and big data [19] are some current development in computing field that allows the organization to enhance their information technology infrastructure. Accordingly, taking the advantage of current technologies [20] proposed the idea of automating the inventory process applied in warehouses. The purpose was to reduce the human efforts and automate the whole process. In addition, the forecasting models for inventory system suggested in a research by combining the forecasting parameters with inventory policy metrics [21]. The inventory models can work using batch approach. The batch approach means, where the demand received in a bunch. A research discussed this scenario where they check the execution of this approach using two models known as single server and multi-server [22]. In addition, the advanced version of moving average method applied in the research to implement the idea of downstream interface for supply chain management by reducing inventory levels [23]. That research is very new and the journal just published pre-proof for that research yet, which shows the importance and newness in the research idea undertaken in this study. Another latest research conducted the simulation of inventory management on a real vending network system [24]. The model used stochastic approach for dealing with demand attribute, which is one of the major constraints in inventory system.

This section discussed and put stressed on presenting the comprehensive overview and research context of ICMS. The inventory system is widely used for controlling the product's in and out records in the organization. Regarding ICMS, the following assumptions are extracted from the literature review:

- ICMS is one of major components in almost every organization.
- The major attributes of ICMS are demand, order, supply, shortage, product's cost, and availability of the products.
- Deterministic and Stochastic are two kinds of model used for dealing with demands.
- The major limitation founds in ICMS are uncertainty of demand and orders, cost of products, seasonal purchasing, product's return, and risk involvement in each process.
- Simulation is very commonly applied for improving the ICMS.
- The major tool has been used for ICMS simulation are, Spreadsheet Simulation and Arena.

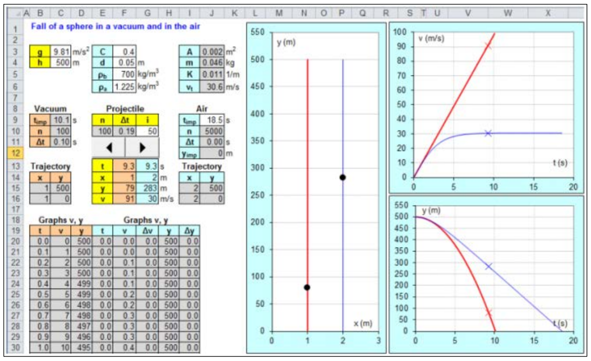
III. Simulation Tools

Simulation is used for doing an experiment based on abstract model, to analyze, measure, and understand its real implementation [25]. It has many phases to execute this process, started from building a conceptual model using different kinds of variables. The variables can be independent, mediating, and dependent based on the requirements of the system. Whereas, the dependent variable is also known as decision variable, which directly associated with independent variables [26]. This section describes the common tools used for simulation experiment as follows.

A. *Simulation using Spreadsheet*

Spreadsheet is one of the traditional program used for various purposes. First time, the concept of spreadsheet initiated by VisiCalc in 1979, which had restricted functionalities but the idea was prominent that further leads to progressive development till now [27]. The latest version of Microsoft Excel has the efficacy to collaborate with different other applications such as visual basic, statistical tool, simulation, decision support system and others [28]. The idea of spreadsheet simulation executed by [3] where they performed the implementation using inventory management system's example. In addition, there are other published work as well discussed and used the concept of spreadsheet simulation [29], [30]. A graphical representation of spreadsheet simulation is shown in Fig. 3. There are number of factors and characteristics highlighted for spreadsheet tool related to simulation as discussed below [3].

- It provides flexible way to initiate variables and create association between the variables.
- The strong formula and equation development option is there to use and generate results.
- Generating of random number and other number distribution option can facilitate to perform simulation.
- Use of data table to recreate the situation and repeat the experiment multiple times.
- The visualization techniques are there to illustrate the input, process and output procedures.
- Easy access to external application using add-in option.
- Overall, automated, integration, and flexible environment.



Fall of a sphere in a vacuum and in the air

g	9.81	m/s²
h	500	m

C	0.4	
d	0.05	m
ρ_b	700	kg/m³
ρ_a	1.225	kg/m³

A	0.002	m²
m	0.046	kg
K	0.011	1/m
v_t	30.6	m/s

Vacuum

t_{imp}	10.1	s
n	100	
Δt	0.10	s

Projectile

n	Δt	i
100	0.19	50

Air

t_{imp}	18.5	s
n	5000	
Δt	0.00	s
y_{imp}	0	m

Trajectory

x	y
1	500
1	0

t	9.3	9.3	s
x	1	2	m
y	79	283	m
v	91	30	m/s

Trajectory

x	y
2	500
2	0

Graphs v, y — Graphs v, y

t	v	y	t	v	Δv	y	Δy
0.0	0	500	0.0	0.0	0.0	500	0.0
0.1	1	500	0.0	0.0	0.0	500	0.0
0.2	2	500	0.0	0.1	0.0	500	0.0
0.3	3	500	0.0	0.1	0.0	500	0.0
0.4	4	499	0.0	0.1	0.0	500	0.0
0.5	5	499	0.0	0.2	0.0	500	0.0
0.6	6	498	0.0	0.2	0.0	500	0.0
0.7	7	498	0.0	0.3	0.0	500	0.0
0.8	8	497	0.0	0.3	0.0	500	0.0
0.9	9	496	0.0	0.3	0.0	500	0.0
1.0	10	495	0.0	0.4	0.0	500	0.0

Fig. 3. Simulation Experiment using Spreadsheet [31].

Apart from simulation, there are many additional operations such as decision support system, programming facilities, data organization, statistical analysis and others are available in excel. The list of major functionalities provided by spreadsheet, but not limited provided by [32] as defined below.

- There are big number of built-in functions available in excel such as financial, statistical, pivot table, database.
- Easy and integrated tool for visualizing the dataset using graphs and charts options.
- The facility for linear programming, visual basic programming, and mathematical programming.
- The tool has the ability to integrate with different tool like risk analysis, sensitivity analysis, decision analysis and others.

B. Simulation using Arena

Nowadays, a latest tool using for simulation by many industries is known as Arena that has been designed by Rockwell Software, Microsoft [33]. The tool is common for conducting experiment and collecting samples to estimate the chances of the proposed system's implementation. The major industries assisting help from this tool are many such as hospitals, inventory systems, order delivery systems, queue estimating process and others [34]. It provide the flexible environment where all of the operators are readily available and just required to connect them during process development. It also has the facility to develop model and transform in Arena, where the model can easily convert in its environment [4].

The tool has widely used in previous researches, defining the high applicability and use of the tool in academics and industries. Development of inventory control model with all specifications in manufacturing industry applied [14] to analyze and simulate the proposed model. Another research presented the common ideas can be applied using this tool. The research suggested the tool can be useful for the industries such as manufacturing, supply chain, and inventory system [35]. That article presented various models development strategies using Arena. Product manufacturing, consumer market, market demand, and component ordering simulation model presented with all possible steps.

The working environment in Arena has divided in two main layers called "Flowchart" and "Spreadsheet". There is a panel in the tool describing all available processes under the categories of "Basic" and "Advanced". The process development is easy to handle as there is no programming involve. The list of operator can be used by drag and drop, where the further specifications can be filled by double click on the operator. Mainly, the process development in this tool used the same programing strategy and connected as input → process → output. Following are the major characteristics provided by this tool.

- It provides the way to generate number of variables with all parameters.
- It offers to create different types of resources such as machine, materials, and components.
- The entities queue can be created for resources use.
- The entities can arrive as one by one or in batch.
- Different statistical methods can apply for generating number distribution.
- The option is available to define own expression for calculating and building user's own formula.

IV. ICMS Simulation – Proposed Framework

To support the idea presented in this research, this section proposed a framework for ICMS simulation as shown in Fig. 4. The idea of this framework is based on the inventory model presented and simulated using spreadsheet tool [3]. The original idea was applied in spreadsheet simulation, while this research implemented the simulation using Arena tool. The inventory model example used in this study is associated with sport club players. Every day, number of players are presenting in the club for different activities. Based on the experience data, it has been found that most of time the demand is either got short or excess based on the attendance.

Due to this fact the sports club management always got in trouble for not fulfilling the requirements of players. Moreover, the shortage and excess of demand cause them financial and other loss. The proposed model shown in Fig. 4 illustrating the simulation approach will be applied in this study using Arena. There are six major phases mentioned in the model. The definition of each phase is defined below, whereas the actual list of variables and values applied in Arena is described in the next section.

1) Variable Identification: This phase is to identify number of variables to be used in the simulation. It mainly depends on system's specification.

2) Input Values to the Variables: This research used the demand variable as stochastic, where the distribution will be known but uncertainty will be involved to make it dynamic variable. For other variables can be taken as dynamic or fixed.

3) Computation for each Variable: The calculation of each variable using a prescribed formulae and notation. This calculation is prior to simulation, means if any variable required any extra computation or assignment.

4) Calculation if Demand is More or Less: What happen if demand is less or more? Here we will use a decision statement for both conditions. But the calculation will be different. It will help to compute loss for both conditions.

5) Run Simulation and Replication: But in both ways simulation will run. For more clear understanding we can replicate the simulation several times.

6) Result Analysis: The result can show the impact of shortage and excess in demand. It will show the financial loss due to not managing the inventory system properly. The result will be displayed using statistical and visualization tools.

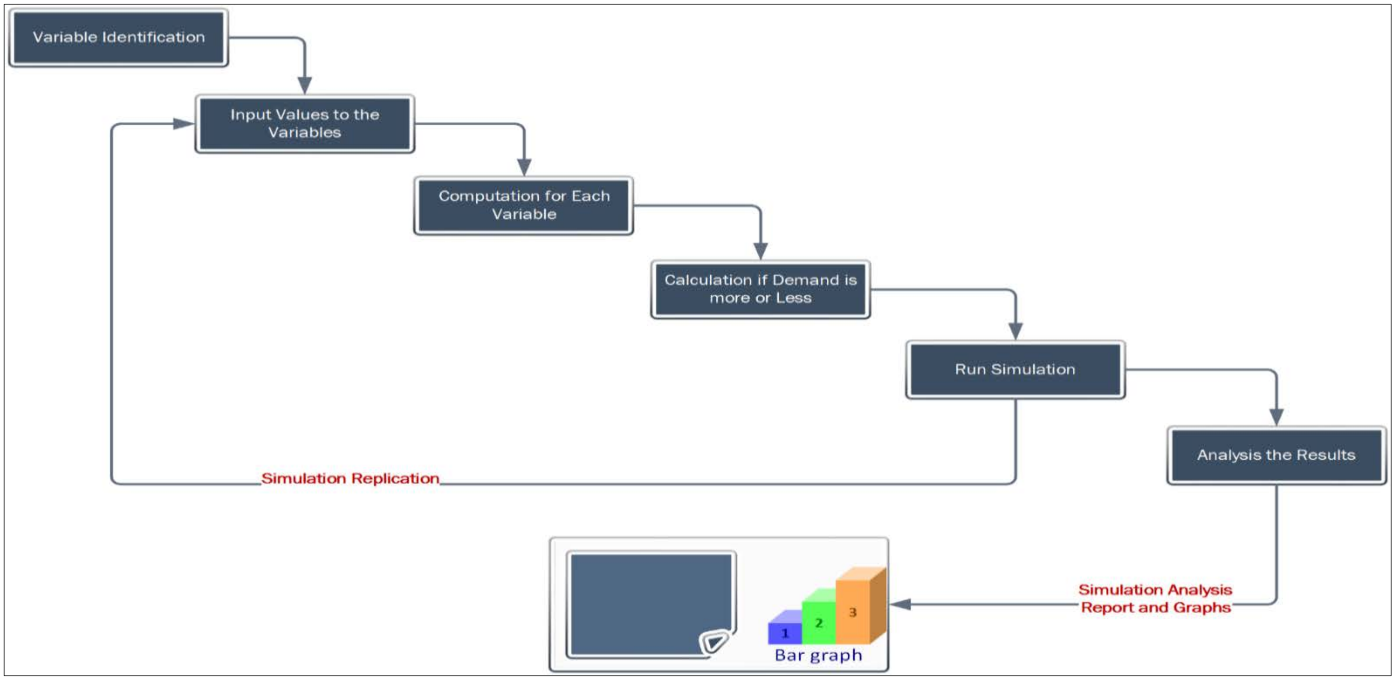

Variable Identification
Input Values to the Variables
Computation for Each Variable
Calculation if Demand is more or Less
Run Simulation
Analysis the Results
Simulation Replication
Simulation Analysis Report and Graphs
1
2
3
Bar graph

Fig. 4. Proposed Model for Inventory System Simulation.

V. Framework Implementation using Arena

The main research problem undertaken in this study is to support organization by developing inventory control model using dynamic values of demands. Another aspect of this study, to automate the overall process of inventory control and management system. Therefore, simulation technique used to test and validate the model. The Arena software is used for implementation of the model. Normally, the model should have quantitative or qualitative values associated with one or many attributes identifying the operational work of the proposed system. Finally, the output of the experiment can highlight the strength and weaknesses of the proposed system. The output can lead the organization to analyze the probability of system implementation in real world. The subsequent section is defined the model specification, variable identification, model implementation, and result analysis.

A. Inventory Model - Variable Identification

Mainly, the inventory model based on the variable like demand generates by the customer. The model considered in this study is based on controlling the inventory of a sport club. The scenario of club is, there are many registered club members visiting club to do sports activities. One of the services provided by the club is to offer them food on every visit. The main complexity in arranging the food is the unknown number of visitors. Therefore, to solve this issue they need a system which can control and manage their food arrangement inventory system. The same problem is implemented using spreadsheet simulation, while in this study we proposed a model to be implemented in Arena. The main problem in this scenario is the financial loss because of not matching the actual demand and items availability. The number of variable used in this model illustrated in Table I. There are two different situations that organization is facing within this problem also mentioned in this table.

TABLE I. Variables Specification

Variable	Specification
Players Arrival	**Entities per Arrival:** 10 (Batch) **Distribution:** Random **Simulation Time:** 1 hour
Players Demand	**Assign:** Demand and inventory stock variable Initiation **Distribution:** Random
Check Inventory?	**Decision Box:** Check if inventory has the available demand?
If Demand Excess than Inventory, Lost?	**Computation:** It will cost $60 per item
If Demand Shortage than Inventory, Lost?	**Computation:** It will cost $160 per item
Reprocess and Record Lost Amount for Excess	**Reprocess:** by putting another resource machine for recovery. **Record:** the total amount lost for excess demand.
Reprocess and Record Lost Amount for Shortage	**Reprocess:** by putting another resource machine for recovery. **Record:** the total amount lost for shortage demand.

B. Model Implementation and Result Discussion

As shown in the Fig. 5, the simulation started after the players arriving to the premises in random number. The arrival is in batch, means group of 10 members arriving in every event. The simulation run for 1 hour and replicated for 10 times. The arrival of the players, by default create the variable “order”, which shows that each players will required a meal box. The next step in the simulation is the assignment of variable “demand” using player’s arrival information. Therefore, the demand variable will be created here same as number of players arriving after prescribed time using random distribution.

The purpose of this assignment is that demand variable will be used at different location in later simulation step. Meanwhile, at the same step another variable will be initiated called "inventory" stock available. It will further validate the difference between demand and stock availability. The processing of players orders then execute based on the availability of machine (resource) provided at next step known as "process". There will be a different processing time for each order. The working time averages is discussed in further result section. Therefore, the processing machine will use some time-delay in processing the orders. The next step is the decision box for validating the demand value with inventory stock. Here the condition applied that "if inventory is greater and equal to Demand". If the statement is "true" it will go to "demand excess" node. The decision box implementation is the fourth step in this simulation.

Moving forward, the next step is to reprocess the simulation in both condition, either demand is less or more. Adding a machine over here is indicating that the additional time is required to process the request for both scenario. The final step in this simulation to record the different variables executed in this simulation. At last, we run the simulation 10 times to analyze the different situations, calculate the averages, and check the average lost in 10 simulations.

The results indicates the expected cost computed for both situation, if demand is less than the inventory, or if demand is more than the inventory. As discussed earlier, the spreadsheet simulation applied earlier by [3], is used in this study to apply the simulation using Arena. Therefore, the result generated based on the following modified equation.

$$z = c_e(y - D)\ if\ y \geq D \qquad (1)$$

If inventory (y) is greater than equal to the demand (D). It represents that there will be excess cost (C_e) applied, to calculate the realized cost (z). In this scenario, the realized cost, will be computed by subtracting the inventory from demand, then multiply it by cost of excess that is $60 as mentioned in Table I. We calculated the values for each simulation to know the realized cost in all simulation, is shown in Table II. The next formula applied is for computing the shortage cost as described below.

$$z = c_s(D - y)\ if\ y < D \qquad (2)$$

If inventory (y) is less than to the demand (D). It represents that there will be shortage cost (Cs) applied, to calculate the realized cost (z). In this scenario, the realized cost, will be computed by subtracting the demand from inventory, then multiply it by cost of shortage that is $160 as mentioned in Table I. We calculated the values for each simulation to know the realized cost in all simulation, is shown in Table II.

After all discussion, and how the values are computed to understand the financial loss using a sample experiment conducted in this study. The table representing the summary of simulation, which denotes the samples collected and realized cost. The objective of this simulation is to find out the optimal association between demand and inventory stocks. Searching of ideal values cannot generate by single run, therefore, the experiment conducted several times to create multiple values and analyze the results. The experiment used a sample cost in each replication to calculate the shortage and excess found in inventory and demand. The purpose was to idealize the best scenario and association between both variables. One of the screenshots of using the variables in the model and assigning some input values is shown in Fig. 6.

It can be evident from Table II that out of 10 simulation run, replication number 4 gives the lowest financial lost (60$), whereas replication number 10 provides the maximum lost to the sports club that is 11680$. Another ideology to be considered from this result is, if inventory is more than the demand the lost amount is less. On the other side, if inventory is less than demand the lost amount is high. Based on the result generated, the sports club can create a strategy, to keep the inventory stock as high as they can do, as the financial loss is minor in this situation. Although, the experiment used a sample of data generated randomly, but the idea was to validate the inventory model for sport club, supported by different variables.

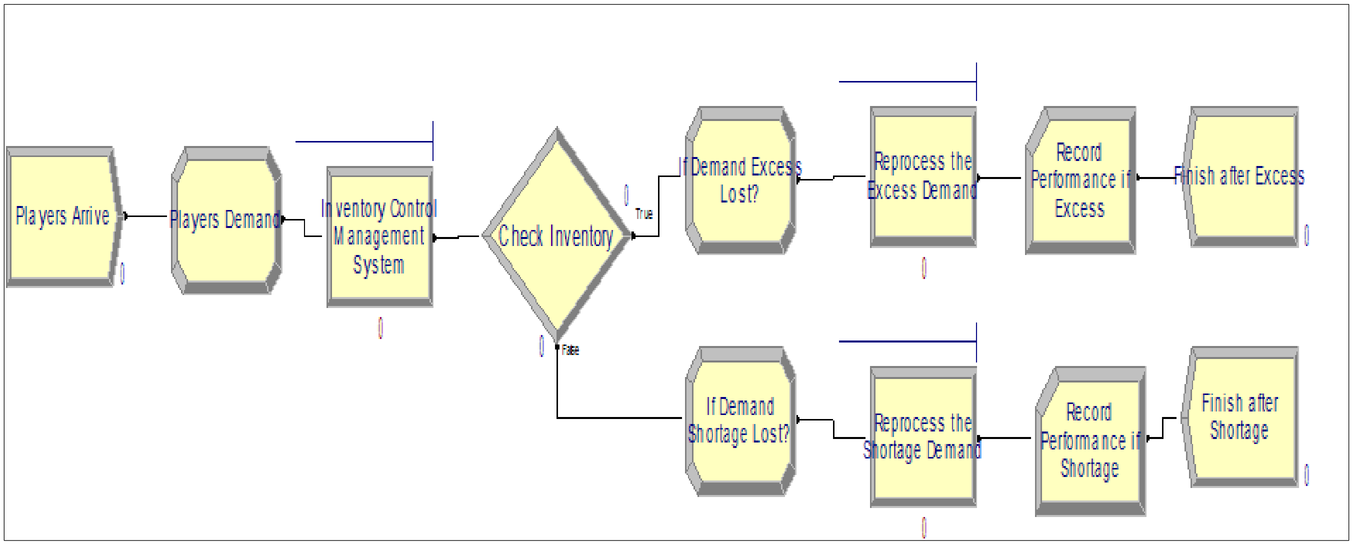
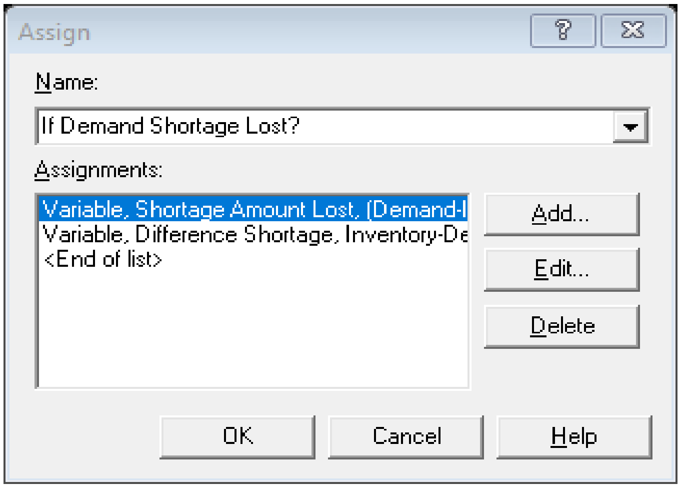

Players Arrive
0
Players Demand
Inventory Control Management System
0
Check Inventory
True
0
False
If Demand Excess Lost?
Reprocess the Excess Demand
0
Record Performance if Excess
Finish after Excess
0
If Demand Shortage Lost?
Reprocess the Shortage Demand
0
Record Performance if Shortage
Finish after Shortage
0

Fig. 5. Screenshot of Model Implementation in Arena Simulator.



Name:
If Demand Shortage Lost?

Assignments:
Variable, Shortage Amount Lost, (Demand-I
Variable, Difference Shortage, Inventory-De
<End of list>

Fig. 6. Input and Assignment to the Variables.

TABLE II. SIMULATION RESULTS

Replication #	**Demand**	**Inventory**	**Realized Cost**
1	57	13	7040$
2	24	63	2340$
3	26	52	1560$
4	84	85	60$
5	40	48	480$
6	53	56	180$
7	40	2	6080$
8	26	47	1260$
9	59	12	7520$
10	95	22	11680$
Mean Values	**50.4**	**40**	**2644$**

In addition, the mean values generated in the simulation indicating that, overall the demand value is higher than the inventory stock. In 10 replications, the mean value of customer's demand recorded as 50.4, and store stock are 40, while the average lost recorded is 2644$. The number of replication is few in this experiment but based on the result, it can be suggested to the organization to increase the inventory stock to match it with the demand. Although, it should be clear that if there is a difference in stock and demand, it will always damage the organization reputation.

Finally, to understand the different machine's performance at each station, Table III is representing the details of each machine. In this table, the recorded time is representing the average working time during all replications. It can be seen that, the average working time is recorded at "inventory control system" is the highest, as this was the first machine and all requests received at this station. After that the request was distributing and forwarding to two different nodes. Like in some replication it went towards the "Excess Demand Reprocess" or sometime forwarded to "Shortage Demand Reprocess". Therefore, their working time is considerably lower than "inventory control machine".

TABLE III. OVERALL WORKING TIME AT EACH STATION

Station	**Overall Average Values**	**Overall Minimum Values**	**Overall Maximum Values**
Inventory Control System	0.87 minutes	0.36 minutes	0.90 minutes
Reprocess the Excess Demand	0.085 minutes	0.05 minutes	0.11 minutes
Reporcess the Shortage Demand	0.074 minutes	0.03 minutes	0.13 minutes

VI. CONCLUSION

The research proposed a model to support organization in dealing with inventory control and management system. The idea is useful for those organizations to take decision based on the generated results. The implementation of proposed model executed with the help of simulation strategy. Specifically, the model used stochastic model of demand, where the variable were uncertain and selected randomly in each replication. The inventory optimization is supposed to be a complex situation, where all organizations accepting unknown damage every time. The research is proposed a model with limited variables to support this idea, which can be enhanced by adding more variables, machines, and stations. The result of the simulation guided that availability of stock will cause less damage, while having lower stock than demand is always giving lost to the organization. It will not only cause the financial lost but it will also degrade the company's reputation.

VII. FUTURE WORK

The model can be improved using demand and supply variables with multi-server strategy. In this study, the researcher used the single server model, where the multiple of entities served with one machine. The idea can be amended by implementing multiple server approach on different scenario. The multiple server approach can provide the analysis and performance of different machine, which can also be used to compare the performances of different machines. The simulation results can provide the statistics to the organization in order to take the decision based on the performance of the model.

REFERENCE

[1] G. S. Fishman, Monte carlo concepts, algorithms and applications. New York: Springer, 1996.

[2] R. D. Smith, "Simulation," Encycl. Comput. Sci., 2000.

[3] A. F. Seila, "SPREADSHEET SIMULATION," in Proceedings of the 2006 Winter Simulation Conference, 2006, pp. 11–18.

[4] Y. Merkuryev and A. Grinbergs, "Design of UML models and their simulation using ARENA," WSEAS Trans. Comput. Res., vol. 3, no. 1, pp. 67–73, 2008.

[5] H.-J. Bungartz, S. Zimmer, M. Buchholz, and D. Pflüger, Modeling and Simulation: An Application-Oriented Introduction. Springer-Verlag Berlin Heidelberg, 2014.

[6] S. Raczynski, Modeling and Simulation: The Computer Science of Illusion. Wiley Online Library, 2006.

[7] B. Zeigler, A. Muzy, and E. Kofman, Theory of Modeling and Simulation, 3rd ed. Elsevier, 2018.

[8] D. Gao, X. Xu, J. Yin, H. Zhang, and B. Zhang, "SDG and qualitative trend based model multiple scale validation," in IOP Conference Series: Materials Science and Engineering, 2017, vol. 231, no. 1.

[9] A. J. Schmitt and M. Singh, "A quantitative analysis of disruption risk in

a multi-echelon supply chain," Int. J. Prod. Econ., vol. 139, no. 1, pp. 22–32, 2012.

[10] T. Peirleitner, A. J., Altendorfer, K., & Felberbauer, "A simulation approach for multi-stage supply chain optimization to analyze real world transportation effects.," in 2016 Winter Simulation Conference (WSC), 2016, pp. 2272–2283.

[11] G. Xu, J. Feng, F. Chen, H. Wang, and Z. Wang, "Simulation-based optimization of control policy on multi-echelon inventory system for fresh agricultural products," Int. J. Agric. Biol. Eng., vol. 12, no. 2, pp. 184–194, 2019.

[12] B. S. S. Tejesh and S. Neeraja, "Warehouse inventory management system using IoT and open source framework," Alexandria Eng. J., vol. 57, no. 4, pp. 3817–3823, 2018.

[13] E. Saha and P. K. Ray, "Modelling and analysis of inventory management systems in healthcare: A review and reflections," Comput. Ind. Eng., p. 106051, 2019.

[14] P. Thammatadatrakul and N. Chiadamrong, "Optimal inventory control policy of a hybrid manufacturing–remanufacturing system using a hybrid simulation optimisation algorithm," J. Simul., vol. 13, no. 1, pp. 14–27, 2019.

[15] H. Rachih, F. Z. Mhada, and R. Chiheb, "Simulation of a stochastic inventory model for a hybrid manufacturing-remanufacturing system," in 2019 International Colloquium on Logistics and Supply Chain Management (LOGISTIQUA), 2019, pp. 1–6.

[16] A. E. E. J. F. Moritz, J. M. Bloemhof-Ruwaard, V. der Laan, E. Van Nunen and L. N. Van Wassenhove, "Quantitative models for reverse logistics: A review," Eur. J. Oper. Res., vol. 103, no. 1, 1997.

[17] A. Bartoszewicz and P. Latosiński, "Sliding mode control of inventory management systems with bounded batch size," Appl. Math. Model., vol. 66, pp. 296–304, 2019.

[18] A. Elragal and M. Haddara, "The Future of ERP Systems: look backward before moving forward," Procedia Technol., vol. 5, no. 2212, pp. 21–30, 2012.

[19] A. L'Heureux, K. Grolinger, H. El Yamany, and M. Capretz, "Machine Learning with Big Data: Challenges and Approaches," IEEE Access, 2017.

[20] A. M. Vamsi, P. Deepalakshmi, P. Nagaraj, A. Awasthi, and A. Raj, "IOT Based Autonomous Inventory Management for Warehouses," in EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing, 2020, pp. 371–376.

[21] N. Kourentzes, J. R. Trapero, and D. K. Barrow, "Optimising forecasting models for inventory planning," Int. J. Prod. Econ., p. 107597, 2019.

[22] S. R. Chakravarthy and A. Rumyantsev, "Analytical and simulation studies of queueing-inventory models with MAP demands in batches and positive phase type services," Simul. Model. Pract. Theory, p. 102092, 2020.

[23] Y. Tliche, A. Taghipour, and B. Canel-Depitre, "An Improved Forecasting Approach to Reduce Inventory Levels in Decentralized Supply Chains," Eur. J. Oper. Res., vol. Journal Pr, no. Pre-Proof, 2020.

[24] H. Grzybowska, B. Kerferd, C. Gretton, and S. T. Waller, "A simulation-optimisation genetic algorithm approach to product allocation in vending machine systems," Expert Syst. Appl., vol. 145, p. 113110, 2020.

[25] C. F. Cheung, C. M. Cheung, and S. K. Kwok, "A knowledge-based customization system for supply chain integration," Expert Syst. Appl., vol. 39, no. 4, pp. 3906–3924, 2012.

[26] E. Turban, R. Sharda, and D. Delen, Decision Support and Business Intelligence Systems, 10th ed. Pearson, 2010.

[27] D. Cassel, "How VisiCalc's Spreadsheets Changed the World," The New Stack, 2019.

[28] Microsoft, "Microsoft Excel.".

[29] W. L. Winston, "Simulation modeling using @risk," Belmont, California: Duxbury, 1996.

[30] A. F. Seila, V. Ceric, and P. Tadikamalla, "Applied simulation modeling," Belmont, California: Brooks-Cole, 2003.

[31] J. Benacka, "Numerical modelling with spreadsheets as a means to promote STEM to high school students," EURASIA J. Math. Sci. Technol. Educ., vol. 12, no. 4, pp. 947–964, 2016.

[32] C. T. Ragsdale, Spreadsheet modeling and decision analysis. Thomson south-western, 2004.

[33] R. S. Inc, "Arena User's Guide," USA, 2005.

[34] "Arena Simulation Software." 2000.

[35] G. E. Vieira, "Ideas for modeling and simulation of supply chains with arena," in Proceedings of the 2004 Winter Simulation Conference, 2004.

Investigating the Awareness and Usage of Moodle Features at Hashemite University

Haneen Hijazi[1], Ghadeer Al-Kateb[2]
Faculty of Information Technology
Hashemite University
Zarqa, Jordan

Eslam Alkhawaldeh[3]
Department of e-learning
Hashemite University
Zarqa, Jordan

***Abstract*—E-learning plays a vital role in the educational process. Learning management systems are being essential component of e-learning. Moodle learning management system is being widely used in Higher Education Institutions due to the rich features it provides that support the learning process. Standard Moodle comprises 21 features (14 activities and 7 resources). Little research has been carried out to examine these features in particular. In this research, the awareness and usage of Moodle features among faculty members at Hashemite University, Jordan are investigated. A sample of 140 instructors were surveyed. Then, the responses were analyzed to find the overall awareness and usage of each feature. Furthermore, the correlation between awareness and usage and how the awareness of Moodle features is associated with their usage were analyzed through correlation and regression analysis. The study revealed that instructors expressed highest awareness towards File, Folder, Assignment, URL and Quiz features whilst the least awareness was towards SCORM package and IMS content package features. Regarding usage, the study identified the File, Folder, Assignment and URL features as the most heavily used features whereas the least commonly used features have been IMS Content Package, SCORM package, Wiki, Glossary, Workshop, Database, Survey, External tool and Choice. Moreover, the study statistically demonstrated a strong correlation between the awareness and usage of features and that changes in the awareness of Moodle features are significantly associated with changes in their usage. In other words, the study revealed that features with low awareness tend to have low usage and that the usage would increase as the awareness increases. The study would help Moodle administrators in Higher Education Institutions decide about the most important features that should be installed in their customized instance of Moodle. Furthermore, the study would help Hashemite University responsible parties in identifying the least commonly used and the least well-known features, allowing them to focus on increasing the levels of awareness and usage of those features in a way that might reflect positively on the learning process.**

***Keywords*—Moodle; learning management system; features; awareness; usage; activities; resources; tools; correlation; regression**

I. Introduction

E-learning is being involved considerably in higher education. It is about the use of Information and Communications Technology (ICT) in delivering education either fully online or partially (i.e. blended learning). In either form, two major components constitute the infrastructure of any e-course. These components are the e-content and the Learning Management System (LMS).

A learning management system is a virtual learning environment that allows communication between students and instructors, and between the students themselves. The instructor through the LMS can upload material, assignments, quizzes and answer questions, while students can read material, ask questions, communicate with each other and do assignments and quizzes. All these activities are performed online using the features that are embedded in the LMS.

Moodle is a “learning platform designed to provide educators, administrators and learners with a single robust, secure and integrated system to create personalized learning environments” [1]. “Moodle” stands for “Modular Object-Oriented Dynamic Learning Environment” [2]. It is an open-source LMS, hence, its use to support the learning process is popular in HEIs all over the world. An administrator can download it for free, customize it, and participate in several public forums to discuss Moodle issues.

Latest version of Moodle (3.8) was released in Nov 2019 [3]. Standard Moodle offers several learning features. The feature that a student uses [4] to interact with other students and/or the teacher is called “Activity”. In contrast, the feature that is presented by the teacher [4] to the students in purpose of supporting learning is called “Resource”.

According to [5], standard Moodle encompasses 14 different types of activities, these activities are: “Assignment” activity; which allows students to do the created assignment online and allows instructors to grade and give feedback. “Chat” activity; that allows participants to communicate synchronously. “Choice” activity to allow students to answer a multiple-choice question. “Database” activity; which allows participants to create, maintain and search a set of record entries. “Feedback” activity; that allows instructors to create custom surveys through a variety types of questions to collect feedback from students. “Forum” activity; allows participants to communicate asynchronously. “Glossary” activity; that enables participants to create and maintain a list of definitions. “Lesson” activity; to enable instructors to design and deliver instructional content. “(LTI) External tool”; enables students to interact with learning resources and activities on other web sites. “Quiz” activity; which enables instructor to design a test comprising questions of various types, grade it automatically or manually, and give feedback. A “SCORM” activity that is used to present multimedia content and to assess students according to an agreed standard for learning object. “Survey” activity allows instructors to assess their courses through several verified survey instruments. “Wiki” activity; enables participant to create and edit a collection of web pages which can be either collaborative or individual. And lastly, a “Workshop” activity the enables peer assessment of students' work.

Also, as per [6], standard Moodle provides 7 types of resources: “Book” resource; that is a multi-page resource in a book-like format. “File” resource; which is used to display images, documents, spreadsheets, presentations, sound files and videos. A “Folder” resource to organize files. “IMS content package” resource; that is used to add material from other sources as a collection of files packaged according to an agreed standard. “Label” resource which uses text and multimedia to separate course elements or to give descriptive information and instructions. “Page” resource; that is used to create a web page using the text editor or html editor, such pages may display text, media files, web links and embedded code, such as Google maps. Finally, a “URL” resource to provide a web link for a resource that is available outside.

In Jordan, almost all public universities are using Moodle as a LMS platform. Hashemite University (HU) is a public university in Jordan that was established in 1995. It has 15 faculties: Tourism and Heritage, Science, Applied Health Sciences, Nursing, Arts, Physical Education and Sport Science, Information Technology, Engineering, Medicine, Pharmaceutical Sciences, Economics and Administrative Science, Educational Sciences, Childhood, Natural Resources and Environment, and Arid Lands in addition to the Graduate Studies faculty.

HU adopted e-learning to support the educational process early. HU payed valued efforts to capitalize and employ e-learning technologies in the learning process [7]. As well, HU is one of the universities of Jordan that employed LMS in the learning process very early. In the early beginning, HU used Blackboard then, in the early 2012 [8], HU migrated to Moodle for being open-source and many other issues. The current version of Moodle that is installed at HU is 3.1.1. HU Moodle hosts over 1500 course. Most courses are traditional face to face courses that uses LMS and e-content to support the learning process while a few courses are being delivered online. In both forms, LMS plays a vital role in the learning process.

In HU, Moodle is configured so that students are enrolled automatically when they register for a course. This customized version of HU includes all features that are available in the standard Moodle. The usage of Moodle features depends on several factors; it depends on its usefulness and usability and how lecturers perceive them [4]. The choice and usage of a certain feature depend on its impact on the lecturer workload and how much time it saves [9] [10]. Moreover, feature usage depends the available equipment, Internet connection and the availability of other tools and features [10].

Principally, this research investigates the awareness and usage of Moodle features among HU faculty. The awareness could be described as the degree of perception that the instructors have towards each Moodle feature and its specific use. The most well-known features and the least well-known will be highlighted (RQ1). On the other hand, the usage describes how frequent each feature is used by instructors. The most commonly used features and the least commonly used ones will also be emphasized (RQ2). Theoretically, it is believed that the awareness of Moodle features is associated with their usage. Features with low awareness is expected to have low usage and this usage would increase as the awareness increases. However, this research aims to statistically investigate the relationship between the awareness and usage of Moodle features and how changes in the awareness are associated with changes in the usage (RQ3).

The main research questions are:

- RQ1: Which Moodle features have the highest awareness, and which features have the least awareness among HU faculty?
- RQ2: Which Moodle features are used most frequently, and which features are used least frequently by HU faculty?
- RQ3: Is there a statistical relationship between faculty’s awareness of Moodle features and their usage?
- RQ4: Based on (activities/resources) classification of Moodle features, which category has highest/least awareness and which category is most/least frequently used?
- RQ5: Based on the adopted functional classification of features, which categories have the highest/least awareness and which categories are the most/least frequently used?

The rest of this paper is organized as follows: Section 2 highlights the related work, Section 3 introduces the method, Section 4 presents the results, Section 5 discusses the results and limitations and Section 6 concludes the paper.

II. Related Work

The literature is rich of studies that aimed at evaluating Moodle as a learning management system, but very few studies examined its features [4]. The main purposes for using Moodle at Kajaani University of Applied Sciences, Finland were reported by KC [4]. KC also evaluated Moodle features (i.e. activities and resources) that were implemented in the Moodle instance they used. The author conducted an online questionnaire that targeted teaching staff and used qualitative weight and sum (QWS) approach. The customized instance of Moodle in Kajaani University embedded 12 activities: assignment, choice, database, feedback, forum, glossary, lesson, quiz, SCORM package, survey, wiki and workshop, and 6 resources: book, file, folder, label, page and URL. The results show that among all features, assignment, feedback, quiz and workshop modules were considered very essential and were heavily used. A comparison between Moodle and Facebook in delivering learning in LPU-L was conducted in [11]. The authors compared the overall acceptance, accessibility and usage of the common features (i.e. chat, groups, search engines, announcements and downloading/uploading files) among both instructors and students. They found that group feature is the most used feature in Facebook, whereas download/upload feature is the most used feature in Moodle. They recommended that Moodle should be improved in order to improve students' perception towards it.

Combined Qualitative Weight and Sum (QWS) and Analytic Hierarchy Process (AHP) approaches were conducted in [12] evaluate and compare 9 open-source LMSs based on 5 criteria, namely, social networking, productivity, administration, presentation, and management. The authors used 35 features classified into subcategories of the adopted 5 criteria. They found that the highest evaluated LMS is Intelligent Web Teacher (IWT), followed by Claroline and in the third place is the Moodle, whilst the most frequently used open source LMS in Jordan universities is Moodle. They claimed that the result of their study would help HEI to choose the proper LMS and to define the most important features to be activated according to students. In [13], Hasan investigated 24 features of Moodle on both desktop and mobile devices from students' perspective in a university in Jordan. She found that among the 24 Moodle features, only 6 features were installed and frequently used by students and 18 other features were required by students but not installed. Moreover, she evaluated the usability of the installed Moodle instance and came out with 17 usability problems. She also proposed 10 improvements according to students to enhance the usability of Moodle.

The use of Moodle tools and functionalities by the University of Aveiro students was analyzed in [14]. The authors used content analysis, structured interview with the Moodle administrator and a questionnaire that targeted the university students. They found, according to the questionnaire, that 98% of respondents used Moodle to download materials, 84% to see news, and that Moodle is least commonly used to deliver assignments, communicate with teachers and Ask questions. Teachers' individual information, teachers' frequency of use of Moodle activities (12 activity), and teachers' perception of how the use of Moodle impacts learning in secondary schools in Catalonia (Spain) were investigated [15]. The researchers found that assignment, quiz, forum, lesson and external tools were the most commonly used activities, while workshop, database and Wiki were the least commonly used. In their research, they did not tackle Moodle resources. The researchers in [16] explored satisfaction levels and usage of 8 features (Assignment, news/announcement, gradebook, events, online assessment, calendar and forum) of Moodle LMS for 47 faculty at CHS. Among the 8 features they investigated, assignment was the most commonly used and forum was the least.

A trial to improve student collective work for the "Multimedia for web-based e-learning" track in Trakia University, Stara Zagora, using the improved collaborative features in Moodle like glossary, wiki and forum was presented in [17]. Purbojo [18] collected Moodle logs files, reports, learning outcomes data, and interview data and performed quantitative and qualitative statistical analyses. Purbojo found that several behavioral characteristics exist for the instructor's role in utilizing Moodle features.

Researchers proposed several categories that can be used to classify LMS features based their functionalities. As per [19], LMSs features can be creation, organization, delivery, communication, collaboration and assessment. In [10], the authors adopted a 4-categories classification to group LMS features: distribution, communication, interaction and course administration. Another categorization by Hamtini and Fakhouri [12] involves Social networking, productivity and software installation, administration and security, presentation and material distribution, and management. Based on these categories, several researchers classified Moodle activities, particularly, based on their perceived functionalities. Costa, Alvelos and Teixeira in [14] adapted Piotrowski [19] categories and added the reusability category to classify Moodle features. They classified Moodle activities as follows: creation (database), organization (lessons), delivery (assignments, workshops), communication (chats, forums, news), collaboration (glossary, wikis), assessment (choice, quiz, survey, and feedback) and reusability of learning resources (SCORM, and external tools). Similarly, Badia, Martín, and Gómez in [15] adopted the categorization of Moodle features in [14] except that they did not tackle 3 features (i.e. news, feedback and SCORM). University of Massachusetts at Amherst [20] classified activities available in their Moodle into: assignments (Moodle Assignments, Turnitin Assignment, External Tool), communication and collaboration (chat, database, forum, glossary, wiki, workshop), Assessment and surveys (quiz, questionnaire, choice), management (attendance, checklist, group self-selection), and interactive delivery of content (lesson and SCORM).

Accordingly, the literature lacks such studies that have been concerned with evaluating the awareness and usage of all Moodle features that are available in the standard Moodle installation. Consequently, the relationship between the awareness and usage of Moodle features have never been investigated. Hence, evaluating Moodle features awareness and usage among HU faculty and the relationship between them is the focal goal of this research.

III. Method

An online survey that targeted HU's instructors was designed and distributed. The survey was developed using Google Forms. It started with the contact information of respondents, then for each Moodle feature two questions were asked:

- Describe your "Awareness" of "feature name"?
- Describe your "Usage" of "feature name"?

For the "awareness" questions, 5 choices were given (i.e. extremely aware, very aware, moderately aware, slightly aware, not at all aware). For the "usage" questions, 5 choices were given (i.e. always, often, sometimes, rarely, never).

As a first step, the researchers communicated with the university administration to ask faculty members to facilitate the process of distributing the survey stating that it is for research purposes. Then the survey was distributed via email among the majority of HU faculty who are currently on their work. The survey was conducted in the interval Feb 10th to Apr 7th from the year 2020.

Data were analyzed using IBM SPSS Statistics and Microsoft Excel. Survey items were classified into two groups. One group includes the items that assess the awareness of Moodle features and the other group includes the items that assess the usage of Moodle features.

In order to measure the reliability of the survey items, Cronbach’s alpha was used. Cronbach’s alpha is used to measure the internal consistency of a scale with a value lies between 0 and 1 [21]. As both the awareness and usage of Moodle features are evaluated in this survey, two Cronbach’s alpha values were calculated. The first Cronbach’s alpha measures the internal consistency of the first group (i.e. the items that evaluate the awareness). As the internal consistency describes the degree to which all the items in a test measure the same concept [21]; a higher value of alpha for awareness items would indicate that the items actually measure the awareness of the Moodle features. The other alpha is used to measure the internal consistency of the second group (i.e. the items that evaluate the usage). Likewise, a higher value of alpha for usage items would indicate that the items actually measure the usage. Several qualitative descriptors were reported in the literature to interpret the value of alpha, the majority agreed that a value around 0.7 or above is desirable [22].

In order to describe the overall awareness and the overall usage for each Moodle feature, mean value was calculated for each item. To advocate the results, the median and mode were also calculated. For these statistical purposes, a numerical rate between 5 and 1 was given for each response. For the “awareness” questions; extremely aware=5, very aware=4, moderately aware=3, slightly aware=2, not at all aware=1. For the “usage” questions, always=5, often=4, sometimes=3, rarely=2, never=1. The mean was calculated for each item by finding the summation of the numeric values of all responses, and then dividing by the number of respondents.

The features were ranked according to the overall awareness and usage based on means. Moreover, the median (i.e. the response for which 50% of responses are higher and 50% are lower) and the mode (i.e. the most common response) for both awareness and usage were reported.

Calculating Pearson correlation coefficient between two variables requires a linear relationship to be exist between these variables [23]. A scatter plot is considered a good way to check for linearity [23].

Hence, in order to check whether there is a linear relationship between the awareness and usage, a scatterplot was used. Correlation analysis was conducted to investigate the degree to which the two variables (i.e. awareness and usage) are closely related [24].

A correlation coefficient (r) was calculated. Typically, a correlation coefficient value lies between -1 and +1 [25]. A value of 0 implies no correlation. A positive value indicates a positive correlation (i.e. as one variable increases, the other variable increases too). A negative value indicates a negative correlation (i.e. as one variable increases, the other variable decreases). Only correlations that are significant (i.e. with p-value less than the significance level (α=0.05)) should be considered [25]. Hence, if the test results with p-value less than 0.05, then this correlation is considered statistically significant at the population level.

For further investigation about how the awareness is associated with usage, simple linear regression test was conducted. The awareness was assumed the independent variable (predictor) and the usage was assumed the dependent variable (response).

The p-value for the for the coefficient of the independent variable (i.e. awareness) is used to assess whether changes in the independent variable are really associated with changes in the dependent variable [26]. A (p-value <= 0.05) for the coefficient of the independent variable (i.e. awareness) means that this relationship is statistically significant. In other words, changes in the independent variable (i.e. awareness) value are related to changes in the dependent variable (i.e. usage) value at the population level.

In order to assess the trustworthy of the regression results [27] and the suitability of this regression model to the dataset, residuals analysis was also conducted.

Moreover, as Moodle features are either activities (14) or resources (7), the overall awareness and the overall usage for the activities and resources were calculated by finding the mean of means for both categories.

Furthermore, the 21 Moodle features implemented in HU installation were classified based on their main functionalities into seven categories. This 7-categories classification is based on [19] [14] [20] with some modifications. The “Content delivery” category includes the simple features that are used to communicate content to students such as file, book and page. “Standardized Content Packages” category includes the features that the instructor can leverage to create multimedia content according to an agreed standard such as SCORM and IMS packages. “External Resources” that provide access to resources outside the Moodle instance like URL and external tool. “Communication” features that allow discussions synchronously such as chat and asynchronously such as forum. “Collaboration features” that enable participants to work together such as database, glossary, wiki and workshop. “Direct Assessment” features that allow instructors to measure how well students have mastered course learning outcomes [28] such as Assignment and Quiz features. “Indirect Assessment” features that are used to gather data from students usually to evaluate their understanding or to evaluate the course based on students’ perceptions and satisfaction such as choice, feedback and survey features. Then, the overall awareness and usage for each category were calculated by finding the mean of means.

IV. Results

A total number of 140 responses were returned. The respondents were from different faculties, namely, Science (22.86%), Information Technology (18.57%), Engineering (18.57%), Arts (and languages center) (10.00%), Economics and Administrative Sciences (7.14%), Nursing (6.43%), Applied Health Sciences (5.71%), Medicine (3.57%), Educational Sciences (2.86%), Natural Resources and

Environment (2.14%), Physical Education and Sport Science (0.71%), Pharmaceutical Sciences (0.71%) and Arid Lands (0.71%). Fig. 1 shows the distribution of participants among different faculties in HU.

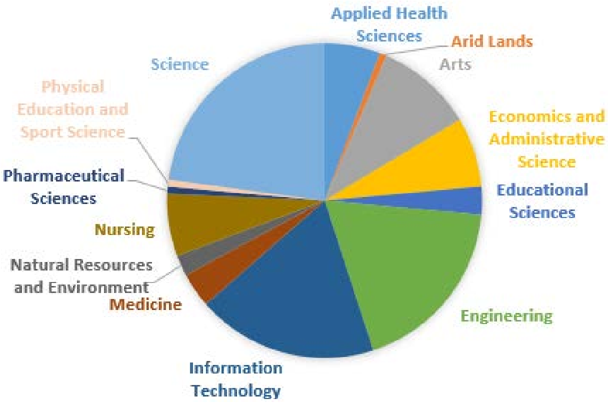


Fig. 1. Respondents Distribution among Faculties.

Reliability tests resulted with a Cronbach's alpha value of 0.94 for the group of items that assess the awareness of Moodle features. This high value implies that the items in this group are internally consistent and actually measures the awareness. Another Cronbach's alpha was calculated to assess the internal consistency among the items that assess the usage. This value of 0.93 implies a high consistency and that the items actually measures the usage.

- **RQ1: Which Moodle features have the highest awareness, and which features have the least awareness among HU faculty?**

Features awareness analysis results are displayed in Table I. For each feature, mean, median and mode are displayed. The features in the table were sorted descending according to the mean, which was used to describe the overall awareness. As per the table, HU's instructors exhibited the highest awareness towards the File feature (4.59), followed by Folder (3.85), Assignment (3.62), URL (3.56) and Quiz (3.46). These were the features that had the highest awareness among HU instructors. This result is also supported by the median and mode values as displayed in the table. The least awareness was towards SCORM package (1.75) and IMS content package (1.76). This result is also supported by a median and mode value of 1 for both.

- **RQ2: Which Moodle features are used most frequently, and which features are used least frequently by HU faculty?**

Table II depicts the usage analysis for Moodle features among HU instructors. Mean, median and mode were also calculated for each feature. The overall usage was represented by the mean. Hence, the features in the table were also sorted descending based on mean values. According to the table, the most frequently used feature was File (4.64), followed by Folder (3.61), Assignment (3.18) and URL (3.16). These were the most commonly used features. This result is also advocated by the median and mode values illustrated in the table. The least frequently used features were IMS content package (1.49), SCORM package (1.51), Wiki (1.57), Glossary (1.62), Workshop (1.67), Database (1.79), Survey (1.81), External tool (1.88) and Choice (1.96).

TABLE I. FEATURES AWARENESS

Feature	Mean	SD	Median	Mode
File	4.59	0.88	5	5
Folder	3.85	1.26	4	5
Assignment	3.62	1.27	4	5
URL	3.56	1.38	4	5
Quiz	3.46	1.39	4	5
Chat	3.09	1.33	3	3
Page	2.96	1.32	3	2
Label	2.96	1.45	3	2
Lesson	2.85	1.27	3	4
Forum	2.74	1.36	3	3
Book	2.66	1.18	3	3
Feedback	2.63	1.40	2	2
Survey	2.42	1.31	2	1
Choice	2.35	1.34	2	1
External tool	2.29	1.23	2	1
Database	2.22	1.19	2	1
Glossary	2.11	1.17	2	1
Workshop	2.08	1.16	2	1
Wiki	2.01	1.19	2	1
IMS content package	1.76	1.08	1	1
SCORM package	1.75	1.06	1	1

TABLE II. FEATURES USAGE

Feature	Mean	SD	Median	Mode
File	4.64	0.79	5	5
Folder	3.61	1.30	4	5
Assignment	3.18	1.44	3	5
URL	3.16	1.40	3	5
Quiz	2.73	1.44	2.5	1
Label	2.69	1.49	2	1
Page	2.66	1.30	2	2
Lesson	2.61	1.41	2.5	1
Chat	2.45	1.24	2	2
Book	2.32	1.23	2	1
Forum	2.14	1.25	2	1
Feedback	2.07	1.27	2	1
Choice	1.96	1.15	2	1
External tool	1.88	1.08	2	1
Survey	1.81	1.08	1	1
Database	1.79	1.02	1	1
Workshop	1.67	0.92	1	1
Glossary	1.62	0.91	1	1
Wiki	1.57	0.89	1	1
SCORM package	1.51	0.93	1	1
IMS content package	1.49	0.90	1	1

- **RQ3: Is there a statistical relationship between faculty's awareness of Moodle features and their usage?**

In order to statistically answer this question, scatterplot, correlation and regression analyses were conducted.

A. Scatterplot

A scatter plot that illustrates the relationship between the overall awareness and the overall usage of Moodle features is depicted in Fig. 2. In this scatterplot, the overall awareness appears on the horizontal axis, and the overall usage appears on the vertical axis. The location of dots on the plot depends on each feature's overall awareness and overall usage.

As per the figure, the overall direction of the relationship implies a positive relationship between the awareness and usage since as the awareness increases the usage also increases. Moreover, since the dots closely resemble a straight line [29] the relationship is obviously a linear relationship. Thus, Pearson's correlation coefficient could be calculated.

B. Correlation Analysis

A correlation analysis was conducted to examine the relationship between the awareness and the usage. The correlation analysis resulted with a Pearson's correlation coefficient value of (r=0.98) and a significance value of (p-value=0.00). A Pearson correlation coefficient value of (r=0.98) indicates a strong positive correlation according to [30] and a very strong positive correlation according to [24] between instructors' awareness of Moodle features and their usage. A significance value of (p=0.00), which is less than the significance level (α=0.05), implies that this correlation is significant; has not come by chance and could be generalized to the entire population [31].

C. Regression Analysis

The conducted correlation analysis stated that there is a strong positive correlation between the instructors' awareness of features and their usage. Furthermore, to investigate how changes in the awareness are associated with changes in usage, simple linear regression test was conducted.

Regression test of the awareness as an independent variable and usage as a dependent variable resulted in a relation with R^2=0.95 (see Table III) and a coefficient value of the awareness (slope) equals 1.06, see Fig. 3.

This regression analysis resulted with a (p-value=0.00) that is less than the significance level (α=0.05), which implies statistical significance (i.e. changes in awareness are associated with changes in the usage at the population level).

To assess the appropriateness of this linear regression model for the data, residuals analysis was conducted. A residual is the difference between the observed value and the predicted value by the model for that observation. The resulted residuals plot is depicted in Fig. 4. As residuals are scattered randomly around the x-axis and are normally distributed, then this linear regression is appropriate for the data.

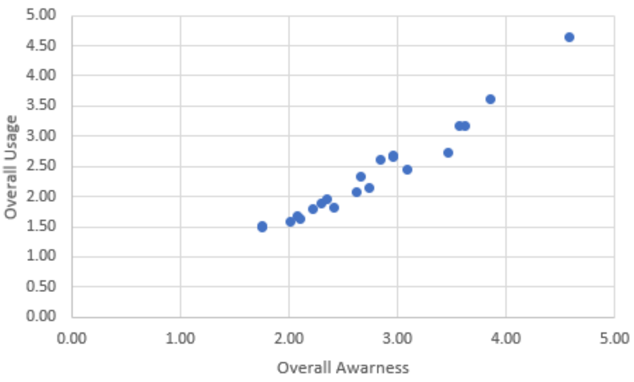


Fig. 2. Overall Awareness-usage Relationship.

TABLE III. REGRESSION STATISTICS - RESULTS

Parameter	Value
Multiple R	0.98
R Square	0.95
Adjusted R Square	0.95
Standard Error	0.18
Observations	21

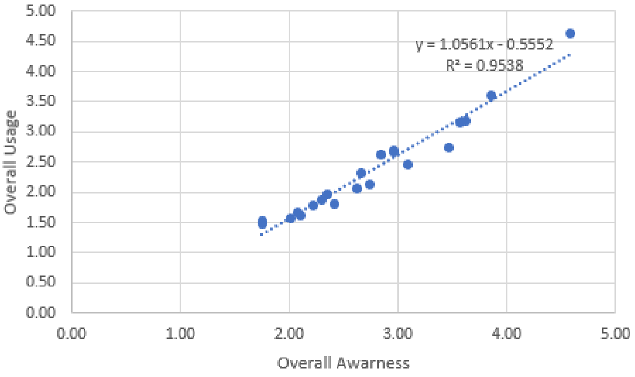


Fig. 3. Awareness-usage Regression Analysis.

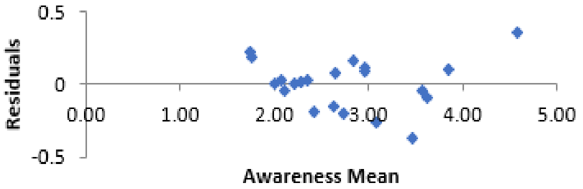


Fig. 4. Awareness mean Residuals Plot.

- **RQ4: Based on (activities/resources) classification of Moodle features, which category has highest/least awareness and which category is most/least frequently used?**

The awareness and usage among activities and resources were separately examined in Table IV. As per the table, HU instructors exhibited considerable higher awareness and usage (3.19, 2.94) towards resources than that for activities (2.55, 2.07).

TABLE IV. AWARENESS AND USAGE OF ACTIVITIES AND RESOURCES

		Awareness	Usage
Assignment	Activities	2.55	2.07
Quiz			
Chat			
Lesson			
Forum			
Feedback			
Survey			
Choice			
External Tool			
Database			
Glossary			
Workshop			
Wiki			
SCORM package			
File	Resources	3.19	2.94
Folder			
URL			
Page			
Label			
Book			
IMS content package			

- **RQ5: Based on the adopted functional classification of features, which categories have the highest/least awareness and which categories are the most/least frequently used?**

For more convenience, awareness and usage were examined for Moodle features based on their functionalities. Table V displays the functional categories sorted by the category's awareness. HU instructors exhibited highest awareness towards "Direct Assessment" (3.54) and "Content Delivery" (3.4) features and least awareness towards "Standardized Content Package" (1.75).

TABLE V. AWARENESS OF FEATURES - GROUPED BY FUNCTIONALITY

Category	Features	Mean	Category awareness
Direct Assessment	Assignment	3.62	3.54
	Quiz	3.46	
Content Delivery	File	4.59	3.40
	Page	2.96	
	Book	2.66	
External Resources	URL	3.56	2.93
	External Tool	2.29	
Communication	Chat	3.09	2.92
	Forum	2.74	
Indirect Assessment	Feedback	2.63	2.47
	Survey	2.42	
	Choice	2.35	
Collaboration	Database	2.22	2.11
	Glossary	2.11	
	Workshop	2.08	
	Wiki" activity	2.01	
Standardized Content Packages	IMS content package	1.76	1.75
	SCORM package	1.75	

Table VI displays the functional categories sorted by the category's usage. HU instructors exhibited highest usage towards "Content Delivery" (3.21) then towards "Direct Assessment" (2.95) features and least usage towards "Standardized Content Package" (1.5), "Collaboration" (1.66) and "Indirect assessment" (1.95) features.

TABLE VI. USAGE OF FEATURES - GROUPED BY FUNCTIONALITY

Category	Feature	Mean	Category usage
Content Delivery	File	4.64	3.21
	Page	2.66	
	Book	2.32	
Direct Assessment	Assignment	3.18	2.95
	Quiz	2.73	
External Resources	URL	3.16	2.52
	External tool	1.88	
Communication	Chat	2.45	2.29
	Forum	2.14	
Indirect Assessment	Feedback	2.07	1.95
	Choice	1.96	
	Survey	1.81	
Collaboration	Database	1.79	1.66
	Workshop	1.67	
	Glossary	1.62	
	Wiki	1.57	
Standardized Content Packages	SCORM package	1.51	1.50
	IMS content package	1.49	

V. DISCUSSION AND LIMITATIONS

The findings of the study state that the most commonly used Moodle features at Hashemite University are File, Folder, Assignment and URL while the least commonly used features are IMS content package, SCORM package, Wiki, Glossary, Workshop, Database, Survey, External tool and Choice. Indeed, cautious comparison with similar works in the literature should be carried out as some of them investigated a subset of the standard Moodle features [4] [15], some investigated tools and features that are available in Moodle other than activities and resources [11] [13] [16], some investigated general features, criteria and tasks that could be supported by a LMS [12] [14]. Furthermore, some works targeted different population other than HEI instructors like school instructors [15], some targeted students rather than instructors [13].

Regarding awareness, it was found that HU instructors are highly aware of File, Folder, Assignment, URL and Quiz features while they are least aware of SCORM package and IMS content package.

Moreover, the study demonstrates statistically that the awareness of Moodle features and their usage are highly correlated and that changes in the awareness are associated with changes in usage. However, neither the correlation nor the regression implies causation [32]. In another words, statistically, it cannot be concluded that the awareness of features causes its usage, or the usage of features causes its awareness. Though, it is believed that many Moodle features are not being used that much by HU instructors due to their unawareness of them, nobody can say that awareness causes usage as other factors may exist too.

Furthermore, the study discloses that the usage and awareness of Moodle “resources” are higher than “activities”. Based on their functionalities, the study also revealed that “Content Delivery” and “Direct Assessment” features are most widely used and well-known amongst HU instructors, whilst the “Standardized Content Packages” features and “Collaboration” are the least.

Generally, the results are largely consistent with the researchers’ expectations.

The study might be limited by the sample size. The survey was distributed via email among the majority of HU faculty who are currently on their work. To increase number of responses, some instructors were reminded later by email and other communication facilities. A total number of 140 responses were returned. The sample represents around 22% of the HU faculty staff who are currently on their work and around 19% of the overall faculty staff.

As the population is limited and in order to control survey distribution process, respondents were asked for their contact information in the survey. This could affect the responses, though a quick look to the responses reveals that most respondents were rational.

One more issue, the survey was opened on Feb 10, 2020 (before COVID-19 has appeared in Jordan) and closed on April 7, 2020 after around 3 weeks from being moved gradually to distance learning due to COVID-19 quarantine. This may not affect the results of this study significantly as the survey was closed in the early weeks. However, instructors’ behavior towards some Moodle features may slightly differ after the end of this quarantine. This could be an interesting dimension for future work.

VI. Conclusion

In this paper the awareness and usage of 21 Moodle features (activities and resources) among HU instructors have been investigated. The study has highlighted the features that have been well-known to HU instructors, these features are: File, Folder, Assignment, URL and Quiz. On the other hand, HU instructors have expressed the least awareness towards SCORM package and IMS content package features. The study also has revealed that the most frequently used features have been File, Folder, Assignment and URL while the least frequently used features have been: IMS content package, SCORM package, Wiki, Glossary, Workshop, Database, Survey, External tool and Choice. Moreover, this study has demonstrated a significant, positive, strong correlation between instructor’s awareness of Moodle features and their usage. Particularly, changes in the awareness of features are significantly associated with changes in their usage.

The study also has found that Moodle resources have received higher awareness and usage than Moodle activities.

Furthermore, Moodle features have been classified based on their functionalities into seven categories, namely, Content delivery, Communication, Collaboration, External Resources, Direct assessment, Indirect Assessment and Standardized Content Packages features. Among all these functional categories, the study has indicated that “Content Delivery” and “Direct Assessment” features have been most widely used and well-known amongst HU instructors whilst the “standardized content packages” and “Collaboration features” have been the least.

At first place, this study would help responsible parties and Moodle administrators in HEIs decide about the most important features that should be installed in their customized instance of Moodle, and even any LMSs, based on the functional categories. Also, the study would help HU responsible parties in identifying the least commonly used and the least well-known features in purpose of conducting activities that aim at increasing the level of awareness and usage of Moodle features. However, conducting activities that focus on enhancing the awareness solely may help in increasing usage based on the conducted regression, but since causation is not established, this cannot be guaranteed. Further, such activities that aim at enhancing the awareness and usage of Moodle activities may reflect positively on the learning process. If such activities were conducted, another post-activity research could be conducted in the future.

References

[1] "About Moodle," 4 Dec 2018. [Online]. Available: https://docs.moodle.org/38/en/About_Moodle. [Accessed 22 Jan 2020].

[2] R. Jeljali, L. Al Naji and K. Khazam, "A Comparison Between Moodle, Facebook, and Paper-based Assessment Tools: Students’ Perception of Preference and Effect on Performance," International Journal of Emerging Technologies in Learning, vol. 13, no. 5, pp. 86-99, 2018.

[3] "Moodle Releases," 13 Jan 2020. [Online]. Available: https://docs.moodle.org/dev/Releases. [Accessed 22 Jan 2020].

[4] D. KC, "Evaluation of Moodle Features at Kajaani University of Applied Sciences – Case Study," in Procedia Computer Science 116, 2017.

[5] "Moodle Activities," 28 Dec 2019. [Online]. Available: https://docs.moodle.org/38/en/Activities. [Accessed 22 Jan 2020].

[6] "Moodle Resources," 28 Dec 2019. [Online]. Available: https://docs.moodle.org/38/en/Resources. [Accessed 22 Jan 2020].

[7] E. Fayyoumi, S. Idwan, K. AL-Sarayreh and R. Obeidallah, "E-learning: challenges and ambitions at Hashemite University," International Journal of Innovation and Learning, vol. 17, no. 4, pp. 470-485, 2015.

[8] A. Al-Khasawneh and R. Obeidallah, "E-Learning in the Hashemite University: Success Factors for Implementation in Jordan," in Advanced Online Education and Training Technologies, IGI Global, 2019, pp. 135-145.

[9] H. Mahdizadeh, H. Biemans and M. Mulder, "Determining Factors of the Use of E-learning Environments by University Teachers," Computers & Education, vol. 51, no. 1, pp. 142-154, 2008.

[10] R. G. Jurado, T. Pettersson, A. R. Gomez and M. Scheja, "Classification of the Features in Learning Management Systems," in XVII Scientific Convention on Engineering and Architecture, Havana, 2014.

[11] V. G. Avila Jasmine , N. G. Hembra, J. M. Mueco and F. G. Zamora, "Moodle and Facebook as A Tool for Delivering Instruction and Attainment of Learning," LPU Laguna Journal of Arts and Sciences, vol. 2, no. 1, pp. 227-250, 2015.

[12] T. M. Hamtini and H. N. Fakhouri, "Evaluation of open-source e-Learning Platforms based on the Qualitative Weight and Sum Approach and Analytic Hierarchy Process," in proceedings of the 10th international conference on education ang information systems, technologies and applications, orlando, florida,USA, 2012.

[13] L. Hasan, "Investigating Students' Perceptions of Moodle LMS In Terms of Its Features and Usability," International Arab Journal of e-Technology, vol. 5, no. 3, pp. 110-122, 2019.

[14] C. Costa, H. Alvelos and L. Teixeira, "The use of Moodle e-learning platform: a study in a Portuguese University," in Procedia Technology, 2012.

[15] A. Badia, D. Martín and M. Gómez, "Teachers' Perceptions of the Use of Moodle Activities and Their Learning Impact in Secondary Education," Technology, Knowledge and Learning, vol. 24, no. 3, p. 483–499, 2019.

[16] F. Ali, A. A. E Al-Mallah and M. Al-Sehlawi, "Exploratory study on Moodle Usage and Satisfaction Level for the Academic Faculty of CHS," in Medical Education in the GCC Countries Conference :Needs, Challenges & Opportunities, 2013.

[17] V. Nedeva, G. Shivacheva, H. Zheleva and V. Atanasova, "Improving Cooperative Learning Activities by New Moodle Features," Applied Researches in Technics, Technologies and Education, vol. 3, no. 3, pp. 224-233, 2015.

[18] R. Purbojo, "Role of the University Lecturer in an Online Learning Environment: An Analysis of Moodle Features Utilized in a Blended Learning Strategy," in In: Persichitte K., Suparman A., Spector M. (eds) Educational Technology to Improve Quality and Access on a Global Scale. Educational Communications and Technology: Issues and Innovations., 2017.

[19] M. Piotrowski, "What is an E-Learning Platform?," in Learning Management System Technologies and Software Solutions for Online Teaching: Tools and Applications, IGI global, 2010, pp. 20-36.

[20] UMASS, "Activity Types in Moodle," [Online]. Available: https://www.umass.edu/it/support/moodle/activity-types-moodle. [Accessed 23 Jan 2020].

[21] M. Tavakol and R. Dennick, "Making sense of Cronbach's alpha," International Journal of Medical Education, pp. 53-55, 2011.

[22] T. S. Keith, "The Use of Cronbach's AlphaWhen Developing and Reporting Research Instruments in Science Education," Research in Science Education, no. 48, p. 1273–1296, 2018.

[23] J. Chee, "Pearson's Product Moment Correlation: Sample Analysis," 2015.

[24] S. Senthilnathan, "Usefulness of Correlation Analysis," SSRN, 2019.

[25] F. Gagné, "Descriptive Statistics and Analysis in Biochemical Ecotoxicology," in Biochemical Ecotoxicology, 2014.

[26] H. J. Seltman, "Simple Linear Regression," in Experimental Design and, 2018, pp. 213-240.

[27] J. Frost, "Check Your Residual Plots to Ensure Trustworthy Regression Results!," [Online]. [Accessed June 2020].

[28] M. J. Allen, "Strategies for Direct and Indirect Assessment of Student Learning," 2008.

[29] D. S. MOORE, W. I. NOTZ and M. A. FLIGNER, "Scatterplots and Correlation," in The Basic Practice, 2013, pp. 97-123.

[30] N. Gogtay and U. Thatte, "Principles of Correlation Analysis," Journal of The Association of Physicians of India, vol. 65, pp. 78-81, 2017.

[31] D. B. F. Filho, R. Paranhos, E. C. da Rocha, M. Batista, J. A. da Silva Jr., M. L. W. D. Santos and J. G. Marino, "When is statistical significance not significant?," Brazilian political science review, vol. 7, no. 1, pp. 31-55, 2013.

[32] J. Frost, "Causation versus Correlation in Statistics," [Online]. [Accessed June 2020].

The Role of ICT Projects in Enterprises: Investments, Benefits and Evaluation

Khaled H. Alyoubi[1]
Information Systems Department, Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

Abstract*—The enterprise's dependency on Information and Communication Technologies (ICT) resources is essential, which cover their several business and operational activities. Enhancing operational capabilities, advancing working environment, and improving employees skills are major benefits provided by modern ICT resources. The real pressure is on organizations to upgrade the ICT infrastructure with latest development to compete in the market. This research investigates the role of ICT projects in an organization from investment, benefits and evaluation perspectives. Based on the literature review, the conceptual framework proposed to understand the relationship between ICT project's investments, benefits, and evaluation. The main purpose of this study is to investigate the approach of enterprises toward ICT investments. Moreover, to understand the type of ICT evaluation strategies that practicing by organizations. Therefore, the proposed framework is applied and validated through multiple case studies to confirm the list of variables collected from literature review. The conducted investigation will help to certify the findings of literature review through selected case studies. The analysis of responses presented in different format to understand the current role and status of ICT projects; investment, benefits and evaluation performed in different organizations. The outcome of this study will addresses substantial factors and offers references for the organization to build their ICT investment and evaluation model. The type of ICT investment, benefits, and measurement models extracted in this research can act as a reference for the organization to develop their own ICT investments policies.

Keywords—ICT projects; ICT investment; ICT benefits; ICT evaluation strategies

I. Introduction

Information and Communication Technologies (ICT) resources are covering major part in executing organizational operational activities and in increasing business growth. According to the facts and figures the enterprises are investing large amount on implementing and enhancing ICT applications to take competitive advantage. The Gartner reported in January 2020 that overall global ICT spending will reach around $3.9 Trillion in the end of this year [1]. This highlights the dependency, importance, and impact of ICT investments over enterprises. The ICT investment is reasonably increasing every year due to latest technological development and new business strategies. It is ultimately improving the business growth and reducing the human efforts, through automating the business processes [2].

The main purpose behind big ICT investments are many as discussed in previous studies. Supporting business architecture [3], increase productivity [4], creating electronic services [5], digital marketing [6], enterprise and decision support systems [7] are some of the factors that require more investments from the organizations. The ICT investments has been categorized in different sectors such as devices, data center, enterprise systems as illustrated in Fig. 1. The list of sectors defined by Gartner, by mentioning the amount of investment for each category. It can be seen from the figure that out of five categories the maximum amount invested on "communication services", whereas the least amount is investing on "data services".

The ICT investments creates numerous benefits to the organizations. The list of benefits identified in literature has divided in two different types: (i) tangible benefits and (ii) intangible benefits [8]. Tangible benefits are those, which can be measured using quantitative strategies such as return on investment, cost and revenue ratio, return on assets and by using other financial formulae [9]. On the other side, intangible benefits are related to those factors, which cannot be measured using direct equation. Indeed, these factors are of qualitative nature that requires multiple questions to be answered during evaluation process. In one of the report, intangible benefits are also known as value on investment, illustrates as the number of values ICT investment generated in the organization [10], [11]. The list of benefits discussed in previous researches are known as data quality and decision making process [12], and communication skills [13].

The evaluation of ICT investment is a complex but essential procedure regularly applied in the organization. It can provide the status of ICT resources in the organization before and after implementation. There are different evaluation strategies proposed in earlier researches. The traditional approach is known as financial approach, which provide the direct benefits generated from the particular ICT application [14]. Other type of evaluation methodologies discussed by [15] are formative and summative approach. Recently, the value on investment is commonly used and applied by different organization. The idea behind this kind of evaluation is to measure the data quality generated by any ICT investment [16]. For measuring the benefits from value's perspective has been modified and applied various time [11], [17] as this provide the new way for measuring the ICT investments. Still the measurement process of ICT benefits evaluation is under progress due to its vast implications on organization can be categorized as financial and non-financial benefits.

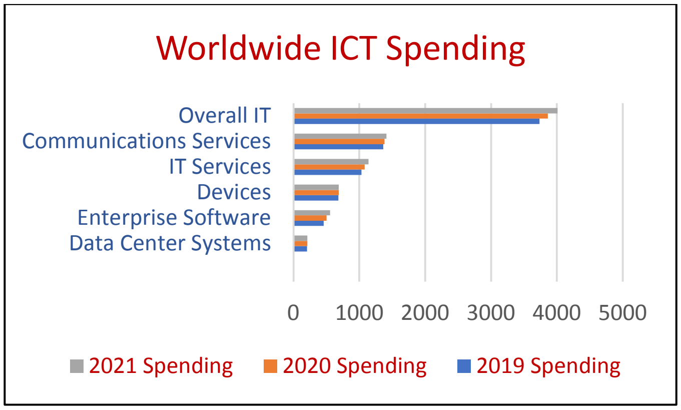


Fig. 1. Worldwide IT Spending Forecast (in Billions of U.S. Dollars) [1].

The main problem undertaken in this research is “What are the major ICT investments organizations are doing and in which sector? and how they evaluate those investment”? In addition, the study examined the most and least invested ICT sector in the organization. As companies are investing large amount on ICT but still there is complexity involves in evaluation approaches. A strategy that can evaluate the both; return and value generated from those investments [18]. Likewise, the problem sub-categorized in this study to observe the most common approaches organization are employing for measuring the benefits created by the particular ICT investment. The other supported questions that will be explored in case studies are risk involvement and prefer time of measurement such as pre or post.

The remaining paper has structured as follows. The next section explained the investment and evaluation status in previous studies. It will help to extract different questions to be asked in case studies during investigation. Section III discussed about the methodology applied in this study. In addition, the proposed framework is explained in that section. Section IV will talk about the implementation steps applied. It will also provide the detailed analysis on number of questions asked from different case studies. Finally, the last section will conclude this paper followed by proposing the future work.

II. ICT Projects: Investment and Evaluation

The role of ICT in an organization is important to run different information systems. Apart from big investments on information technology (IT) infrastructure, utilization of those resources is the main concern of the organization [19]. To know the status of ICT resources organization used different type of evaluation strategies. Traditionally, financial approaches such return on investment and return on assets are commonly known for measuring the return from any investment [20]. Currently, the progression in evaluation techniques suggested different other methodologies such as non-financial approach [21], portfolio management approach [22], economic approach [23], balanced scorecard [24], and others.

The development in ICT evaluation strategies is progressed due to high integration of ICT resources in business communities. The increase of ICT resources in organizations created positive impact over developing new ICT evaluation strategies. The use of ICT applications can be realized by reviewing the previous work presented in different countries. For example, ICT involvement in Spanish class room system for learning activities [25], ICT household resources evaluation in Polish [26], and assessment of ICT tools in Switzerland [27]. In addition, different researches presented to measure the ICT involvement in communities such as relationship between ICT and environment quality [28], ICT and health industry [29], ICT and productivity [30]–[32], ICT and teaching environment [33]–[35], last but not the least ICT dependency in building electronic services [5], [36].

The above paragraph highlights the major development in organizations are highly depending on ICT resources. Therefore, this study investigated the organization’s point of view on measuring those investment. The summary of common evaluation strategies presented in Table I. The traditional financial approach does not provide the comprehensive report as it will just calculate the return from particular investment [9]. On the other side value on investment (VOI) is new addition in evaluation strategies, which can describe what type of values ICT resources provided to the industry. There is a progressive development in VOI based strategies presented in previous work such as moving from ROI to VOI [10], discussed the purpose and objectives of measuring non-financial factors. Those factors can be helpful for the organization that are public and non-profitable, where they do not consider more about return on investment rather their major concern is how much value generated by the ICT resources.

TABLE I. ICT Project Evaluation Approaches

Approach	Major Category	Pre or Post Evaluation	Methodology
Return on Investment [37]	Financial Approach	Post	By calculating the ratio between the investment and generated returns.
Value on Investment [10]	Non-Financial Approach	Post	Generally, it uses different dimensions for measurement such as; organizational impact, community impact and others.
Giga Information Framework [38]	Portfolio Management Approach	Pre	It’s a project based approach used by decomposing the objectives of the project and predict the potential impact.
VMM [39]	Multi Dimension Approach	Pre	Evaluating the potential performance of the project by adding multiple dimensions and phases.
IDA-VOI Approach [11]	Multi Dimension Approach	Pre and Post	Depends on multiple domain and dimensions like cost-benefits analysis and value analysis.
Information Economics [23]	Integrated Approach	Pre	Integration of multiple approached used in this techniques such as ROI, NPV. In addition also combining dimensions for measuring value on investment.

The ICT investment are generating numerous benefits need to be evaluated using proper technique. According to [40] the ICT values are directly related with the main objective of ICT investment. For example, if the organization implemented new strategy for data integration and backup system, the benefit from this project will be known as "informational benefits". Whereas, these benefits cannot be assessed using ROI formula. It will require different type of questions to be examined properly. Meanwhile, there are different issues highlighted by the scholars while measuring these type of benefits. For example, due to their implications, there is no direct formula for evaluating the list of informational benefits [41]. Therefore, this research took a small step towards proposing a framework for measuring different types of investment as discussed in following section.

III. Methodology

The research methodology is a procedure to develop detailed and step-wise approach for attaining the answers of problems discussed in this study [42]. Therefore, the research framework as shown in Fig. 2 illustrating the overall approach taken in this study to discover the solution for the research questions. As evident from the figure the research considered the case study strategy in this study for data gathering. This type of approach is commonly used, which provides the latest experiences of the major stakeholders in the case studies [43]. Moreover, the case study is appropriate in this kind of research, where detailed analysis and company's procedures are investigated on behalf of research question [44]. The case study approach is commonly used in previous researches, specifically in the field of ICT investment and evaluation [45]–[47].

Therefore, based on the findings from literature review the list of questions have been extracted. The questions asked from each case study for inquiring about their approach on ICT investment, benefits, and assessment procedures. According to the Fig. 1, the first two phases has covered with the help of literature review. The questionnaire developed with the help of previous work, therefore, each question is properly referenced to the original work as mentioned in the next section. The purpose of referencing the questions from literature review, is to validate the construct developed in this study, which has been presented earlier from different scholars. Another main idea was to integrate multiple theories presented by different scholars in one framework.

The next phase in the model is the selection of case studies and inquire about basic idea of ICT project investment; known as Phase-3. The first question investigates in this phase is to understand the objectives behind each investment. To identify the purpose of investment is much associated with getting the idea of potential impact of ICT resources. Secondly, stakeholder's analysis will be performed to estimate the list of beneficiaries. Both of these questions are asked prior to major investigation as this phase will provide the basic information about the ICT applications and investment. The idea of inquiring about ICT project's objectives and stakeholder's analysis is presented and validated in previous researches [8], [48]. ICT applications facilitates different kind of stakeholders, this step will help in investigating the further research questions. The purpose of investment will also identify the number of services it will provide and who will receive those benefits. This phase will lead us towards the final phase of this model that is Phase-4.

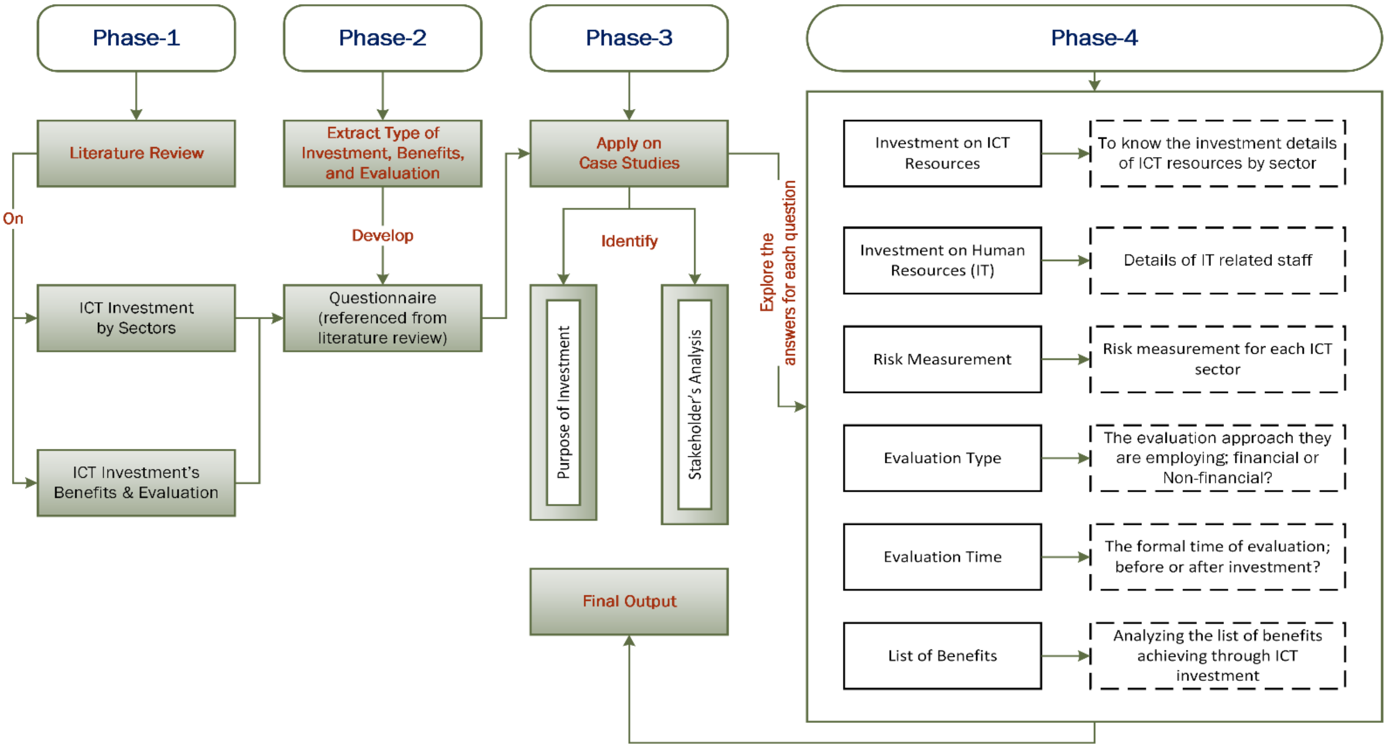

Phase-1
Phase-2
Phase-3
Phase-4
Literature Review
On
ICT Investment by Sectors
ICT Investment's Benefits & Evaluation
Extract Type of Investment, Benefits, and Evaluation
Develop
Questionnaire (referenced from literature review)
Apply on Case Studies
Identify
Purpose of Investment
Stakeholder's Analysis
Final Output
Explore the answers for each question
Investment on ICT Resources
To know the investment details of ICT resources by sector
Investment on Human Resources (IT)
Details of IT related staff
Risk Measurement
Risk measurement for each ICT sector
Evaluation Type
The evaluation approach they are employing; financial or Non-financial?
Evaluation Time
The formal time of evaluation; before or after investment?
List of Benefits
Analyzing the list of benefits achieving through ICT investment

Fig. 2. A Framework for Analyzing the Role of ICT Projects in Enterprise.

Finally, the last step in the framework is about to investigate the main research problem, described in this study. There are number of questions asked as extracted form literature review. The first two questions are related to gather the information about the type of investment. It will further provide the details, whether the amount invested for implementing ICT resources or hiring IT staff. The technical resources can be data center or ICT infrastructure, and IT human resources such as technical staff or IT consultants. Furthermore, the third question is about to assess risk involvement in those investment, and what is the level of risk. The next two queries are related with evaluation type and time. Some organizations are only interested in financial return, while others can consider qualitative assessment tool as well. Evaluation before project initiation or after implementation is another research question asked in this study. The last question illustrating in the framework is about number of benefits achieved from the ICT investment. For more clear understanding, the questions related to assessment approaches were asked to evaluate the latest ICT project implemented in the case studies. From the implementation steps it can be understand that the main reason and purpose of this study in analyzing the ICT project impacts and benefit on selected case studies.

IV. Implementation, Results and Discussion

As discussed in the previous section, the proposed framework is divided in multiple stages. The first two phases was already executed by reviewing vast literature review. During the review process, there are multiple finding discovered as follows. The ICT investment are normally divided into two categories; (i) technical investments and (ii) human resource (IT) investments [49]. The findings suggests that both types of investments can be evaluated using same strategy. As the evaluation process start by investigating number of objectives and list of beneficiaries. The identification of both can help us to apply the pre or post evaluation smoothly. The research identify the motivation behind each ICT investment executed in case studies. The implementation steps mentioned in Fig. 2 applied in sequence on each study. The generated results from each step are discussed in following sub-sections.

A. Case Studies - Overview

Data collected from literature review further applied, validated, and refined through case studies. As this research used qualitative approach, in which single source of data cannot provide the detailed analysis [44]. Furthermore, as described by [50] that validation of construct using single case study does not provide sufficient results if responses collected from same organization. However, the idea of selecting more than one case study is encouraged several time due to repeating the same question and justify the results several time by asking from multiple organization [51], [52]. Therefore, the framework validation applied on five case studies located in Pakistan.

Based on the data extracted from literature review, the structured questions was designed and sent to the organizations. The same questionnaires were sent to each case study for clear understanding and analysis on collected responses. The collected sample size (65) was feasible as required sample size in case study research [53]. Table II explains the details of each case study, where due to privacy issues, the company's names were kept confidential. The table is illustrating the characteristics of collected data from each organization.

TABLE II. Selected Case Studies Characteristics

Organization	Employee ≤ 3 Years' of IT Experience	Employee > 3 Years' of IT Experience	Total Participants
Educational Institute	4	16	20
Bank	2	8	10
Hotel	7	8	15
Restaurant	5	7	12
Manufacturing Industry	3	5	8

Table II representing the total number of participants selected in this case study. As this research does not sent the questionnaire to anyone, but the participant's selection were based on their involvement with ICT project investment and evaluation. Therefore, the ICT users were not part of this study, but only middle and higher management employees were asked to participate in the research. The data showing in above table highlighting the average experience of the respondent is higher than three years. It indicates the validity of the responses is reflecting through their experience. Moreover, the maximum participants belong to "educational institute", also it was the biggest organization in this survey.

B. ICT Projects – Sector Wise Investment

The first thing that examined by the case studies is to understand their sector wise ICT investment. Basically, this question is referring to the published report by Gartner, where they provided the worldwide ICT investment in different sectors [1]. Originally there are five categories, but for making it clear we converted into three main categories as shown in Table III. The table is representing the yearly investment in all categories. The data suggests that maximum investment are for the enhancement of ICT infrastructure. The collected data is validated through Gartner report also that the communication infrastructure is the largest invested sector as shown in Fig. 1. According to the collected data, out of the three categories mentioned in the table, "software" is the lowest investment sector. It can have different reasons: (i) software is one time installment, and most of the time they run for many year with some updates, (ii) some organizations outsource their main software, and (iii) software consist of few requirements than ICT infrastructure and electronic services. Overall, findings suggests the three major categories are very common in the organizations, where they are investing enough amount.

C. Investment on Human Resources (IT)

The next investigation performed using a similar question but for investment on human resources. Scholars have discussed this type of investments in previous researches, where they suggested that IT consultant, IT staff, and user training are part of ICT project's implementation [18], [54]. Therefore, this question is referred to those researches, and we asked the status of investment in case studies. The purpose is to estimate overall status of ICT project's investment. The responses were collected in the form of qualitative approach, where three options were provided for each category as mentioned in Table IV. As showing in the table, IT staff are playing major role in the organization and it seems the hiring is not happened frequently. It shows that the experience is important, and organizations are willing to be consistent with the experienced employees and do not want to replace them. The table is representing that consultants services are commonly taken in the case studies for ICT projects. Commonly, they are responsible for the whole project starting from the idea till the project implementation. Some organizations also hire them for post evaluation and auditing services. In addition, some enterprises are interested to conduct training workshop often. The table suggested that most of them are hiring trainers to train their employees for new systems.

TABLE III. ICT PROJECTS – SECTOR WISE INVESTMENT (PERCENTAGE)

Organization	Software	ICT Infrastructure	Electronic Services
Educational Institute	30%	40%	30%
Bank	20%	30%	50%
Hotel	20%	30%	50%
Restaurant	20%	40%	40%
Manufacturing Industry	30%	60%	10%

TABLE IV. INVESTMENT ON HUMAN RESOURCES (IT)

Organization	IT Staff	Consultancy	Trainer
Educational Institute	Frequently	Sometime	Frequently
Bank	Sometime	Rarely	Frequently
Hotel	Sometime	Frequently	Rarely
Restaurant	Rarely	Frequently	Sometime
Manufacturing Industry	Sometime	Frequently	Frequently

D. *ICT Projects – Risk Measurement*

The risk involves in every project as with ICT project. Most of the time risk assessment are measured using percentage or by using values between "low" and "high", whereas based on the situation the values of risk are modified. The chances of risks are telling also the size of damage after occurring of any bad event. There are many researches that have discussed about risk management in ICT projects [9], [55]. Accordingly, in this study we inquire about organization's point of view about risk involvement under each ICT sector and the chances of damages as shown in Table V. To make it clear the study asked about the three categorized ICT sector. Based on the collected data, it can be clearly seen that "electronic services" are on high risk in all organizations except "Manufacturing Industry". The "electronic services" are directly connected with the internet and may have vulnerability as dealing with customers using online environment. Hacker's attack on online services is the challenging threat for the organization. Whereas the "manufacturing industry" has provided few online services, therefore, they have low risk involvement there, but "very high" for "software" and "ICT infrastructure" as there major work related with these two categories. As per the analysis, low risk means, the organization has backup for those resources or not fully dependent on it. However, the very high, high and low risks refer to the chances as well as the size of damage, if any bad event happened with those resources.

TABLE V. ICT PROJECTS – RISK MEASUREMENT

Organization	Software	ICT Infrastructure	Electronic Services
Educational Institute	Low	High	Very High
Bank	High	High	Very High
Hotel	Low	Very High	Very High
Restaurant	Low	Very High	Very High
Manufacturing Industry	Very High	Very High	Low

E. *ICT Projects – Evaluation Type*

The next phase of this study is to explore the enterprise's approach towards evaluation of ICT project. For this, we have provided the major categories of ICT evaluation approaches to the participants to know what they are practicing commonly. The main reference to this question are [9], [56] where researchers provided different types of categories for assessing the ICT investment. Although, the time of evaluation also play a major role in measuring the impact of ICT resources that has discussed in next section. The selected common ICT project evaluation approaches with the responses are shown in Table VI. The question arose in front of respondents by asking their opinion about what they are practicing. The financial approach used to calculate the return from any investment. Most common approach used by all is "ROI". Same is not the case with non-financial approaches, some of them used "VOI" based approach for some projects but not very common. Normally, VOI based approach are useful to know about how much value the users or customers received from any ICT project. The example of values are discussed by [13] in the form of organizational values generated by ICT resources. Finally, some organization also apply strategy based approach for evaluation, such as portfolio management. Hiring of consultant agency for assessment is the next option asked in this study, which is not very common strategy in the case studies.

TABLE VI. ICT PROJECTS – EVALUATION TYPE

Organization	Financial Model	Non-Financial Model	Strategic Approach	Consultant Agency
Educational Institute	ROI	VOI	No	Based on Project
Bank	ROI	Based on Project	Sometime	Yes
Hotel	ROI	Based on Project	Sometime	No
Restaurant	ROI	Based on Project	No	No
Manufacturing Industry	ROI	VOI	No	Yes
*ROI: Return on Investment *VOI: Value on Investment				

F. ICT Projects – Evaluation Time

This section is directly associated with the previous section, where the time of approach explored. ICT project evaluation strategies are commonly divide in two categories; pre-evaluation and post-evaluation. Mainly, the question is extracted and referenced from the work applied by [57] where they categorized the evaluation strategies into four different types as mentioned in Table VII. According to the collected data, it is evident that all organization are curious about measuring the ICT investment using any approach. Therefore, the following table is showing the acceptance from all cases and highlighting the positive use of evaluation strategy. Most of the respondents agreed that there is a proper mechanism for evaluating the ICT project. Sometime, they do it before implementing it and sometime after. The project based evaluation is also a popular way in case studies, where impact of project evaluated after its execution. It is more direct approach when the implications of ICT resource can be measured based on the project's objectives. Finally, most of the organization practicing ICT evaluation on yearly bases as well. This type of evaluation is not related with particular project, rather it provides the comprehensive analysis on a performance of ICT resources in a complete lunar year.

G. ICT Projects – List of Benefits

Finally, the last questions asked in this study was related to number of benefits ICT project can generate. The question is referred to the study where they identify different type of benefits such as tangible benefits and intangible benefits [58]. Therefore, we select a mixed list of benefits extracted from [58] and asked from the case studies to find out their consent on each, if they agreed or not as illustrated in Table VIII. The purpose was to identify if case studies are supporting the variable and consider those variables as one of the benefits of ICT project or not. Regarding, "information quality" and "assets security" got full confidence from all case studies as the performance of these resources can be enhanced by ICT project. The remaining two variables, "employee skills" and "customer satisfaction" received partial agreement from case studies. Both of these benefits are kind of intangible benefits, which can have conflict of agreement. In previous studies, all of these benefits have been discussed and validated through number of items [40], [59] and confirmed as ICT benefits. The results presented in this section provide the comprehensive analysis on collected data, purely based on participant's responses from case studies. The list of benefits selected in this study were four, but it can be many, which can be identified based on the objectives of ICT project.

TABLE VII. ICT PROJECTS – EVALUATION TIME

Organization	Pre-Evaluation	Post-Evaluation	Project Based	Yearly Bases
Educational Institute	Yes	Sometime	Yes	Sometime
Bank	Yes	Yes	Yes	Yes
Hotel	Yes	Yes	Yes	Yes
Restaurant	Yes	Yes	Yes	Yes
Manufacturing Industry	Yes	Sometime	No	Yes

TABLE VIII. ICT PROJECTS – LIST OF BENEFITS

Organization	Information Quality	Assets Security	Employee Skills	Customer Satisfaction
Educational Institute	√	√	X	√
Bank	√	√	X	√
Hotel	√	√	√	√
Restaurant	√	√	√	X
Manufacturing Industry	√	√	√	X

V. CONCLUSION

ICT project are playing major role in organizational development and business growth. The study presented the idea of examine the current status and practice enterprises are doing for ICT project's investment, evaluation, and benefits. The list of questions are extracted from related work, which further validated through multiple case studies. Therefore, the variable collected from related work has confirmed by selected case studies. The results indicated that organizations are doing most of the investment on building ICT infrastructure. Whereas electronic services considered on high risk zone. Return on investment selected as foremost evaluation strategy for all organization. Apart from using any strategy, but organization are willing to do evaluation before and after implementation to assess the impact and return value from ICT project. Finally, all organizations are agreed on the argument that ICT project creates positive impact on organizations.

In future, the result can be refined by implementing the framework on different case studies selected from multiple regions. This area of research still requires some quantitative measuring framework to understand the visible impact of ICT projects on organizations.

REFERENCES

[1] STAMFORD, "Gartner Says Global IT Spending to Reach $3.9 Trillion in 2020," 2020.

[2] L. Juhaňák, J. Zounek, K. Záleská, O. Bárta, and K. Vlčková, "The relationship between the age at first computer use and students' perceived competence and autonomy in ICT usage: A mediation analysis," Comput. Educ., vol. 141, no. June, 2019.

[3] A. A. Al-ghamdi and F. Saleem, "The Impact of ICT Applications in the Development of Business Architecture of Enterprises," Int. J. Manag. Stud. Res., vol. 4, no. 4, pp. 22–28, 2016.

[4] M. Kante, R. Oboko, and C. Chepken, "An ICT model for increased adoption of farm input information in developing countries: A case in Sikasso, Mali," Inf. Process. Agric., vol. 6, no. 1, pp. 26–46, 2019.

[5] M. K. Anser, Z. Yousaf, M. Usman, and S. Yousaf, "Towards strategic business performance of the hospitality sector: Nexus of ICT, e-marketing and organizational readiness," Sustain., vol. 12, no. 4, pp. 1–17, 2020.

[6] J. M. Millán, S. Lyalkov, A. Burke, A. Millán, and A. van Stel, "'Digital divide' among European entrepreneurs: Which types benefit most from ICT implementation?," J. Bus. Res., no. December 2018, pp. 1–15, 2019.

[7] F. Saleem and A. S. Al-Malaise, "Implementation of Data Mining Approach for Building Automated Decision Support Systems," in Information Society (i-Society), International Conference on (pp. 127-130). IEEE., 2012, pp. 127–130.

[8] F. Saleem, N. Salim, A. G. Fayoumi, and A. Alghamdi, A General Framework for Measuring Information and Communication Technology Investment: Case Study of Kingdom of Saudi Arabia, vol. 322. 2012.

[9] L. Dadayan, “Measuring return on government IT investments,” in Proceedings of the 13th European Conference on Information Technology Evaluation, 2006, no. September, p. 12.

[10] D. Hurley, “Changing the View of ROI to VOI—Value on Investment,” 2001.

[11] IDA-VOI, “IDA Value Of Investment,” 2003.

[12] S. Gregor, M. Martin, W. Fernandez, S. Stern, and M. Vitale, “The transformational dimension in the realization of business value from information technology,” J. Strateg. Inf. Syst., vol. 15, no. 3, pp. 249–270, 2006.

[13] S. Shang and P. B. Seddon, “Assessing and managing the benefits of enterprise systems: the business manager’s perspective,” Inf. Syst. J., vol. 2000, pp. 271–299, 2002.

[14] C. A. Magni, “Internal Average Rate of Return and Aggregate Return on Investment,” Invest. Decis. Log. Valuation, Springer, pp. 555–611, 2020.

[15] M. Hersh, “Evaluation framework for ICT-based learning technologies for disabled people,” Comput. Educ., vol. 78, pp. 30–47, 2014.

[16] B. Heinrich, M. Klier, A. Schiller, and G. Wagner, “Assessing data quality – A probability-based metric for semantic consistency,” Decis. Support Syst., vol. 110, no. April, pp. 95–106, 2018.

[17] E. Jed, J. Wachowicz, and S. Chun, “Social Return on Investment:Exploring Aspects of Value Creation in the Nonprofit Sector,” 2000.

[18] M. Carcary, “ICT Evaluation in the Irish Higher Education Sector,” Electron. J. Inf. Syst. Eval., vol. 12, no. 2, pp. 129–140, 2009.

[19] E. Rodríguez-Crespo and I. Martínez-Zarzoso, “The effect of ICT on trade: Does product complexity matter?,” Telemat. Informatics, vol. 41, no. April, pp. 182–196, 2019.

[20] W. Van Grembergen and S. De Haes, “Measuring and demonstrating the value of IT,” IL 60008 USA, 2005.

[21] Z. Irani, J.-N. Ezingeard, and R. J. Grieve, “Costing the true costs of IT/IS investments in manufacturing: a focus during management decision making,” Logist. Inf. Manag., vol. 11, no. 1, pp. 38–43, 1998.

[22] M. Kozina and D. Popović, “VAL IT Framework and ICT benefits,” in European Conference on Information and Intelligent Systems, 2010, pp. 221–227.

[23] M. Parker and R. Benson, “Information Economics,” Inf. Econ., no. C, pp. 1–15, 1989.

[24] R. Kaplan and D. Norton, “Using the balanced scorecard as a strategic management system,” Harv. Bus. Rev., pp. 75–85, 1996.

[25] M. Arrosagaray, M. González-Peiteado, M. Pino-Juste, and B. Rodríguez-López, “A comparative study of Spanish adult students’ attitudes to ICT in classroom, blended and distance language learning modes,” Comput. Educ., vol. 134, no. October 2018, pp. 31–40, 2019.

[26] A. Karczmarczyk, J. Wątróbski, J. Jankowski, and E. Ziemba, “Comparative study of ICT and SIS measurement in Polish households using a MCDA-based approach,” in Procedia Computer Science, 2019, pp. 2616–2628.

[27] J. C. K. H. Riedel and M. Vodicka, Evaluating the Melodie ICT tool for supporting idea generation, vol. 43, no. 17. IFAC, 2010.

[28] D. Avom, H. Nkengfack, H. K. Fotio, and A. Totouom, “ICT and environmental quality in Sub-Saharan Africa: Effects and transmission channels,” Technol. Forecast. Soc. Change, vol. 155, no. March, p. 120028, 2020.

[29] U. P. Dutta, H. Gupta, and P. P. Sengupta, “ICT and health outcome nexus in 30 selected Asian countries: Fresh evidence from panel data analysis,” Technol. Soc., vol. 59, no. July, p. 101184, 2019.

[30] N. Abramova and N. Grishchenko, “ScienceDirect ScienceDirect ICTs , Labour Labour Productivity Productivity and and Employment : Employment : Sustainability Sustainability in in Industries Industries in in Russia,” Procedia Manuf., vol. 43, pp. 299–305, 2020.

[31] P. Koutroumpis, A. Leiponen, and L. D. W. Thomas, “Small is big in ICT: The impact of R&D on productivity,” Telecomm. Policy, vol. 44, no. 1, 2020.

[32] Q. Zhou, P. Gao, and A. Chimhowu, “ICTs in the transformation of rural enterprises in China: A multi-layer perspective,” Technol. Forecast. Soc. Change, vol. 145, no. June 2018, pp. 12–23, 2019.

[33] A. Habibi, F. D. Yusop, and R. A. Razak, “The dataset for validation of factors affecting pre-service teachers’ use of ICT during teaching practices: Indonesian context,” Data Br., vol. 28, 2020.

[34] J. M. Fernáández Batanero, M. M. Reyes Rebollo, and M. Montenegro Rueda, “Impact of ICT on students with high abilities. Bibliographic review (2008–2018),” Comput. Educ., vol. 137, no. April, pp. 48–58, 2019.

[35] R. Scherer and F. Siddiq, “The relation between students’ socioeconomic status and ICT literacy: Findings from a meta-analysis,” Comput. Educ., vol. 138, no. 0317, pp. 13–32, 2019.

[36] I. O. Adam, “Examining E-Government development effects on corruption in Africa: The mediating effects of ICT development and institutional quality,” Technol. Soc., vol. 61, no. December 2019, p. 101245, 2020.

[37] ROI, “Return On Investment (ROI) - Investopedia,” 2016. [Online]. Available:http://www.investopedia.com/terms/r/returnoninvestment.asp. [Accessed: 02-Nov-2015].

[38] C. Gliedman, “The Foundation of Sound Technology Investment : The Total Economic Impact TM Methodology,” 2003.

[39] VMM, “The Value Measuring Methodology,” 2002.

[40] A. C. G. Maçada and M. M. Beltrame, “IT business value model for information intensive organizations,” BAR-Brazilian …, pp. 44–65, 2012.

[41] R. Mirani and A. Lederer, “An instrument for assessing the organizational benefits of IS projects,” Decis. Sci., vol. 29, no. 4, pp. 803–838, 1998.

[42] M. Saunders, P. Lewis, and A. Thornhill, Research methods for business students, 5th ed. Financial Times Prentice Hall, 2011.

[43] R. Yin, Case study research: Design and methods, 5th ed. SAGE Publications, 2013.

[44] K. F. Punch, Introduction to Social Research: Quantitative and Qualitative Approaches, Third. SAGE Publications, 2013.

[45] J. Torrent-Sellens, P. Ficapal-Cusí, J. Boada-Grau, and A. Vigil-Colet, “Information and communication technology, co-innovation, and perceived productivity in tourism small and medium enterprises: an exploratory analysis,” Curr. Issues Tour., no. ahead-of-print, pp. 1–14, 2015.

[46] J. Manuel and C. Pérez, “Assessing the impact of Information and Communication Technologies on the Portuguese hotel sector : an exploratory analysis with Data Envelopment Analysis,” Tour. Manag. Stud., vol. 11, no. 1, pp. 35–43, 2015.

[47] N. M. Ongaki and F. W. Musa, “A Framework for Evaluating Ict Use in Teacher Education in Kenya,” Int. J. Res., vol. 2, no. 4, pp. 65–95, 2015.

[48] AGIMO, “Demand and Value Assessment Methodology,” Canberra, Australia, 2004.

[49] A. Rai, R. Patnayakuni, and N. Patnayakuni, “Refocusing where and how IT value is realized: An empirical investigation,” Omega, vol. 24, no. 4, pp. 399–412, 1996.

[50] T. Winston, “Introduction to Case Study The Qualitative Report,” vol. 3, no. 2, pp. 1–11, 1997.

[51] R. E. Stake, Multiple Case Study Analysis. Guilford Press, 2013.

[52] K. Eisenhardt, “Better stories and better constructs: The case for rigor and comparative logic,” Acad. Manag. Rev., vol. 16, no. 3, pp. 620–627, 1991.

[53] J. Hair, W. Black, B. Babin, R. Anderson, and R. Tatham, Multivariate data analysis, 7th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2010.

[54] H. W. Chou, H. H. Chang, Y. H. Lin, and S. Bin Chou, “Drivers and effects of post-implementation learning on ERP usage,” Comput. Human Behav., vol. 35, pp. 267–277, 2014.

[55] G. Westerman, “IT Risk Management : From IT Necessity to Strategic Business Value,” Massachusetts Inst. Technol., p. 15, 2006.

[56] E. Khakasa, “Evaluating Information Technology Investments - A Survey of Kenyan Commercial Banks,” Proc. 10th Annu. Conf. IAABD, no. 1992, pp. 473–480, 2009.

[57] A. S. A.-M. Al-Ghamdi and F. Saleem, “General characteristics and common practices for ICT projects: Evaluation perspective,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 1, 2018.

[58] B. Ranti, “A Review of Information Technology Investment Evaluation Methodologies: The Need for Appropriate Evaluation Methods,” Pros. Konf. Nas. Teknol. Inf. Komun. untuk Indones., no. 4, pp. 112–115, 2006.

[59] S. Shang and P. B. Seddon, “A comprehensive framework for classifying the benefits of ERP systems,” AMCIS 2000 Proc., p. 39, 2000.

Serious Games Requirements for Higher-Order Thinking Skills in Science Education

Siti Norliza Awang Noh[1], Nor Azan Mat Zin[2], Hazura Mohamed[3]
Faculty of Information Science and Technology
University Kebangsaan Malaysia, Malaysia
43600 Bangi Selangor, Malaysia

***Abstract*—Education in the 21st century emphasises the mastery of higher-order thinking skills (HOTS) in the pursuit of developing competitive human capital globally. HOTS can be taught through science education. However, science education is considered very challenging leaving students feeling less interested and less motivated. Apart from that, students are found to be weak in mastering their thinking skills based on the decline in students' achievements in the Trend in International Mathematics and Science Study (TIMSS) and the Programme International Students Assessment (PISA) tests. This situation highlights the need to change the approach of teaching and learning science in line with the current technological changes to meet the challenges of globalisation. Previous studies showed that the use of serious games in learning can enhance students' thinking skills. Thus, serious games can be used to develop higher-order thinking skills among students. This paper presented results of preliminary study using interviews, document analysis, and questionnaire survey. Findings have shown that there are several issues and challenges of teaching and learning in implementing HOTS in science education, in addition to game design requirement for science education. The requirements will be used to design a serious game implementing HOTS in science education.**

Keywords—Higher-Order Thinking (HOT) skills; educational games; serious games; interface design; science education

I. Introduction

21st-century learning skills require students to have high-level thinking skills to remain competitive, creative, and innovative. Higher-order thinking skills (HOTS) are defined as the expanded use of the mind to meet new challenges and the process occurs when a person must interpret, analyse or manipulate information because a problem cannot be resolved through the application of previously learned knowledge [1]. Concisely, HOTS is a form of quality thinking that enables one to think more deeply, more productively, and more effectively.

HOTS is a key goal of today's education system worldwide and is also outlined in the country's Education Development Blueprint 2013–2025. An effective method of learning should be implemented to attract students and enhance their HOTS in science. The goal of science education is to help students develop HOTS to meet the challenges of everyday life through enhancing students' cognitive skills such as critical thinking, reasoning, reflective thinking, and science process skills [2].

According to [3] games are a social interaction tool that can stimulate early learning and cognitive development in optimal early childhood education. In addition, play activities bring fun and entertainment as well as encourage individuals to learn something informally. Therefore, the development of computer technology such as serious games can be implemented in the learning process as they offer a strong format for educational environment [4].

Serious games are interactive digital games developed not just for entertainment as the primary purpose but rather as entertainment games that can educate players, also known as educational games. In the field of education, serious games are developed to achieve learning objectives set and are student-oriented. It is in line with the needs of today's students who are born in the age of digital technology known as the Net generation. The Net generation has different skills, interests, and needs, therefore the teaching approach in the education system has to change to meet their requirements. The approaches, methods, techniques, and teaching materials for the current Net generation need to be different from those of the previous generation [5] as the current generation's information processing and thinking are more advanced compared to previous peers. Technology can be used to teach the Net generation.

The biggest challenge of a teacher today is to create an exciting teaching and learning approach to guide students in developing their thinking skills to compete globally. However, the current teaching and learning strategies are exam-oriented that focuses on memorization rather than thinking skills [2]. Lower-order thinking still dominates teaching methods and learning outcomes [6]. Furthermore, the Malaysia Education Blueprint 2013-2025 [7] reported that student encountered difficulty in responding to HOTS exam question in providing complete information along with proof of logic and inability to think critically and logically.

Previous studies discovered that serious games enable the conception of attractive teaching and learning environment, and the development of thinking skills. However, some criteria of the games need to be emphasised so that they are in line with the needs of pedagogy and thinking skills. Balance between the needs of the curriculum and the game structure is important for producing a game that is interesting and does not affect the learning outcomes and functionalities. Apart from that, interesting games often have privileges in terms of accuracy and completeness of learning contents.

Past studies found plentiful games in the market, but most do not meet the criteria of games for education. This is because game developers only emphasise the gaming aspects rather than learning. According to [8] most private companies have developed digital games without regard to students' learning methods and learning strategies. Besides, games developed for education only cover contents that the developers felt necessary without considering the local education syllabus, which are Standard Based Curriculum for Secondary Schools (KSSM) and Standard Based Curriculum for Primary (KSSR). Hence, our study focused on the requirements of serious game design for HOTS in science education to ensure that educational objectives can be achieved.

This paper discusses the results of preliminary study to identify issues, challenges and requirements for implementing HOTS in science education from educator's perception as well as game design requirements from student's perception. Section II presents a literature review which include theory and related works. Section III describes methods used while Section IV presented result and discussions. The paper concludes with a summary of the study findings and future work.

II. Literature Review

A. Higher-Order Thinking Skills in Science Education

According to [3] Thinking skills are the management of mental processes that occur in the mind or the cognitive system including knowledge, observation, and production. Apart from that, thinking is a mind activity to make decisions in solving problems based on existing information and experiences. Thinking aims to find meaning and understanding of something, explore various ideas or creations and make judgements, and then reflect and be metacognitive about the process. In other words, thinking skills are the ability of an individual to master the potential of his/her mind to control, empower, and adapt to a difficult environment.

According to the definition given in the cognitive domain of Bloom's Taxonomy, there are six cognitive levels of which the first two are known as lower-order thinking skills (LOTS) and the subsequent four levels are higher-order thinking skills (HOTS). LOTS is defined as a simple application and routine steps while HOTS challenges students to interpret, analyse, or manipulate information [9] as shown in Fig. 1.

HOTS is defined by the Curriculum Development Division (2013) of the Ministry of Education Malaysia as the ability to apply knowledge, skills, and values in reasoning and reflection to solve problems, make decisions, innovate, and be creative. According to [11] high-level learning means the capacity to go beyond the information given, to develop critical stance, to have metacognitive awareness, and to solve problems. Thinking critically is thinking at a higher level of cognition that increase the probability of a desirable outcome [12].

The thinking skills elements based on Bloom's Taxonomy consist of knowledge, understand, apply, analysis, synthesis and evaluate. Meanwhile, the elements in Bloom's Taxonomy as reviewed by Anderson and Krathwohl consist of remembering, understanding, applying, analysing, evaluating, and creating. According to the Ministry of Education Malaysia (2013), there are four elements at the highest level outlined in HOTS implementation in schools which are applying, analysing, evaluating, and creating as shown in Table I.

These four elements are important aspects that should be implemented in teaching and learning science to encourage thinking skills among students. However, studies have shown that Malaysian students have moderate level of motivation in science [13]. According to the reports by Trends in International Mathematics and Science Study (TIMSS) in 1999, 2003, 2007, 2011, and 2015, the science scores were declining. Moreover, the Programme for International Student Assessment (PISA) reports also showed similar results in 2009, 2012, and 2015. The deterioration of students' achievements in science subjects at international assessment level reflects that Malaysian students have difficulties in answering HOT-based questions. The application of knowledge according to the four highest levels of cognitive taxonomy does not occur due to most lessons in schools did not sufficiently engage students in constructive thinking[6].

Apart from that, the mastery level of critical thinking and problem-solving skills of science stream students was at a moderate level consistent with the current education that tends to be exam-oriented and fact-based [14], which focused on LOT skill such memorization or "Remember". According to [15] teachers lack knowledge and understanding in methods of teaching for thinking. Moreover, teachers do not know the strategies and methods that can be used to incorporate thinking skills in the teaching and learning process [14]. The teachers are also not familiar with the processes and skills due to their lack of confidence to teach HOTS in science [15].

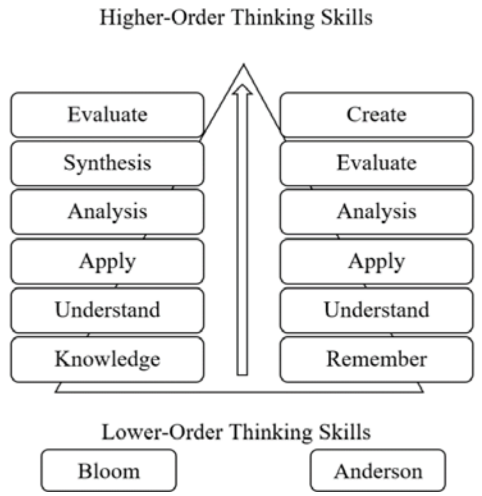


Fig. 1. Reviewed Bloom's Taxonomy (Anderson and Krathwohl, 2001) [10].

TABLE I. HOTS Applied in School

Level of Thinking	Description
Create	Use information to create something new
Evaluate	Critically examine information and make judgements
Analysis	Clarify information and investigate relevance
Apply	Use information in new situation (Almost similar)

Furthermore, TIMSS 2007 reported science lessons are mostly teacher-centric. In addition Malaysian students were reported watching teachers demonstrate experiments and investigation rather than doing it themselves. TIMSS 2011 also reported that 47% of Malaysian students were involved in science investigation in less than half of the lessons.

Most students are unable to solve problem-solving questions that require HOTS [16]. This is because understanding science requires students to be able to apply science concepts and principles in solving problems in different situations and contexts. Therefore, scientific inquiries and problem-solving skills need to be cultivated and practised continuously [17].

Serious games are an appropriate learning tool for implementing HOTS in science education as they are not just games involving story, art, and software, but they are a pedagogical activity for educating as well as constructing thinking related to memory skills and learning [18]. Serious games can stimulate learning and cognitive development leading to student-centred learning. In addition, serious games enable educators to attract students' attention and involve them in educational experiences to achieve learning goals [19].

B. Serious Game

The first concept of serious game was introduced by Clark Abt in 1970, categorised as an entertaining interactive game that allowed players to experience a variety of situations that are impossible to be experienced in the real world [20]. Popular digital games include action games, adventure games, strategy games, and puzzle games.

Serious games are games developed not just for entertainment but also used in business, treatment, medicine, education, health, advertising, and military training. Apart from entertainment, the main purpose of serious games is to train and educate players. One of the most popular serious games is America's Army utilised by the U.S. Army's as a tool in the process of military recruitment and training in the United States. Serious games have also been used in healthcare such as clinical decision-making and patient interaction [21].

Serious games have benefited through interactive learning strategy that provides learning experiences with enjoyment, pleasure, motivation, ego gratification, and emotional change in terms of where, what, and how they learn [22]. Serious games can be considered as a proper educational tool which enhances learning and fulfils students' needs and requirements. According to [21] there is a variety of serious game applications that are applied in computer science, physic, engineering, science, and language learning. Serious games challenge students to think critically in problem-solving.

C. Serious Game for Education

Digital games in learning are also known as game-based learning, educational games, and serious games aimed at teaching specific subjects to attract and interest students while providing a fun learning experience. Digital game-based learning is the integration of educational content with computer and video games to engage and motivate beyond leisure activities [21] compared to traditional learning methods. Serious games have been proven to not only attract students to play but also to interest them to interact with the games that create a real learning experience and help them achieve their mission and learning goals of a given subject [23]. According to [20] digital games are very flexible and have the advantages of applying the learning principles and orientations that provide the theoretical foundation as to the efficacy of digital game as a pedagogical tool. One of the most popular games among students today is Minecraft [24]. Previous studies found that Minecraft has benefited students to collaborate as learning progresses.

To implement the use of serious games in education, they must have the advantage of being an interesting, enjoyable[25], motivating, and effective learning tool [26]. Serious games have several advantages in education such as providing a more attractive learning environment for achieving better learning outcomes [27] that can attract students' attention, interest [25], motivation [28] and engage them in educational experiences with a view to achieving specific learning goals and outcomes [25]. Additionally, serious games can increase the efficiency and ability of students [29] and help them understand the content of lessons [30] improve high productivity skills and self-efficacy [31]. Serious games implement the principles of learning by applying various elements as described in the behavioural, cognitive [21], and constructive theories [32]. They also enable development of thinking strategies such as problem-solving, critical thinking, collaborating, creative, automaticity and a host of other higher order thinking skills [25].

The acceptance of digital games by the current net generation is better because they are born in the era of digital technology industry. According to [33] the perception, interest and tendency to video game have a high inclination of children. Furthermore, computer games are today important part of children's leisure and this generation has game-based learning skills [34]. The use of digital games is capable of building the skills required by the net generation, therefore, they should be embedded in school learning. In addition game-based interaction makes their learning more effective [34] and interesting especially among young learner. Furthermore, good serious game design for education has great potential to train or educate students in improving thinking skills which is the main agenda of education in Malaysia in particular. This is because the aspects of digital games involve the selection of strategies and resolution of conflicts as well as problems which are related to one's thinking skills [3]. Playing digital games also embeds critical thinking skills among students to obtain and find solutions in gamified situations.

III. Method

This study was conducted through semi-structured interview sessions with two School Improvement Specialist Coaches (SISC+) the officers from the District Education Office (DEO) and eight science teachers to identify the application of the HOTS elements in a science subject. Respondent selections were based on their knowledge and experience in implementing HOTS in science. In addition, document analysis was conducted to investigate teachers' knowledge of HOTS and the applicability of this element through documents used by teachers in teaching and learning science. Data from these documents were used to verify the information obtained from the interviews. The documents analysed in this study were the daily lesson plans (DLPs). Further investigation to identify students' expectations of interface game design was conducted involving 60 primary school students using two educational games - *Electric Circuit* and *Learning Circuit.* The features for interface game design identified from previous studies were used in a questionnaire.

IV. Results and Discussion

A. Interview

Based on the interviews conducted, the teachers reported that students had difficulties understanding the concepts of science due to several factors. One of the main factors is that science education is said to be abstract, indirectly affecting the attitude of students in science education. The implications of this situation have left students feeling less interested, less motivated, anxious, and worried about the science subject. This situation has led to a decline in students' achievement in science.

Moreover, the elements of HOTS are difficult to be implemented in science teaching and learning. Although the learning approach used is student-centred, the students are still exposed to memorisation and exam-oriented techniques. Teaching and learning activities are limited to the classroom setting only which cause students to not remember the things they have learned. However, students like to perform experiments, but they still do not understand the related science concepts. They are finding it difficult relating HOTS in science to daily life. HOTS situation experienced at home cannot be applied when visualised in the form of a question. This situation leads to the weakness in the aspect of students' thinking skills as they are less trained to think broadly and critically.

The study has also found that teachers' knowledge of HOTS are still low as some of them could not explain all the HOTS elements. Despite the exposure to the HOTS elements of teaching, it is evident that teachers are still unclear about the use of the elements in teaching and learning. The findings also showed that teachers are not sure of the strategies and methods to incorporate thinking skills and how to train students for HOTS. Additionally, teachers are bound by the school schedule and still tied to the syllabus provided in the textbook.

Currently, HOTS implementation in teaching and learning can also be applied using technology tools. However, interviews with SISC+ officers found that developers of educational digital games only cover contents that they felt necessary without considering the local syllabus and do not implement all HOTS elements. Moreover, existing games do not encourage users to think and do not generate inference skills. The finding is supported by previous study [8] which existing digital games are developed without considering students' learning styles, differentiation, and learning strategies. Table II summarises some of the most significant findings derived from the preliminary study related to the issues, challenges, and requirements in teaching HOTS in science from educators' perception.

TABLE II. The issues, Challenges, and Requirements in Teaching HOTS in Science Education from Educators' Perception

Category	Issues / Challenges	Requirements
Student attitude	1. Less interested in science 2. Less motivated in science 3. Have anxiety of science 4. Low achievement in science	1. Use digital game 2. Use interesting interface design based on students' perceptions 3. Use of ARC theory (motivation)
HOTS in Science	1. Student cannot answer the HOTS questions 2. Difficult to understand the concept of HOTS in Science 3. The HOTS situation experienced at home cannot be applied when visualised in the form of question. 4. Students are not able to retain in memory the things they have learned 5. Students like to perform experiment, but still do not understand the related science concepts. 6. Difficulty relating science to daily life.	Implementing HOTS in game designs through tasks/challenges
Teaching strategy	1. Teaching students skill to answer exam questions only (low thinking skill) 2. Teachers' knowledge of HOTS are still low 3. Teachers are not sure the strategies and methods that can be used to incorporate thinking skills and how to train students for HOTS.	Designing HOTS in context of Science subject
Technology approaches	1. Technology tools being used in teaching and learning such as Courseware, Video, Audio, Internet, Simulation, and Game digital 2. Existing contents that developer felt necessary without considering the local syllabus. 3. Existing games do not encourage users to think 4. Not all HOTS elements are implemented in the existing games 5. The content of existing games do not relate with real life situations 6. Existing games do not generate inference skills	1. Implementing digital game approach focused on HOTS 2. Considering local education syllabus, which are KSSM and KSSR. 3. Implementing HOTS elements in digital games

B. Document Analysis

The results of the document analysis show that teachers did not specify the HOTS elements planned in writing in their daily lesson plans. Students were not trained to perform high-level thinking during teaching and learning. The teaching and learning methods such as experiments are still tied to the syllabus provided in the textbook. The findings have also shown that the teachers did not use varied teaching aids to implement the HOTS elements. The teachers only used textbooks and no teaching aids. In conclusion, teachers still have difficulties in implementing HOTS in teaching and learning. Furthermore, the implementation of all HOTS elements is not yet done in the teaching and learning of science.

C. Questionnaire

Studies showed that the students have very high interests in using digital games for learning purposes [4]. However, there are some requirements for more interactive learning experiences in digital games based on students' expectations.

The results show that students rating is less than 3.0 (mean score) for the interface game design elements, text, image, audio, and animation. These are important elements to attract students to play games while learning science. Based on these games, the students reported that they find it difficult to understand the language, terms and the meaning of the sentences. The language used does not match their age although the role of language is important to understand the content of the game. Therefore, the use of familiar language and the means of delivering information in the form of sentences should be concise to facilitate students' understanding of the game's content.

Additionally, these games have shown the use of unrelated and uninteresting images for science learning, indirectly making it difficult for students to grasp the content. Science is a subject that involves real-life environment, therefore, the animated images in games need to be interesting and relatable to real situations that link science to everyday life.

The use of unattractive audio has also bored the students. This shows that the use of audio is important to attract students in this age group to play the game. Therefore, using interesting and appropriate background music and sound effects related to the realistic story, environment, character, component and movements for science learning are important so that students do not get bored but are more excited to play while learning.

Moreover, the use of uninteresting animations has left students bored and less interested in the game. The use of interesting and compatible animations in term of story, environment, character, component and movements that relate to the real situation for science learning are important for engaging students at play and at the same time learning science.

The findings have shown that even though these games are developed for learning science, the requirements of the games selected must be aligned with users' expectations of interface game design. It is important to attract students' interest to play and enhance their motivation and thinking skills in science by using digital games. Table III shows results of the students' game requirement survey.

TABLE III. THE GAME DESIGN REQUIREMENTS FROM STUDENT'S PERCEPTION

Requirement	**Mean**	
	Electric Circuit	**Learning Circuit**
The use of word is easy to understand	2.9	2.9
The use of language is easy to understand science learning	2.35	2.65
The meaning of the sentence is easy to understand science learning	2.35	2.65
The use of images are related to science learning	2.98	2.95
The use of background music for science learning is interesting	2.3	1.56
The use of sound effects is suitable for science learning	1.56	1.96
The use of animation for science learning is interesting	2.7	1.81
The use of animation is appropriate (compatible) with science learning	2.98	2.9

V. CONCLUSION

HOTS elements are essential requirements to train students for higher thinking skills to meet the challenges of the future. Teaching and learning using serious games can encourage thought processes while playing. The requirements gathered from this study will be used to design a serious game to teach science. For future work will involve designing a serious game for teaching.

ACKNOWLEDGMENT

The authors acknowledge the help given by the Sekolah Kebangsaan Sultan Ibrahim Dua, Sekolah Kebangsaan Dato Hamid Satu, Sepang District Education Office and Pasir Mas District Education Office. We also thank the Malaysian Ministry of Education for funding the under program MyBrain15 (MyPhD) and UKM Publication sponsorship under the grant PP-FTSM-2020.

REFERENCES

[1] R. Nagappan, Teaching and Acquiring Higher-Order Thinking Skills Theory and Practice. Penerbitan Universti Pendidikan Sultan Idris, 2016.

[2] N. A. N. Y. Mohd Nazri Hassan, Ramlee Mustapha, Rosnidar Mansor, "Pembangunan Modul Kemahiran Berfikir Aras Tinggi di dalam Mata Pelajaran Sains Sekolah Rendah: Analisis Keperluan Guru," Sains Humanika, vol. 9, no. 1–5, pp. 119–125, 2017.

[3] M. Noor, B. Madjapuni, and J. Harun, "Kemahiran Berfikir Kritis melalui Permainan Digital Dalam Persekitaran Kemahiran Berfikir Kritis melalui Permainan Digital Dalam Persekitaran Pembelajaran Konstruktivisme Sosial," no. January, 2019.

[4] H. Supeno, M. Liyanthy, and E. H. N. Huda, "Game development to train critical thinking in science subjects using model of digital game based learning-instructional design," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 8, pp. 192–195, 2019.

[5] M. N. Masran, L. F. Md Ibharim, and M. H. Mohamad Yatim, "Pendekatan Pembelajaran Melalui Reka Bentuk Permainan Digital dalam Proses Pengajaran dan Pembelajaran Kanak-kanak: Isu dan Cabaran," Pendekatan Pembelajaran Melalui Reka Bentuk Permainan Digit. dalam Proses Pengajaran dan Pembelajaran Kanak-kanak Isu dan Cabaran, no. October, pp. 1–10, 2014.

[6] S. Y. Tan and S. H. Halili, “Effective Teaching of Higher-Order Thinking (HOT) in Education,” Online J. Distance Educ. e-Learning, vol. 3, no. 2, pp. 41–47, 2015.

[7] Ministry of Education Malaysia, “Malaysia Education Blueprint 2013 - 2025,” Education, vol. 27, no. 1, pp. 1–268, 2013.

[8] A. Koptelov and S. Taube, “Learning Mathematics and Critical Thinking via Computer Games Design,” J. Math. Sci., vol. 2, pp. 88–96, 2015.

[9] Y. Abosalem, “Assessment techniques and students’ higher-order thinking skills,” ICSIT 2018 - 9th Int. Conf. Soc. Inf. Technol. Proc., no. March, pp. 61–66, 2016.

[10] D. Scully, “Constructing Multiple-Choice Items to Measure Higher-Order Constructing Multiple-Choice Items to Measure Higher-Order Thinking,” Pract. Assessment, Res. Eval., vol. 22, 2017.

[11] A. Endah Retnowati, Anik Ghufron, Marzuki, Kaisyan, Adi Cilik Pierawan, Character Education for 21st Century Global Citizen. Routledge, 2018, 2018.

[12] O. May, O. May, and K. Kamp, “Metacognitive Awareness and Critical Thinking Abilities of Pre-service EFL Teachers,” J. Educ. Learn., vol. 7, no. 5, pp. 116–129, 2018.

[13] C. H. Norlizah, “Students ’ Motivation towards Science Learning and Students ’ S cience Achievement,” Int. J. Acad. Res. Progress. Educ. Dev., vol. 6, no. 4, pp. 174–189, 2017.

[14] M. N. Hassan, R. Mustapha, N. A. Nik Yusuff, and R. Mansor, “Pembangunan Modul Kemahiran Berfikir Aras Tinggi di dalam Mata Pelajaran Sains Sekolah Rendah: Analisis Keperluan Guru,” Sains Humanika, vol. 9, no. 1–5, 2017.

[15] W. Mazwati, W. Yusoff, S. C. Seman, W. Mazwati, and W. Yusoff, “Teachers ’ Knowledge of Higher Order Thinking and Questioning Skills : A Case Study at a Primary School in Teachers ’ Knowledge of Higher Order Thinking and Questioning Skills : A Case Study at a Primary School in Terengganu , Malaysia,” Int. J. Acad. Res. Progress. Educ. Dev., vol. 7, no. 2, pp. 45–63, 2018.

[16] S. Nur, D. Mahmud, N. M. Nasri, M. A. Samsudin, and L. Halim, “Science teacher education in Malaysia : challenges and way forward,” Asia-Pacific Sci. Educ., pp. 0–11, 2018.

[17] M.-T. Cheng, J.-H. Chen, S.-J. Chu, and S.-Y. Chen, “The use of serious games in science education: a review of selected empirical research from 2002 to 2013,” J. Comput. Educ., vol. 2, no. 3, pp. 353–375, 2015.

[18] P. E. Turner, E. Johnston, M. Kebritchi, S. Evans, and D. A. Heflich, “Influence of online computer games on the academic achievement of nontraditional undergraduate students,” Cogent Educ., vol. 42, no. 1, pp. 1–16, 2018.

[19] W. Westera, “Why and How Serious Games can Become Far More Effective: Accommodating Productive Learning Experiences, Learner Motivation and the Monitoring of Learning Gains,” Int. Forum Educ. Technol. Soc., vol. 22, no. 1, pp. 59–69, 2019.

[20] K. Madani, T. W. Pierce, and A. Mirchi, “Serious Games on Environmental Management,” pp. 1–44, 2007.

[21] D. Vlachopoulos and A. Makri, “The effect of games and simulations on higher education : a systematic literature review,” Int. J. Educ. Technol. High. Educ., pp. 1–33, 2017.

[22] A. S. Drigas, “Serious Games in Preschool and Primary Education : Benefits And Impacts on Curriculum Course Syllabus Serious Games in Preschool and Primary Education : Benefits And Impacts on Curriculum Course Syllabus,” Int. J. Emerg. Technol. Learn., no. January, 2017.

[23] R. Hodhod, H. Fleenor, and S. Nabi, “Adaptive Augmented Reality Serious Game to Foster Problem Solving Skills,” Work. Proc. 10Th Int. Conf. Intell. Environ., no. January, pp. 273–284, 2014.

[24] M. Pusey and G. Pusey, “Using Minecraft in the Science Classroom,” Int. J. Innov. Sci. Math. Educ., vol. 23, no. 3, pp. 22–34, 2015.

[25] T. Anastasiadis, G. Lampropoulos, and K. Siakas, “Digital Game-based Learning and Serious Games in Education,” Int. J. Adv. Sci. Res. Eng., vol. 4, no. 12, pp. 139–144, 2018.

[26] Noor Azli Mohamed Masrop et al., “Kesan Permainan Digital Dalam Pendidikan,” Proceeding Int. Conf. Inf. Technol. Soc. 2015), vol. 2003, no. June, pp. 1–7, 2015.

[27] P. Lameras, S. Arnab, I. Dunwell, C. Stewart, S. Clarke, and P. Petridis, “Essential Features of Serious Games Design in Higher Education: Linking Learning Attributes to Game Mechanics,” Br. J. Educ. Technol., vol. 48, no. 4, pp. 972–994, 2017.

[28] Mercy Trinovianti Mulyadi and N. A. M. Zin, “MMORPG Game Framework Based on Learning Style for Learning Computer Networking,” Asia-Pacific J. Inf. Technol. Multimed., vol. 8, no. 1, pp. 63–77, 2019.

[29] K. Nisa, C. Z. Zulkifli, N. A. A. Aziz, and N. M. Nordin, “Reka bentuk gamifikasi pembelajaran geografi berasaskan Permainan Geoplay,” Geografi, vol. 5, no. 1, pp. 46–61, 2017.

[30] K. Fu, T. Hainey, and G. Baxter, “A Systematic Literature Review to Identify Empirical Evidence on The Use of Computer Games in Business Education and Training,” 10th Eur. Conf. Games Based Learn., vol. 1, pp. 232–239, 2016.

[31] K. O. Ah-Nam Lay, “Developing 21 st Century Chemistry Learning through Designing Digital To cite this article : Developing 21 st Century Chemistry Learning through Designing Digital Games,” J. Educ. Sci. Environ. Heal., vol. 4, no. 1, pp. 81–92, 2018.

[32] L. Ah-nam and K. Osman, “Developing 21 st Century Skills through a Constructivist-Constructionist Learning Environment,” K-12 STEM Educ., vol. 3, no. 2, pp. 205–216, 2017.

[33] R. Aziz, H. Norman, N. Nordin, F. N. Wahid, and N. A. Tahir, “They Like to Play Games ? Student Interest of Serious Game-Based Assessments for Language Literacy,” Creat. Educ. ERA;, pp. 3175–3185, 2019.

[34] R. D. Agustin, “Serious Games for Effective Learning,” 2017 6th Int. Conf. Electr. Eng. Informatics, 2017.

Transitioning to Online Learning during COVID-19 Pandemic: Case Study of a Pre-University Centre in Malaysia

Ahmad Alif Kamal[1], Norhunaini Mohd Shaipullah[2], Liyana Truna[3]
Centre of Pre-University Studies
Universiti Malaysia Sarawak
Kota Samarahan, Malaysia

Muna Sabri[4]
Faculty of Medicine and Health Sciences
Universiti Malaysia Sarawak
Kota Samarahan, Malaysia

Syahrul N. Junaini[5]
Faculty of Computer Sciences and Information Technology
Universiti Malaysia Sarawak
Kota Samarahan, Malaysia

***Abstract*—In the last decade, online learning has grown rapidly. However, the outbreak of coronavirus (COVID-19) has caused learning institutions to embrace online learning due to the lockdown and campus closure. This paper presents an analysis of students' feedback (n=354) from the Centre of Pre-University Studies (PPPU), Universiti Malaysia Sarawak (UNIMAS), Malaysia, during the transition to fully online learning. Three phases of online surveys were conducted to measure the learners' acceptance of the migration and to identify related problems. The result shows that there is an increased positivity among the students on the vie of teaching and learning in STEM during the pandemic. It is found that online learning would not be a hindrance, but blessing towards academic excellence in the face of calamity like the COVID-19 pandemic. The suggested future research direction will be of interest to educators, academics, and researchers' community.**

***Keywords*—E-learning; STEM; coronavirus: pandemic; education technology; assessment; technology acceptance**

I. Introduction

In the last decade, online learning has expanded rapidly due to its convenience [1]. Online learning attempts to provide flexibility to study ubiquitously for both the instructors and learners [2][3], and it is without its unending challenges. However, the world is shaken with the outbreak of the Coronavirus (COVID-19) outbreak [4]. The situation has forced learning institutions to impose a temporary halt in the academic calendar. Certain level of education are in dilemma whether to abide by the enforcement or to abruptly welcome online learning [5]. The Centre of Pre-University Studies (PPPU), Universiti Malaysia Sarawak (UNIMAS), Malaysia chose to complete its academic calendar, migrating its teaching-learning to online.

Thus, the objective of this paper is to study the students' perceptions of the sudden shift to online learning in terms of participation and examination. This paper presents an analysis of the feedback from the students in PPPU. It is high time for pre-university instructors and learners to readjust their preparedness to tackle the challenges in the migration of offline to online learning. Among the essential elements that need to be addressed include instructors' and learners' readiness of the transition to online teaching and learning. This paper is structured as follows: Section II discusses the literature review. Section III presents the case study of the PPU migration to online learning. Meanwhile, Section IV explains the methodology. Section V presents the results of the survey. Next, Section V discusses the findings. Finally, Section VI summarizes the project and offers future research opportunities.

II. Literature Review

In Malaysia, the effect of the global pandemic has hampered the learning institutions during the mid-semester break of undergraduate programs and the ongoing second semester of pre-university programs [6]. To further tackle the alarming infection rate of the deadly coronavirus, the Malaysian government had issued a movement control order (MCO) [7] that fully dampen the learning institutions' operational activities. Thus, with little to no option left, learning institutions should opt to alter its course of action from the standard norms to an already seemingly positive alternative of embracing online learning. However, the change must be well planned and appropriately designed to avoid further disruptions caused by the MCO. As the situation provides, there would be ample time to prepare a good instructional design of bachelor degree programs to suit the needs of the current learning environment.

However, online learning comes with massive challenges. Firstly, the students need to have technology access as the primary indicator of the online learning readiness [8]. As students also take their learning independently, instructors may also need more time to design their content delivery effectively [9] as learners will most definitely be facing technical and adapting difficulty. Highlighting a report from UNESCO reported that over 87% of the world's student population from more than 160 countries were impacted by the lockdown [10]. In Malaysia, this unprecedented crisis has provided an opportunity to improve online education for almost 5 million school students and 1.2 million university students [11].

Due to the pandemic, particularly when all educational activities are stopped, online and web-based learning platforms have become dramatically popular. It allows universities to adapt their conventional blended-based learning during the pandemic quickly. However, the migration process onto online learning must not be time-consuming and easy to set up. Table I shows the method for institutions with the prospect to adopt online learning [12] during the pandemic.

TABLE I. METHOD FOR THE TRANSITION FROM FACE TO ONLINE LEARNING

Function	Method	Applications software
Teaching delivery	Lectures can be pre-recorded then uploaded (offline) or streamed live (online).	*ZOOM Cloud Meeting, Youtube*
Assignments and evaluation	Students upload their quizzes or assignments online.	*Socrative, Google Docs*
Peer-interaction	Group discussions and projects are conducted online.	*Google Hangouts Meet, Microsoft Teams*
Learning resources sharing	Learning materials are shared in a digital learning environment through a learning management system (LMS)	*Blackboard, Google Classroom, Moodle Cloud*

Besides, web-based training tools have been widely used by physicians in the US as learning resources [13], and has been demonstrated to be successful. Thus, universities, colleges, and schools have resorted to online learning. Meanwhile, technology-enhanced distance learning (TEDL) [14] is linked with the 'modern teaching machines'.

However, adapting to the new normal is not a straightforward process. In response to the MCO, higher education institutions across the nation must revoke in-class teaching methods. They must execute online electronic communication platforms to facilitate teacher-student interaction. Nevertheless, this approach may be inadequate for certain field like hands-on medical [15] or technical education.

Meanwhile, a dataset involving 460 students in Vietnam reported how the students responded to the situation related to e-learning tools and skills during the nationwide school closures due to COVID-19 [16]. Meanwhile, three limitations pertaining to online learning behavior was found among students at a university in Italy during the pandemic lockdown [17]:

- Live classrooms put a lot of workload on the teaching server.
- Students who relied on 4G or lower access experienced restricted throughput.
- Some users reported poor Internet reliability due to increased traffic.

While the pandemic has shocked conventional face-to-face instruction, it has now provided learning institutions with a unique opportunity. Thus, there is a need to refurbish the existing way of delivering learning. Commonly, formal education has always depended on a traditional face-to-face approach. For example, it was reported [18] that despite the current popularity of online learning, only less than 5% of classes utilized it.

Therefore, it is the time for educational institutions to adopt disruptive learning technologies, especially during the disease outbreak, and its consequent recovery period [19]. COVID-19 has given rise to a sheer necessity of online and blended learning approaches. Blended learning is critical to distance and open education mostly during the emergence of the pandemic, as it is especially useful for teaching and learning processes in the rural areas [20]. A good example would be the State of Sarawak, Malaysia where almost half of its population lives in rural areas.

Thus, the unanticipated transition to blended learning nicely fits into the context of the COVID-19 pandemic. It is suddenly of utmost importance to education. Therefore, teachers and educators need to keep up with the evolving learning and tools theories to support learners' needs. As a result, online cloud-based platforms such as ZOOM Cloud Meeting and Google Drive are presently essential to support diverse and geographically disperse learners from all four corners of the globe. The educators' role has changed from the "sage on the stage" to "the guide on the side".

Various teaching and learning modes like media social (Facebook, Whatsapp, Telegram), live video conferences (Zoom, Microsoft Teams) as well as pre-recorded lecture videos (Youtube) were deployed. The situation has appeared as an opportunity for learners to consume and instructors to diversify via the flexibility on the delivery and timing of online learning [21]. The synchronization of online classes can help students feel a stronger sense of connection to their peers and instructors. Students would have full control and freedom [22] to complete their course learning materials at their own time from any location with Internet access.

III. The Case Study

PPPU, UNIMAS offers one-year programs, namely Foundation for Physical Sciences, Foundation for Life Sciences, and the International Foundation in Science. The programs consist of two semesters (18 weeks per semester). For the current session, there are 613 enrolled students with 63 academic and administrative staff members involved.

The COVID-19 pandemic has forced PPPU to a temporary shutdown. The teaching and learning process must be transformed into remote instruction with a learning-from-home approach. The MCO has halted the regular classes in between a running second semester (December 2019 to April 2020), which led to the transition to online learning. Since the MCO, all educational institutions were not allowed to proceed with their face to face learning activities. However, the ministry's directive declares that online educational platforms may only be used if adequate preparation has been done, and students' connectivity to the Internet is satisfactory. Hence, PPPU has launched a survey to gauge the students' accessibility to the Internet to allow for the continuance of online learning. Being the pioneering center to entirely adopt online learning within a short period, the challenges faced by PPPU's students and academicians were significant.

However, with full commitment and excellent support, this initiative has been accomplished. PPPU has gradually initiated full online delivery and online assessment beginning 1 April 2020. The university's learning management system (LMS) named eLEAP was used. This resulted in the migration of fully online learning, involving 613 PPPU students. *eLEAP* allows lecturers and instructors to design their virtual classes and deliver their courses that include various tools to enable teaching and learning to be delivered in synchronous and asynchronous modes.

All PPPU's assessment activities in March to April were conducted online, including the final examination. Over 613 PPPU students have successfully taken their final examination (27th to 30th April 2020). The assessments were conducted through online report submission, take-home tests, and online examination via eLEAP. Other suitable assessments have also replaced the laboratory's practical assessments.

IV. Methodology

Through the time PPPU migrated towards online learning, the surveys were conducted in the first three weeks to measure the learners' acceptance of the migration and identify pressing issues. The online survey was deployed in three phases, namely, Phase 1, Phase 2, and Phase 3 (sample size: N_1 = 557; N_2 = 332; N_3 = 354). The survey consists of a 5-point Likert scale (1-strongly disagree, and 5-strongly agree) for statements representing students' perception of the transitioning of traditional classroom sessions to online learning mode.

V. Results

The analysis shows that the mean score of the students' responses to the statement increases over time, represented by the phases. Students tended to agree more in terms of their personal device and Internet efficiency by the third phase. Moreover, they became more participative, less anxious and were capable of performing tasks while enjoying their online learning sessions. Learners also believed that learning new knowledge and concept online is not hard. Besides, the learners understood its usefulness as the learning empower them to work with their peers.

The online learning content prepared was very interactive, and technical support was also sufficiently provided. As a result, more students felt that online learning was able to replace their face-to-face classes effectively. Statements 13 to 18 represent the score by individual course performances of the students' point of view. The result shows that all courses have an increase in their means from the second to the third phase. The learners' thoughts on online mode of examination also increased significantly over time as they trusted the instructors to prepare, then assess them fairly. This result suggests an overall improvement in the learners' online learning experience and proves that the migration and transition to online learning was a success despite the emerging circumstances. Table II shows the descriptive statistics of the students' perception of UNIMAS PPPU online learning deployment.

TABLE II. Descriptive Statistics of the Students' Perception of UNIMAS PPPU Online Learning

Statements	**Mean**			**Std. dev.**		
	Phase			**Phase**		
	1	**2**	**3**	**1**	**2**	**3**
Device efficiency in accessing online content	3.86	3.82	3.99	0.95	0.91	0.85
Internet connection efficiency in accessing online learning content	3.70	3.64	3.83	0.99	0.97	0.94
Participation in online learning content so far	4.37	4.51	4.74	0.89	0.78	0.56
I feel anxious using online learning to learn	3.50	3.56	3.31	1.03	1.00	1.00
I think I can perform instructions and assignments well in the online learning content	3.26	3.40	3.71	0.98	0.85	0.87
I find that learning via online learning is enjoyable	2.89	2.94	3.37	1.10	1.04	0.96
I believe that learning new things via online learning is easy	2.75	2.78	3.10	1.03	0.99	0.99
I think that the online learning content so far is useful for learning	3.61	3.69	3.96	0.97	0.93	0.86
I feel that online learning content enables me to work well together with my peers	2.76	2.95	3.23	1.13	1.07	1.07
The online learning content that I have participated so far is interactive	3.38	3.53	3.80	0.95	0.91	0.87
I think that there is sufficient technical support to help students access the online learning content	3.39	3.53	3.81	0.95	0.94	0.90
I feel that the online learning content provided is sufficient as a replacement for face-to-face class	3.02	3.04	3.44	1.15	1.11	1.07
Quality of online learning content: [Biology]	-	3.96	4.04	-	0.88	0.90
Quality of online learning content: [Physics]	-	3.80	4.01	-	0.84	0.84
Quality of online learning content: [Chemistry]	-	3.71	3.97	-	0.85	0.88
Quality of online learning content: [Mathematics]	-	3.54	3.73	-	0.91	0.95
Quality of online learning content: [English]	-	3.77	3.90	-	0.87	0.89
Quality of online learning content: [ICT Competency]	-	3.73	4.01	-	0.84	0.84
I think that online examination is a fair approach to replace the traditional examination considering the current situation	2.87	3.40	4.09	1.42	1.29	0.84
Overall satisfaction towards the method of online examination	-	-	4.01	-	-	0.81

The results of the analysis on clusters of students' feedbacks were obtained. This section consists of three questions regarding UNIMAS PPPU online learning. The questions are as follows:

- Reasons for students not having full participation in online learning.
- Reasons for students not agreeing with online learning.
- General comments on online learning contents.

The data collected is categorized into several equivalent factors that students have encountered.

This data was captured from students answering the first question, admitting of having less than 100% participation to the online learning sessions or activities assigned (Ph. 1, Ph. 2 and Ph. 3 are N_1 = 181, N_2 = 87 and N_3 = 56 respectively). The analysis from the students' feedback reveals four main factors of difficulties with online learning faced by students. Technical issues (Ph. 1, 38.67%; Ph. 2, 47.13%; Ph. 3, 30.36%) and self-attitude (Ph. 1, 33.15%; Ph. 2, 31.03%; Ph. 3, 58.93%) contributed as the biggest factor in the learners' absence from online learning. Most of the students confessed that technical issues such as Internet connection caused them to fail in completing online activities and assessments. Compatibility issues with operating systems, browsers or devices were also reported as technical issues by a few students. Online learning content administration and management (Ph. 1, 16.02%; Ph. 2, 11.49%; Ph. 3, 10.71%) such as unclear instructions or procedures, no notification for assignment dues and tight schedule between activities, assignments and consultations are also among the issues reported by the students. Environment distraction (Ph. 1, 3.87%; Ph. 2, 5.75%; Ph. 3, 0.00%) is the last of the four factors identified from the students' feedback.

The rest of the responses were recorded as other factors (Ph. 1, 8.29%; Ph. 2, 4.60%; Ph. 3, 0.00%) consisting of varying feedbacks such as disagreeing to online learning, unreadiness for the transition, felt that online learning is ineffective and proposing to defer the second semester to another date. Overall, Ph. 3 (self-attitude) is the biggest hurdle faced by the students, and contributed to the most significant percentage. This result suggests that learners can slowly adapt and gear themselves to embrace online learning, given the opportunity and the time. Table III shows the percentage of students who did not fully participate in the online learning content by factors and phases.

This data was collected from students who answered disagreed with the implementation of online examination (Ph. 1, Ph. 2 and Ph. 3 are N_1 = 322, N_2 = 137 and N_3 = 54 respectively). Most students expressed their disagreement due to the possibility of technical problems. Thus, they think it was an ineffective final assessment. Besides that, the attitude of students is one of the critical reasons for disagreeing with the online examination.

Surprisingly, peer attitude is considered as another primary concern for the students as they felt that it would be viable for some learners to cheat during the individual examination such as copying others or discussing the answers. But in *Ph. 3*, analysis shows that students expressed a bigger concern for themselves rather than their peers, showing their trust in the instructors to prepare a fair assessment. Course management and learner environment factors were also reported by the students; thus, making them feel the need to disagree with having online examination.

Some of the other students, however, responded with suggestions for improvement if the online examination is to be conducted. By Ph. 3, the number of respondents decreased significantly as students were exposed to a mock online examination conducted by the Mathematics course. Later, the learners applied to their first-hand experience for the actual examination. This helps students to evaluate the effectiveness and relevance of online examination. However, a few of them hoped that the online examination would be replaced with better and more effective assessments in the future. Table IV shows the percentage of students who disagreed with the online examination by factors and phases.

This data was obtained from students who answered the third question to provide general comments on the online learning transition (Ph. 1, Ph. 2 and Ph. 3 are N_1 = 246, N_2 = 83 and N_3 = 68 respectively). In this analysis, the students' feedbacks were clustered into similar factors and quantified accordingly to obtain the percentages. Positive feedback on online learning from the students (Ph. 1, 41.87%; Ph. 2, 37.35%; Ph. 3, 48.53%) slightly increases over phases. This indicates that most of the students can participate in and obtain new experiences with online learning activities and assessments despite some negative feedbacks provided which may be affected by some uncontrolled factors (Ph. 1, 20.33%; Ph. 2, 14.46%; Ph. 3, 7.35%). The good news is that the negative feedbacks display a decreasing trend.

Most of the students suggested improvements for online learning in every phase since they find that online learning is the best method for teaching and learning considering the pandemic (Ph. 1, 28.46%; Ph. 2, 34.94%; Ph. 3, 32.35%). This goes to show that the learners are well aware of their rights and responsibilities in an online learning environment. For other factors (Ph. 1, 9.35%; Ph. 2, 13.25%; Ph. 3, 11.76%), some students suggested ending the semester because they feel that they are not adequately prepared for online learning. Other than that, some feedbacks have little to no relevance with the online learning transition. Table V shows the percentage of the students' general comments on online learning by factors.

TABLE III. PERCENTAGE OF STUDENTS WHO DID NOT FULLY PARTICIPATE IN THE ONLINE LEARNING CONTENT BY FACTORS AND PHASES

Phases	**1**		**2**		**3**	
Sample (N)	181		87		56	
Respondents	*n*	%	*n*	%	*n*	%
Technical Issues	70	38.67	41	47.13	17	30.36
• Internet connection	57		32		14	
• *eLEAP*	7		5		1	
• Device	6		4		2	
Self-Attitude	60	33.15	27	31.03	33	58.93
Learning Management	29	16.02	10	11.49	6	10.71
Environment	7	3.87	5	5.75	0	0.00
Others	15	8.29	4	4.60	0	0.00

TABLE IV. PERCENTAGE OF STUDENTS WHO DISAGREED WITH THE ONLINE EXAMINATION BY FACTORS AND PHASES

Phases	**1**		**2**		**3**	
Sample (N)	322		137		54	
Respondents	*n*	%	*N*	%	*N*	%
Technical Issues	117	36.34	59	43.07	22	40.74
• Internet connection	94		57		15	
• *eLEAP*	6		0		1	
• Device	17		2		6	
Attitude	112	34.78	44	32.12	20	37.04
• Self	42		22		18	
• Peers	70		22		8	
Course Management	16	4.97	7	5.11	4	7.41
Learner Environment	18	5.59	6	4.38	0	0.00
Others	59	18.32	21	15.33	8	14.81

TABLE V. PERCENTAGE OF STUDENTS WHO WERE NOT ADEQUATELY PREPARED TO PARTICIPATE IN ONLINE LEARNING CONTENT BY FACTORS AND PHASES

Phases	**1**		**2**		**3**	
Sample (N)	246		83		68	
Respondents	*n*	%	*N*	%	*n*	%
Positive	103	41.87	31	37.35	33	48.53
Negative	50	20.33	12	14.46	5	7.35
Suggestion for improvement	70	28.46	29	34.94	22	32.35
Others	23	9.35	11	13.25	8	11.76

There is a decline in students having problems and an increase in positivity in online learning during the transition. Most of the students were shocked by the new circumstances that forced them to participate in online learning activities and assessments provided in the environment of eLEAP with zero experience, making them feel less confident and demotivated. In Ph. 2 and Ph. 3, more students became more familiar with the learning environment. Instructors were also provided with more precise procedures related to online learning.

VI. DISCUSSIONS

In overall, a planned transition to online learning is a well-thought idea of improving teaching and learning delivery. In this case, the transition is of a desperate situation. This study is geared to the reporting the students' experiences of this rare occurrences, to curb with their teaching and learning given the sink or swim circumstances. The leaders and instructors of PPPU sought as many resources as possible for the benefits of the learners. The survey deployed and the data collected was originally used to clearly identify the issues amlong learners that need to be addressed as soon as possible to support their learning process, for the completion of their study.

In this section, the major issues commonly and frequently arise among learners are reported. Below are some outlined main challenges that institutions may face when migrating to online learning in similar emergency situations like COVID-19.

A. Factors affecting Online Learning Experiences

Technical challenges in terms of both hardware and software requirements) are the most common issues when delivering teaching and learning online [23]. Neglecting the hardware requirement, students are not required to have above average computer skills [24]. Thus, technical support provided by instructors [25] is compulsory to ensure a smoother teaching and learning process.

Meanwhile, unpleasant emotion may hinder online learning [26, 27]. The feeling of anxiety includes overwhelming apprehension, worry, distress, or fear—and may worsen. Learning without an instructor's interaction makes students less comfortable. As interactivity is a key element for an excellent online learning, it produces positive learning impact [28]. Thus, interactivity is regarded as a vital element for an impactful online learning experience [29]. Moreover, teamwork also plays a huge role to increase positivity in learning experiences [30]. Teaching science and mathematics (STEM)-related subjects is hard through traditional face-to-face means [31], let alone learning new concepts online. Thus, a proper instructional design is needed to enable an appropriate migration to online learning, allowing the students to not falter in learning [32]. In addition, online learning musters the learners' lifelong learning ability [33], especially as the Internet is the biggest library for new skills and knowledge.

B. Students Participation in Online Learning

Learners' participation is crucial in determining excellent online learning results. This can be challenging when the learners may be in an anxious state due to being out of comfort zone, added with technical complexity [34]. Competency level in digital literacy may disrupt the students' motivation to be engaged online. Besides, managing self-learning time independently is also a challenge. Also, the diversity of learners must be considered to ensure active participation in online activities [35]. Besides, online learning enables students to work and study at their time and venue [36], which may lessen the cost of distance education, while tackling learners from rural and remote areas [37]. Online learning has proven to increase the activity of discussions and collaborations [38]. As a result, shyer or introvert learners can participate actively [39]. Lastly, learners' participation can be boosted by the usage of interactive tools [40] such as online forum and videos.

C. Online Examination

Many students feel uncomfortable taking examination online due to the potential of unjustified actions such as cheating [41] or discussing for individual assessment [42]. Furthermore, authentication of examination takers is a big concern [43]. In terms of question design, the preparation can be a major hassle as it can be tedious while raising fairness issues [44][45][46]. Despite the constraints, online examination improvises assessment by reducing the operational time. The management of the examination becomes flexible in terms of planning and executing with various tools [47]. The tools may be embedded with an automated marking process [48], making the students receive their feedbacks faster [49]. Additionally, online examination help institutions to reach remote students [50]. Overall, the online examination may produce an equal level of students' performance [51] and also helps to boost the students' results [52] as compared to the paper-based examination.

D. Limitations

The sudden transition to online not only shocked the students, but the instructors as well, resulting in an immense challenge of implementing it “right”. Prior to the pandemic, online learning was simply a voluntary alternative. Now, policy instructors are imposed on transferring all the planned and available teaching and learning materials to the cloud, resulting in extra burden on their job [53]. Furthermore, on-the-go training is needed for a responsible instructor [54]. This is very important among instructors who are not well-versed in computer skills or ICT tools [55]. There are cases where certain instructors could not convey their ‘learning style’ due to flaws in implementing online learning [56]. Additionally, instructors find it difficult to gauge how much a student can take or handle [57]. By proper planning and outlining, instructors could be more well-equipped technically for a better instructional design, and learners would be well-prepared for a full teaching and learning approach via e-learning. Then, a more rigorous research measurement tool may be applied to study the impact of a comprehensive online learning.

VII. Conclusion and Future Works

In conclusion, the students’ feedback from PPPU, UNIMAS, Malaysia have been presented, during the transition to full online learning through online surveys among 354 respondents. The result shows that there is an increased positivity among the students about online learning during the pandemic. Despite the challenges, online learning leads to better student participation. The present study opens up an insight into the trends on how colleges and universities react to the pandemic. Surely online learning would not be a hindrance, but a blessing towards academic excellence in the face of calamity like the COVID-19 pandemic.

Acknowledgment

This study was supported by the Centre of Pre-University Studies (PPPU) and Research, Innovation & Enterprise Centre (UNIMAS Innovation), Universiti Malaysia Sarawak (UNIMAS).

References

[1] J. A. A. Abe, “Big five, linguistic styles, and successful online learning,” Internet High. Educ., vol. 45, p. 100724, doi: https://doi.org/10.1016/j.iheduc.2019.100724, 2020.

[2] M. T. Fuller, “ISTE Standards for Students, Digital Learners, and Online Learning”, pp. 284–290, 2020.

[3] I. Chirikov, T. Semenova, N. Maloshonok, E. Bettinger, and R. F. Kizilcec, “Online education platforms scale college STEM instruction with equivalent learning outcomes at lower cost,” Sci. Adv., vol. 6, no. 15, doi: 10.1126/sciadv.aay5324, 2020.

[4] V. J. Munster, M. Koopmans, N. van Doremalen, D. van Riel, and E. de Wit, “A novel coronavirus emerging in China—key questions for impact assessment,” N. Engl. J. Med., vol. 382, no. 8, pp. 692–694, 2020.

[5] C. Hodges, S. Moore, B. Lockee, T. Trust, and A. Bond, “The difference between emergency remote teaching and online learning,” Educ. Rev., [Online]. Available: https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning, 2020.

[6] M. Chinazzi et al., “The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak,” Science (80-.)., vol. 368, no. 6489, pp. 395–400, doi: 10.1126/science.aba9757, 2020.

[7] S. Abdullah et al., “Air quality status during 2020 Malaysia Movement Control Order (MCO) due to 2019 novel coronavirus (2019-nCoV) pandemic.,” Sci. Total Environ., vol. 729, p. 139022, doi: 10.1016/j.scitotenv.2020.139022, 2020.

[8] R. A. Rasheed, A. Kamsin, and N. A. Abdullah, “Challenges in the online component of blended learning: A systematic review,” Comput. Educ., vol. 144, p. 103701, doi: 10.1016/j.compedu.2019.103701, 2020.

[9] S. L. Aj and M. Vijayalakshmi, “How Cloud Frameworks Support Blended Learning Environments,” in Emerging Techniques and Applications for Blended Learning in K-20 Classrooms, IGI Global, pp. 114–136, 2020.

[10] F. J. de O. Araujo, L. S. A. de Lima, P. I. M. Cidade, C. B. Nobre, and M. L. R. Neto, “Impact Of Sars-Cov-2 And Its Reverberation In Global Higher Education And Mental Health.,” Psychiatry research, vol. 288. Ireland, p. 112977, doi: 10.1016/j.psychres.2020.112977, 2020.

[11] J. M. Abdullah et al., “A critical appraisal of COVID-19 in Malaysia and beyond,” Malaysian J. Med. Sci., vol. 27, no. 2, pp. 1–9, doi: 10.21315/mjms2020.27.2.1, 2020.

[12] R. Varalakshmi and K. Arunachalam, “Covid 2019 – Role of Faculty Members To Keep Mental Activeness of Students,” Asian J. Psychiatr., vol. 51, no. April, p. 102091, doi: 10.1016/j.ajp.2020.102091, 2020.

[13] J. B. Stambough et al., “The Past, Present, and Future of Orthopedic Education: Lessons Learned From the COVID-19 Pandemic,” J. Arthroplasty, pp. 1–5, doi: 10.1016/j.arth.2020.04.032, 2020.

[14] T. Surma and P. A. Kirschner, “Virtual special issue computers in human behavior technology enhanced distance learning should not forget how learning happens,” Computers in Human Behavior, vol. 110. Elsevier Ltd, p. 106390, Sep. 01, doi: 10.1016/j.chb.2020.106390, 2020.

[15] R. M. Moadel et al., “Remaining academically connected while socially distant: Leveraging technology to support dispersed radiology and nuclear medicine training programs in the era of COVID-19,” Acad. Radiol., 2020.

[16] T. Trung, A.-D. Hoang, T. T. Nguyen, V.-H. Dinh, Y.-C. Nguyen, and H.-H. Pham, “Dataset of Vietnamese student’s learning habits during COVID-19,” Data Br., vol. 30, p. 105682, doi: 10.1016/j.dib.2020.105682, 2020.

[17] T. Favale, F. Soro, M. Trevisan, I. Drago, and M. Mellia, “Campus traffic and e-Learning during COVID-19 pandemic,” Comput. Networks, vol. 176, p. 107290, doi: https://doi.org/10.1016/j.comnet.2020.107290, 2020.

[18] C. H. Li et al., “Virtual Read-Out: Radiology Education for the 21st Century During the COVID-19 Pandemic.,” Acad. Radiol., pp. 1–10, doi: 10.1016/j.acra.2020.04.028, 2020.

[19] R. N. Keswani, A. Sethi, A. Repici, H. Messman, and P. Chiu, “How To Maximize Trainee Education During the COVID-19 Pandemic: Perspectives from Around the World,” Gastroenterology, 2020.

[20] M. Javaid, A. Haleem, R. Vaishya, S. Bahl, R. Suman, and A. Vaish, “Industry 4.0 technologies and their applications in fighting COVID-19 pandemic,” Diabetes \& Metab. Syndr. Clin. Res. \& Rev., 2020.

[21] M. Kebritchi, A. Lipschuetz, and L. Santiague, “Issues and Challenges for Teaching Successful Online Courses in Higher Education,” J. Educ. Technol. Syst., doi: 10.1177/0047239516661713, 2017.

[22] G. Northey, T. Bucic, M. Chylinski, and R. Govind, “Increasing student engagement using asynchronous learning,” J. Mark. Educ., vol. 37, no. 3, pp. 171–180, 2015.

[23] M. Ez-zaouia, A. Tabard, and E. Lavoué, “EMODASH: A dashboard supporting retrospective awareness of emotions in online learning,” Int. J. Hum. Comput. Stud., vol. 139, doi: 10.1016/j.ijhcs.2020.102411, 2020.

[24] L. Yu, Y. Qi, and Z. Qi, “Several Misconceptions About Online Education,” in 2019 3rd International Conference on Education, Economics and Management Research (ICEEMR 2019), pp. 250–253, 2020.

[25] C. J. Tanis, “The seven principles of online learning: Feedback from faculty and alumni on its importance for teaching and learning,” Res. Learn. Technol., vol. 28, 2020.

[26] J. Hilliard, K. Kear, H. Donelan, and C. Heaney, “Students’ experiences of anxiety in an assessed, online, collaborative project,” Comput. Educ., vol. 143, p. 103675, doi: 10.1016/j.compedu.2019.103675, 2020.

[27] J. Yu, C. Huang, Z. Han, T. He, and M. Li, “Investigating the Influence of Interaction on Learning Persistence in Online Settings: Moderation or Mediation of Academic Emotions?,” Int. J. Environ. Res. Public Health, vol. 17, no. 7, doi: 10.3390/ijerph17072320, 2020.

[28] Y. Ha and H. Im, “The Role of an Interactive Visual Learning Tool and its Personalizability in Online Learning: Flow Experience.,” Online Learn., vol. 24, no. 1, 2020.

[29] N. M. Preradovic, T. Lauc, I. Panev, and others, “Investigating interactivity in instructional video tutorials for an undergraduate informatics course,” Issues Educ. Res., vol. 30, no. 1, p. 203, 2020.

[30] D. Gamage, I. Perera, and S. Fernando, “MOOCs lack interactivity and collaborativeness: Evaluating MOOC platforms,” Int. J. Eng. Pedagog., vol. 10, no. 2, pp. 94–111, 2020.

[31] A. A. Kamal and S. N. Junaini, “The effects of design-based learning in teaching augmented reality for pre-university students in the ict competency course,” Int. J. Sci. Technol. Res., vol. 8, no. 12, pp. 2726–2730 [Online]. Available: http://www.ijstr.org/paper-references.php?ref=IJSTR-1219-27322, 2019.

[32] A. Volungevičienė, M. Teresevičienė, and U.-D. Ehlers, “When is open and online learning relevant for curriculum change in higher education? Digital and network society perspective,” Electron. J. e-Learning, vol. 18, no. 1, pp. 88–101, 2020.

[33] D. Passmore, “Transforming From the Classroom to an Online Nursing Educator: A Transformative Learning Experience for New Online Nursing Faculty,” in Enriching Collaboration and Communication in Online Learning Communities, IGI Global, pp. 82–102, 2020.

[34] J. Gillett-Swan, “The challenges of online learning: Supporting and engaging the isolated learner,” J. Learn. Des., vol. 10, no. 1, pp. 20–30, 2017.

[35] R. Kerr, I. Merciai, and M. Eradze, “Addressing cultural and linguistic diversity in an online learning environment,” EMI. Educ. Media Int., doi: 10.1080/09523987.2018.1547546, 2018.

[36] M. Rusli, “The Effects of Various Modes of Online Learning on Learning Results,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 4, pp. 100–105, 2020.

[37] B. Gilbert, “Online learning revealing the benefits and challenges,” Fish. Digit. Publ., pp. 1–32, 2015.

[38] M. Higley, “Reasons Why Collaborative Online Learning Activities Are Effective - eLearning Industry,” eLearning Industry. https://elearningindustry.com/collaborative-online-learning-activities-reasons-effective (accessed May 16, 2020), 2018.

[39] S. Asil, “The Benefits of Online Education for Children,” Tech Afghanistan, 2020.

[40] J. F. Ortega-Morán, B. Pagador, J. Maestre-Antequera, A. Arco, F. Monteiro, and F. M. Sánchez-Margallo, “Validation of the online theoretical module of a minimally invasive surgery blended learning course for nurses: A quantitative research study,” Nurse Educ. Today, vol. 89, p. 104406, doi: 10.1016/j.nedt.2020.104406, 2020.

[41] R. Bawarith, A. Basuhail, A. Fattouh, and S. Gamalel-Din, “E-exam cheating detection system,” Int. J. Adv. Comput. Sci. Appl, vol. 8, pp. 176–181, 2017.

[42] N. I. Nizam, S. Gao, M. Li, H. Mohamed, and G. Wang, “Scheme for Cheating Prevention in Online Exams during Social Distancing,” Preprints, doi: 10.20944/PREPRINTS202004.0327.V1, 2020.

[43] Y. Khlifi, “An advanced authentication scheme for E-evaluation using students behaviors over E-learning platform,” Int. J. Emerg. Technol. Learn., doi: 10.3991/ijet.v15i04.11571, 2020.

[44] P. K. Panda, “Problems Faced by Bachelor of Education Students During Computer Based Online Entrance Test,” Stud. Indian Place Names, vol. 40, no. 60, pp. 3734–3740, 2020.

[45] D. V. Kotwal, S. R. Bhadke, A. S. Gunjal, and P. Biswas, “Online examination system,” Int. Res. J. Eng. Technol., vol. 3, no. 1, pp. 115–117, 2016.

[46] J. Jiang, B. Wu, L. Chang, K. Liu, and T. Hao, “The Design and Application of an Web-Based Online Examination System,” in International Symposium on Emerging Technologies for Education, pp. 246–256, 2019.

[47] P. Kumar, “Review Study on E-Learning in Higher Education Administration and Management,” 2020.

[48] M. Thomas and A. R. Beresford, “Automated marking of free-text questions in STEM,” in Cambridge Computing Education Research Symposium, p. 14, 2020.

[49] N. Alruwais, G. Wills, and M. Wald, “Advantages and Challenges of Using e-Assessment,” Int. J. Inf. Educ. Technol., vol. 8, no. 1, pp. 34–37, doi: 10.18178/ijiet.2018.8.1.1008, 2018.

[50] S. Kausar, X. Huahu, A. Ullah, Z. Wenhao, and M. Y. Shabir, “Fog-Assisted Secure Data Exchange for Examination and Testing in E-learning System,” Mob. Networks Appl., pp. 1–17, 2020.

[51] H. T. Nennig, K. L. Idárraga, L. D. Salzer, A. Bleske-Rechek, and R. M. Theisen, “Comparison of student attitudes and performance in an online and a face-to-face inorganic chemistry course,” Chem. Educ. Res. Pract., 2020.

[52] S. S. Reddy, S. Arumugam, and S. A. Kumar, “Online Examinations to Undergraduate Engineering Students: A Case Study in an Autonomous Institution,” J. Eng. Educ. Transform., vol. 32, no. 2, pp. 61–66, 2018.

[53] J. P. Humiston, S. M. Marshall, N. L. Hacker, and L. M. Cantu, “Intentionally Creating an Inclusive and Welcoming Climate in the Online Learning Classroom,” in Handbook of Research on Creating Meaningful Experiences in Online Courses, IGI Global, pp. 173–186, 2020.

[54] H. Fulford, M. Bailey, S. Mcwhirr, and G. Stephen, “Building Online Learner Communities: an Activity Theory Perspective”, 2017.

[55] U. Verawardina et al., “Reviewing Online Learning Facing the Covid-19 Outbreak,” J. Talent Dev. Excell., vol. 12, no. 3s, pp. 385–392, 2020.

[56] K. Richardson, “Online Tools With Synchronous Learning Environments,” in Handbook of Research on Online Pedagogical Models for Mathematics Teacher Education, IGI Global, pp. 68–78, 2020.

[57] F. F. Watson, A. Pina, and J. Smaall, “A Strategic Framework for Online Learning: Insights for Developing a Masterplan for Online Learning at Your Institution,” in Online Learning Consortium Innovate Conference (2020), pp. 1–4, 2020.

Optimised Tail-based Routing for VANETs using Multi-Objective Particle Swarm Optimisation with Angle Searching

Mustafa Qasim AL-Shammari[1], Ravie Chandren Muniyandi[2]
Center for Software Technology and Management
University Kebangsaan Malaysia, UKM Bangi
Selangor - 43600, Malaysia

***Abstract*—Routing protocols for vehicular ad hoc networks (VANETs) are highly important, as they are essential for operating the concept of intelligent transportation system and several other applications. VANET Routing entails awareness about the nature of the road and various other parameters that affect the performance of the protocol. Optimising the VANET routing guarantees optimal metrics, such as low E2E delay, high packet delivery ratio (PDR) and low overhead. Since its performance is of multi-objective nature, it needs multi-objective optimisation as well. Most researchers have focused on a single objective or weighted average for multi-objective optimisation. Only a few of the studies have tackled the actual multi-objective optimisation of VANET routing. In this article, we propose a novel reactive routing protocol named tail-based routing, based on the concept of location-aided routing (LAR). We first re-defined the request zone to reduce the lateral width with respect to the lateral distance between the source and destination and named it tail. Next, we incorporated angle searching with crowding distance inside the multi-objective optimisation MO-PSO and called it MO-PSO-angle. Then, we conducted optimisation of tail-based routing using MO-PSO-angle and compared it with optimised LAR, which exhibited the superiority of the latter. The best improvement was at the optimisation point with a 96% improvement of PDR and a 313% improvement in E2E delay.**

Keywords—VANETs; Routing; PDR; E2E delay; optimization; multi-objective particle swarm; location based routing; MOPSO

I. Introduction

Vehicular ad hoc networks VANETs routing is one type of ad hoc networks. It involves connecting vehicles in the road environment for easy communications among them. It has big role in the intelligent transportation system ITS which is considered as an advanced application made to provide innovative services related to modes of transport and traffic management. The result of ITS is accomplishing safety, coordination and intelligence vehicles routing. The relying of ITS on having a reliable and robust VANETs routing is an adequate motivation for researchers to work on solving the issues of VANETs routing. Furthermore, such research can serve in economical saving due to the avoidance of traffic jam and the management of hazardous situations in order to limit the damages or the loss of human life because of road accidents. Routing protocols (RPs) play an essential role in the efficacy of an ad hoc network [1], [2], [3] and [4]. Lack of fixed infrastructure for the ad hoc network makes it a tough task to route a message from a source to its destination. While this process is easy in the case of a traditional network due to the prior availability of the network graph, it is a challenging task in the case of the ad hoc network that has a non-fixed topology. Furthermore, RPs become more challenging in the vehicular ad hoc networks (VANETs) because of the high mobility and associated dynamics of the vehicles. The need is to have a reliable and robust routing protocol that facilitates numerous tasks and applications required to operate a routing-dependent application. This becomes more important in the applications of intelligent transportation systems where fast information exchange among vehicles must be enabled. The information exchange is important for providing various traffic and safety services [4]. The literature describes a wide range of routing protocols originated from various views and philosophies. While some approaches are based on the concept of prior preparation of the route or proactive way [5], others are based on the reactive way in which the route is prepared only on demand [6]. In emergency situations or when fast delivery of message is required, proactive routing is preferred. On the other hand, when there is no restriction or constraint on the time of delivery, on-demand or reactive routing protocol is preferred for its low overhead. In either way, the performance is subject to change based on the range of parameters that have to be carefully selected depending on the optimisation approach. Many researchers have adopted or adapted various meta-heuristic optimisation approaches for selecting the best values of parameters. However, a majority of them relied on a single objective optimisation approach which affects the multi-objective nature of the problem [7]. We propose a routing protocol for VANET and optimise it for the multi-objective nature of the problem, offering more control of the performance and flexibility in responding to the need of the user or other applications that are based on the routing.

Optimisation algorithms based on heuristic searching are a big family, including a wide range of approaches such as genetic [7], particle swarm optimization [8], simulated annealing [9], ant bee colony [10], harmony searching [12] and many others [11]. Each of the approaches is inspired by a certain type of world phenomenon or metaphor. Some of them have multi-objective variants such as non-dominated sorting for genetic algorithm, and multi-objective particle swarm optimization [12] [13], sea lion optimization [14], multi-objective evolutionary algorithm [15]. Each approach has its own capability of searching that differs according to the used set of criteria of evaluating candidate solutions and their relations in the solutions and objective space. The goal of this study is to propose a novel on-demand routing protocol based on the concept of location-aided routing (LAR) and then to conduct multi-objective particle swarm optimisation for a subset of its internal significant parameter. Next, we prove the importance of using non-dominated-based optimisation for improving the routing protocols from the perspective of networks metrics, namely packet delivery ratio (PDR), E2E delay, and overhead.

This research focuses on developing multi-objective optimized reactive routing protocol of VANETs from various perspectives. We are concerned with PDR, E2E delay and overhead. The optimization changes the time period of updating the neighbour zone and the radius of the coverage zone R in order to accomplish better performance. To the best of our knowledge, this article provides the first multi-objective reactive routing protocol for VANET.

The remaining of the article is organised as follows. In Section II, we present a previous approach. In Section III, we present the proposed tail-based routing. In Section IV, we provide the results and discussion. Lastly, we provide the conclusion and recommendations.

II. Previous Approaches

The literature contains numerous approaches for developing an optimised routing protocol for VANETs. For this purpose, some researchers have adapted the theories of meta-heuristic searching. Some others have used the meta-heuristic searching for optimising the clustering approach, which will be used as the routing topology. In [16], the enhanced dragonfly algorithm (EDA) was used to minimise the energy utilisation based on the solution of clustering. For avoiding the local optimal, the algorithm was improved by incorporating Cauchy operator. This work can be criticised easily because of the fact that energy is not the top-most priority of VANET network when compared with the more important aspects related to the quality of the found clusters and the performance of routing. The other objective functions need to be incorporated to meet the multi-objective nature of the problem. In [17], meta-heuristic searching was done for optimisation, and a reputation-based weighted clustering protocol was proposed. For optimisation, a vector of various parameters was used, namely Hello_Interval, Election_Interval, ITJ_Interval PRE_Interval, CH_Timeout_ Interval, CM_Timeout_Interval, Cluster_Size, Weight of Distance, Weight of Velocity, Weight of Reputation and Weight of One-Hop Neighbours OHN. They [18] have used weighted sum, which causes the local minima because of the non-convexity of the optimisation curve. Another example of the application of optimisation of routing in VANET is the multi-casting application for countering the broadcast storm that exists in the emergency or hazardous scenarios in VANETs. In [19], an artificial bee colony (ABC) is used for optimising a fuzzy system used for predicting the highest-ranked link for routing the RREQ message. The solution was based on the fuzzy membership function and rules. For optimisation, four objective functions were used: PDR, E2E delay, throughput, and the number of control packets. However, they were used in a single objective function based on the weighted sum formula. This causes a fall in the local optimal because of the likelihood of non-convexity of the optimisation surface. Bello (2020) has used genetic optimisation for optimising the routing of VANET employing a set of parameters such as transmit power, frequency, and path loss. The objective function used has maximised the route from the perspective of a new major metric named route metric. Obviously, the optimisation does not consider the multi-objective nature of the problem. In [20], a clustering algorithm centred on moth-flame optimisation (MFO) is proposed. The approach is inspired by the movement of moths with respect to the light source. The author has used it for clustering purpose; however, the study did not clearly present the formulation of the multi-objective functions. In [21], a hybrid fuzzy logic and genetic algorithm was developed with the aim of using the fuzzy logic for weight calculation of the multi-objective functions. The application was for service provided in 5G VANETs. The approach aims at maximising the capacity and the number of fog controller base band unit controllers (FC-BBUCs) and at minimising the delay, the number of FC-ZCs that one BBUC handles and the traffic load of each FC-ZC, and consequently of each BBUC pool. It optimises connections between the FC-BBUCs and the FZ-ZCs using the hybrid fuzzy genetic. In [22], a clustering algorithm based on ant colony optimisation (ACO) was proposed. The approach uses two objective functions: delta difference value of the clusters and the summation of the distance values of all CHs from their cluster members. It also uses a weighted sum approach for the two objective functions. This causes local optimality because of the weighted aggregation of the objective functions. In [23], an optimisation for optimised link state routing (OLSR) protocol was proposed. The authors have proposed the use of eight variables for this purpose: HELLO-Interval, REFRESH-Interval, TC-Interval, NEIGHB HOLD TIME, HELLO-Interval, TOP HOLD TIME, TC-Interval, MID HOLD TIME, TC-Interval, DUP HOLD TIME, WILLINGNESS. The authors have also recommended an objective function that is formulated as weighted equation for the number of packets sent, E2E delay, the number of packets received and throughput. The problem with this approach is the fall in local minima. Some authors have used an optimisation without weighted sum aggregation of the objective functions; they adopted the concept of non-domination. In [24], a non-dominated sorting genetic algorithm was used for optimising routing protocol in VANET. OLSR protocol was used for optimisation. However, the authors have used only a small number of parameters in the optimisation, namely hello interval, TC interval, and refresh interval. Furthermore, only two objective functions were used: packet loss and E2E delay. In [25], a hybrid leapfrog algorithm and particle swarm optimisation was proposed for estimating the future position of the nodes and predetermining the link breakage. The specific goal of using PSO is to determine the optimum multiple paths for transmission, while the goal of using the leapfrog mechanism is to obtain the update mechanism. In [26], an optimisation framework for OLSR based on meta-heuristic searching algorithm such as particle swarm optimisation, differential evolution, genetic algorithm and simulated annealing was proposed. In order to carry out the optimisation, a formulation of multi-objective function based on weighted summation was done, and a subset of OLSR parameters was selected to optimise the function.

III. Proposed Tail based Routing

This section outlines the developement methodology by first presenting the general block diagram. Further, it provides the neighbour zone update and route request zone. Finally, it describes all the elements of the tail zone, route combining, route reply message, data transmission, route discovery, optimisation, solution space, multi-objective particle swarm optimisation, crowding distance, angle distribution and pseudocode.

A. General Block Diagram

A block diagram that describes TR, along with an explanation of its elements, is given in Fig. 1. The first block is the neighbour update, which is responsible for building the zone around the node with the purpose of updating the information of nodes around it. Next, the process of the route discovery is performed in which it is combined with three main blocks: tail zone update, sending route request (RREQ) message and receiving route reply message. Then, the data transmission is done based on the selected route. At this stage, the optimisation part changes the parameters of neighbour update and route discovery. For each set of parameters, the network measures are generated and used for evaluating the selected routes. The select parameter block is used to choose one operating route from a set of non-dominated solutions.

B. Neighbor Zone Update

Each node in the network sends a hello message to its neighbour to update its information. The interval of sending this message is T_{hello} and the message is periodic. It includes the following information: the ID of the node, the location information of the node x, the velocity of the node v_x, the acceleration of the node a_x and hazardous condition. Every hello message includes a timestamp to indicate the moment at which the message was transmitted is given in Table I.

TABLE I. The Topology of the Hello Message

ID	x	v_x	Timestamp

C. Route Request Message RREQ

The RREQ message is generated to find a route to a certain destination D. It includes the coordinate information of the source node x_s and the ID of the node ID_s. It also contains the ID information $ID_{D,}$ the coordinate information of the node x_D and the time stamp at the last update of the RREQ message. The layout of the message is given in Table II.

TABLE II. The Topology of the RREQ

ID_s	x_s	ID_D	x_D	t_D	Timestamp

The RREQ message will be multi-casted to a subset of nodes in the neighbour zone. The subset includes the nodes with high probability to deliver the message. Hence, we select a percentage (LT) that indicates the high probability nodes for delivery. They include the closest LT to the destination based on the expected location of the node. In the case of having a number less than LT in the neighbour zone, the node will send the message to all the nodes in the zone.

D. Tail Zone

Let us assume that the nodes are traveling in the highway environment as depicted in Fig. 2. A source node S decides to send data message to a certain destination node named D. Now, an RREQ message will be initiated. The RREQ message will be sent to the neighbour list of nodes S . However, for preventing the redundant transmission of messages that could cause flooding, the node will only re-broadcast the message if it exists within the tail zone of the node. The tail zone is defined as the moving rectangular zone that follows the node. As can be seen in Fig. 2, the source node S is supposed to send a packet to the destination node D. The request zone is a tail that starts from the destination node and ends at the proximity of the source node. However, its width is lower than the lateral distance between the source node and the destination. It helps to minimise the broadcast storm of the route request packet and minimise the overhead.

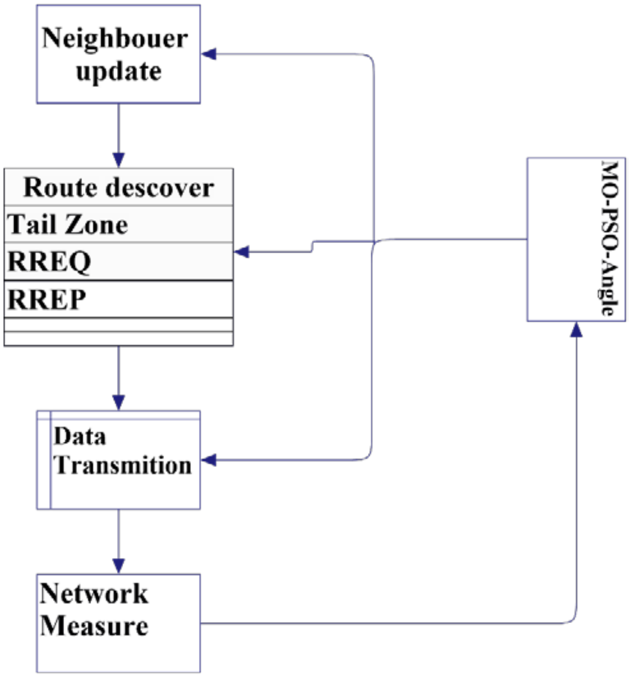


Fig. 1. Block Diagram of Tail-based Routing.

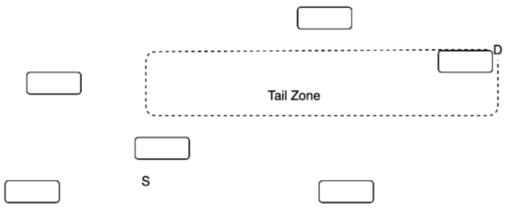


Fig. 2. Examples of Source Vehicle S and Destination Vehicle D and the Tail Zone between them in Tail-based Routing.

In order for each node to decide whether it is located inside the request zone or outside, it applies the equation.

$$x_{D,t} = x_{D,t_o} + (t - t_o)v_{D,t_o}$$

Where

$x_{D,t}$ the location of node D at t

x_{D,t_o} the location of node D at t_o

v_{D,t_o} the speed of node D at t_o

This means that any intermediate node X that receives the RREQ message will check the time t_o of the last update of the location of D and compare the time with its own time $t_{o,1}$. It will then update its own information if $t_{o,1} > t_o$. Next, the node will use the location information of the source node S with its own location to make one of two decisions: either to drop the message if the node X is outside the tail zone or to rebroadcast the message if the node X is inside the tail zone.

E. Route Combining

At each node, X receives the RREQ and broadcasts it, and new information is added to RREQ in the form of the ID of the node X and its current coordinate information. This is for combining a whole path from the destination towards the source for the route reply message.

F. Route Reply Message RREP

When the RREQ message arrives at its destination, the route reply message is sent from the destination towards the source. This message will go through the reversed route provided in the RREQ. The destination node will consider the first RREQ message that arrives for RREP. Other RREQ messages will be discarded.

G. Data Transmission

Once a route is established and the route reply message reaches the source, the data transmission will begin from the source towards destination through the path.

H. Route Discovery

Route discovery is initiated when the node needs to use a route for data transmission. However, the node will use the last used route to the destination if no route error message (REE) is generated for its last use. The REE is generated at the last node that cannot reach its neighbour in the route. In the case of non-validity of the route due to REE generated, the node will initiate a new route discovery. The new route initiation will include a new type of RREQ message that contains information towards the last valid node, i.e. the node that generated REE in the last transmission. In this case, the overhead will be reduced because of enabling the valid part in the last discovered route. On the other hand, in the case of non-returning RREP message, the node will not trigger a new route discovery immediately; it will wait for time T_w or when the condition in the neighbour zone changes to have at least a percentage (PR_{min})of new vehicles entered. The pseudocode that shows the route discovery process is provided in Table III. We assume that the node has a generated packet that needs to be sent to a certain destination. The output is the route that will be used for sending the packet. The process starts with checking the location of the destination node. It is assumed that the location is given in any of the hello messages that the node uses to update its location with respect to its neighbours. Once the location is determined, the node will build the tail zone containing all the nodes that are moving in the road behind the destination node and in the same direction of the destination node. Next, the node will send RREQ message and nodes within the tail zone will have the responsibility of forwarding the packet. The node will wait for the first RREP message which contains the route which will be used to send the packet. If no RREP message is received, the node will wait for a time T_w.

TABLE III. PSEUDOCODE OF THE ROUTE DISCOVERY PROCESS IN TAIL BASED ROUTING

```
Input
generatedPacket
subjectNode
Start
check the location of the destination node
build the tail zone                    //starting of RouteDiscovery
sends RREQ message to the nodes in the tail zone
if(receving RREP message)
wait for the first RREP message
sends the packet in the route of the RREP
else
wait for Tw
end
End
```

I. Optimization

The goal of optimisation is to improve the performance of the developed protocol to satisfy various aspects of the routing performance measures. Typically, routing needs to satisfy high PDR, less overhead and E2E delay. The implicit conflict of the objectives requires multi-objective optimisation.

J. Solution Space

For solution space, we consider three variables: the time period of updating the neighbour zone T_{hello} and the radius of the coverage zone R.

$x \in X \subset R^3$ where $x = (T_{hello}, R, LT)$. The goal is to find the best solution x^* that provides the best performance in terms of the packet delivery ratio, the overhead and the E2E delay. Then, each solution is evaluated by three objective functions

$$f_1 = PDR(x)$$

$$f_2 = -overhead(x)$$

$$f_3 = -E2EDelay(x)$$

Adding the constraint of LT = 1 implies the variant of reduced tail-based routing because the lateral distance of the request zone will be constant. However, we propose another variant where we also optimise T_{hello}, R and the lateral distance of the request zone LT. We call it scaled reduced tail-based routing where $x = (T_{hello}, R , LT)$ $x \in X \subset R^3$.

We notice that each of the two parameters has an impact on the values of the objective functions. Our goal is to maximise f_1, f_2 and f_3 for ensuring better performance by increasing the PDR, decreasing the overhead and decreasing the delay.

Increasing the value of LT implies more overhead; however, it decreases the potential of reaching the destination in a short time, which in turn affects the E2EDelay and the PDR. Hence, there is an implicit confliction between the objectives. This motivates us to use multi-objective optimisation for finding the best solutions.

K. Multi Objective Particle Swarm Optimization MOPSO

We use multi-objective particle swarm optimisation, which is an extension to the single objective particle swarm optimisation. In this algorithm, a set of initial solutions named as swarm is created and evaluated based on the three objective functions. Next, the non-dominated set of solutions is found and stored in the repository. The algorithm then selects the best global solution and it moves each solution using a mobility equation that combines three effects: inertia, best personal and best global. This is presented in the following equations:

$$v_t = wv_{t-1} + C_1W_1\left(x_{bg} - x_{t-1}\right) + C_2W_2\left(x_{bp} - x_{t-1}\right)$$

$$x_t = x_{t-1} + v_t$$

$\mathrm{C}_1, \mathrm{C}_2$ constants

$\mathrm{W}_1, \mathrm{W}_2$ random numbers between 0 and 1

L. Crowding -distance CD

The crowd distance is calculated based on the distance between one of the non-dominated solutions and its two adjacent solutions. Having a higher distance means more potential of exploration. Hence, when we select the solutions, we use the ones with the most crowd distance to choose one iteration over another from the repository that provides the non-dominated solutions.

M. Angle distribution AD

In addition to the crowd distance, we find the angle between one solution and the other also as a criterion for exploration. We select the solution that has a higher angle between its vector and the two adjacent solutions. This enables more exploration if it is used along with the crowd distance.

N. Pseudocode for Optimization

For showing how the optimisation works, we present the pseudocode of the optimisation in Table IV. The algorithm receives two inputs: the size of swarm or the number of solutions, NSol and the number of iterations It. The algorithm also receives three coefficients: inertia w, constant of best local C_1 and constant of best global C_2. Here it is important to state that the best local or global can be set of non-dominated solutions. We need to select one of them as the leader, the maximum velocity $\mathrm{V}_{\mathrm{max}}$ and the minimum velocity $\mathrm{V}_{\mathrm{min}}$. The output of the pseudocode is the Pareto front which represents set of non-dominated solutions found by the approach. The algorithm starts with the swarm initialization as depicted in line number 11. Next, the algorithm calculates the objective functions of the swarm which is done in the evaluate command in line 12. Afterwards, the algorithm selects the leader as random solution from the non-dominated solution in order to determine the best global which is done in command line number 13. Next, the algorithm initializes the counter of the index of the iteration in line number 15. Next, the main loop of searching starts in line number 16 and continues until line number 26. This loop represents the main searching loop which is responsible of moving the swarm within the searching space in order to find the non-dominated solutions or the Pareto front that will be reported in line number 29. The loop contains an inner loop that runs on the particles starting from line 17 until line 21. It moves each particle in line 17, mutate the particle in line 18, evaluate the particle in line 19, select the best position from the last and the current in line 20, and update the best personal solution of the particle in line 21. Afterwards, the algorithm selects the global leader from the whole swarm in line 23.

TABLE IV. MULTI OBJECTIVE PARTICLE SWARM OPTIMIZATION WITH ANGLE

```
1-   Inputs
2-   f_1,f_2,…f_n          //set of objectives
3-   It                                //maximum number of
     generations
4-   NSol   // size of solutions of swarm
5-   V_max                             //maximum velocity
6-   V_min                             //minimum velocity
7-   w,  C_1,C_2
8-   AngleRes  //the resolution of the angle
9-   Output
10-  PF    Pareto Front   //
11-  Start
12-  Initialize swarm
13-  Evaluate swarm based on f_1,f_2,…f_n
14-  select leaders
15-  t = 0
16-  While t < tmax
17-  For each particle
18-  newParticle=Update Position (particle,Vmin,Vmax)
19-  particle=Mutation(particle)
20-  Evaluate(particle,f1,f2, ...fm)
21-  particle=selectBest(newParticle,particle,AngleRes)
22-  Update pbest
23-  EndFor
24-  Select(leader,gbest)
25-  t++
26-  EndWhile
27-  Report Pareto Front
28-  End
```

The usage of the angle in the new developed MO-PSO-Angle is provided in the calling of procedure of selectBest which put the new particles after moving and the particles before moving in one repository and selects the solutions based on both the crowding distance and the AngleRes in a probabilistic way using pseudocode presented in Table V.

TABLE V. PSEUDOCODE OF SELECTING SOLUTIONS BASED ON THEIR CROWDING DISTANCE AND ANGLE

```
particle=selectBest(newParticle,particle,AngleRes)
Generate random number $r \in (0,1)$
If $r < 0.5$
Selects the solution with the minimum angle density between newParticle and
particle
Else
Select the solution with the maximum crowding distance between newParticle
and particle
End
```

We mean by angle density, the number of solutions inside the angle after decomposing the solution space into angular sectors according to the angle resolution AngleRes. For each solution we define in addition to the crowding distance the angle density which refers to the number of solutions inside the corresponding sector.

IV. Results and Discussions

For evaluating our developed tail-based routing and MO-PSO-Angle, we have implemented both of them and performed an evaluation of simulation of VANETs environment. The optimization was done based on 10 vehicles. The simulation parameters are depicted in Table VI.

TABLE VI. PARAMETERS OF THE SIMULATION

Parameter name	Parameter value
dataPacketLifeTime	10 [sec]
interArrivalTime	6 [sec]
dataPacketGenerationMean	2 packets
routeRequestTimeOut	2 [sec]
routeReplyBufferSize	100[Byte]
timeExp	300
numberOfNodes	10,20,…150
coverageZoneRadius	100

This section provides the MOO optimization results of LAR, tail based Routing, and Scaled Tail based Routing using MO-PSO-Angle. We have performed 10 experiments for optimizing each of the protocols. We show four of them in Pareto front in Fig. 3. The distribution of the results in the Pareto front show that tail-based routing after optimization was able to provide less E2E delay, less inverse of PDR, and less overhead as the solutions represented by green points are gathered at the corner of the Fig. 3.

However, this is not enough to know quantitatively the performance, then we have to provide the MOO measures.

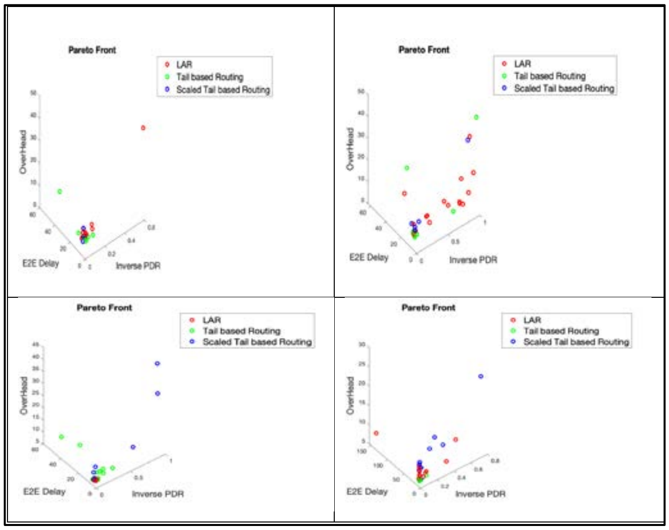


Fig. 3. The Pareto front of LAR, Tail-based Routing and Scaled based Routing.

The visualization of the set coverage in Fig. 4 shows the superiority of tail-based routing over LAR after optimizing, and the superiority of scaled-based routing over LAR after optimizing, however, tail based routing is superior over scaled tail-based routing in terms of obtaining more dominant solutions. This is can interpreted by the fact narrowing the request zone more when we do scaling of it in the scale based tail routing.

In addition to that, we present the results of hyper-volume which shows the spread of the solutions in Fig. 5. We see that tail-based routing has also higher value of hyper-volume. This shows that tail-based routing was able to provide more flexibility of choices to the decision maker for selecting the best point of operation that achieves more domination at the same time in terms of PDR, E2E delay and hyper-volume. In addition, we provide the number of non-dominated solutions for the three protocols after optimization. The results are depicted in Fig. 6.

O. Statistical Evaluation

The last part that was done is the statistical evaluation which is presented in Table VII. We present 10 experiments ranging from 1 until 9. We provide NDS and HV results for each of the experiments. Next, we apply t-test values between the benchmark LAR, our approach tail-based routing and the scaled tail-based routing which is another variant of our approach. We present the statistical t-test values in Table VIII.

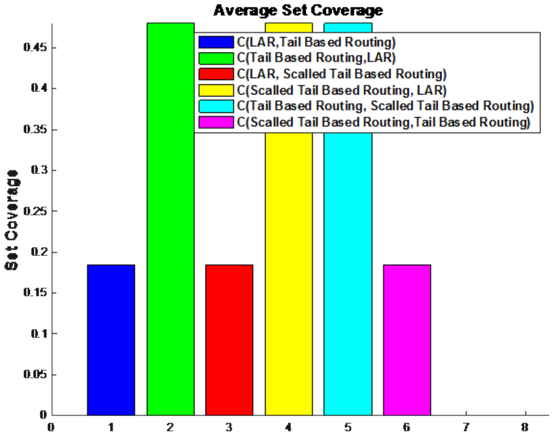


Fig. 4. The Set Coverage of Comparing the Three Protocols: LAR, Reduced Tail based Routing, and Scaled Reduced Tail based Routing.

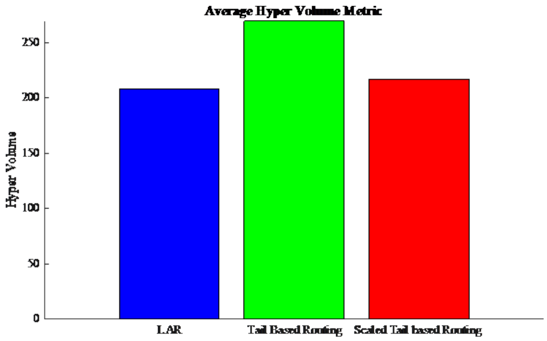


Fig. 5. Hyper-Volume of the Protocols: LAR, tail based Routing and Scaled Tail based Routing.

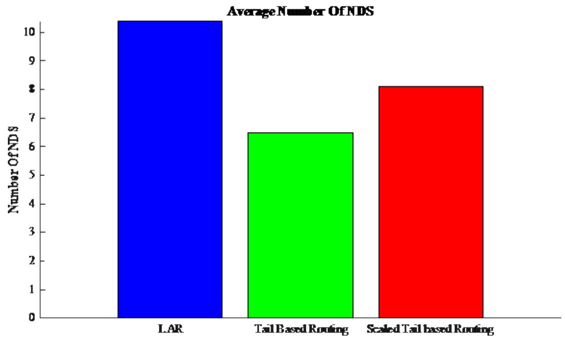


Fig. 6. Number of Non-Dominated Solutions LAR, Tail based Routing and Scaled Tail based Routing.

The results of t-test indicates to superiority in terms of NDS which provides that our developed protocol was competitive from the perspective of NDS considering that is superior in terms of providing more dominating solutions as it is presented in the previous sub-section.

In order to provide the application aspect of the performance of LAR routing and MO-PSO-Angle optimized Reduced Tail based routing, we provide a thorough comparison of the three network evaluation metrics, namely, PDR, E2E delay and overhead for experiments related to the number of nodes which changes from lower value of 30 vehicles up to maximum value of 130 vehicles. Observing Fig. 7, reveals that tail based routing has achieved higher average PDR for all number of nodes that goes from 30 up to 150. This is interpreted by narrowing down the request zone to small rectangular zone named tail that has more potential of delivering the packets to the destination. This also has implied less E2E delay as it is shown in Fig. 8. We notice that tail-based routing has achieved less E2E delay for the whole possible number of vehicles and in the same range. However, for LAR, E2E delay was high when the number of nodes was low, this shows that LAR is not capable of accommodating for the sparsity in the network due to the smaller number of nodes comparing with tail-based routing. The third performance aspect that has been monitored is the overhead that is shown in Fig. 9 which was only less for a smaller number of nodes while it increased when the number of nodes has increased. This is justified by the fact that the optimization was conducted at low number vehicles (only 10 vehicles) and the system was evaluated based the range from 30 until 150. This has led to higher value of coverage radius which causes the high overhead comparing with LAR. However, considering that the overhead itself is not a problem when we have low E2E delay and high PDR, we conclude the superiority of tail based routing over LAR. In order to measure the improvement percentage, we use the formula of

$$IP = \frac{newMeasure - OldMeasure}{oldMeasure}$$

Applying for PDR gives the result of 96% at 10 nodes and gives for E2E delay the percentage of 313%.

TABLE VII. STATICALLY EVALUATION

Exp No	1	2	3	4	5	6	7	8	9
LAR	228.647522	0.16641046	0.57624527	198.983643	58.3471143	0.64042105	371.524098	0.01731366	1151.20476
Tail based Routing	272.888518	25.451955	0	2.36052967	4.66519149	1.14641338	0.11795314	62.9180172	2303.12052
Scaled Tail based Routing	640.749255	72.9114435	271.232878	41.0532603	850.114161	0.37399182	18.8569746	104.647234	169.911114
LAR	17	4	10	5	11	8	12	4	19
Tail based Routing	8	8	2	4	6	5	5	11	9
Scaled Tail based Routing	16	3	12	12	9	4	9	8	5

TABLE VIII. T-TEST COMPARISON BETWEEN OUR TAIL BASED ROUTING AND THE LAR AND SCALED FOR HV AND NDS

		LAR	Scaled tail-based routing
HV	Tail based routing	0.61964725	0.84418149
		LAR	Scaled LAR
NDS	Tail based routing	0.10917176	0.27551771

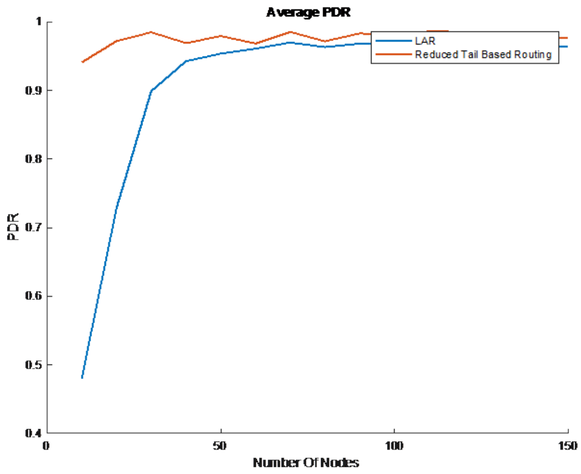


Fig. 7. PDR of LAR and Reduced Tail-based Routing after Optimization using MO-PSO-Angle.

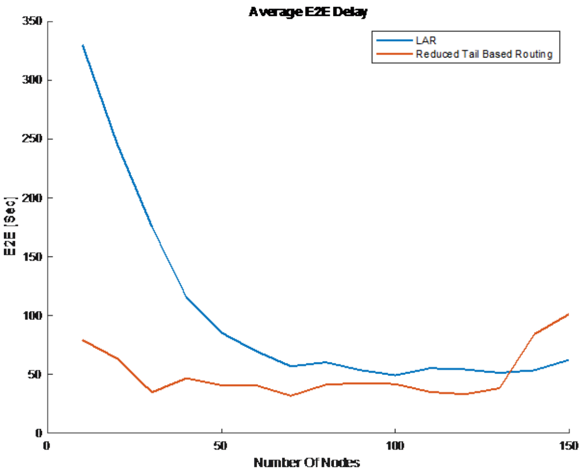


Fig. 8. E2E Delay of LAR and Reduced Tail based Routing after Optimization using MO-PSO-Angle.

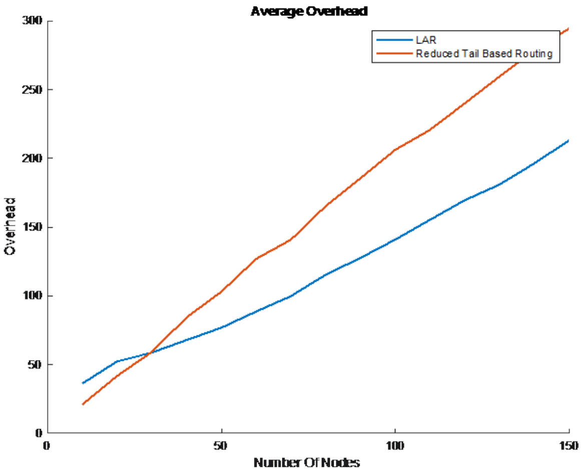


Fig. 9. Overhead of LAR and Reduced Tail-based Routing after Optimization using MO-PSO-Angle.

V. Conclusion and Future Work

This article has tackled the issue of reactive routing protocol in VANETs network. Also, it has provided a new protocol named tail-based routing which is inspired from location aided routing with some modifications. More specifically, it changed the request zone that denotes the region between the source and destination to smaller but more effective region named tail zone which makes it more suitable for highway type of roads. Another aspect that was tackled is the optimization of parameters in the protocols. More specifically, while most researchers conduct single objective optimization with assuming weighted summation is adequate for representing an objective function that combines various needed metrics for optimization in the protocol, we have proposed doing a non-dominated sorting type of optimization. The non-dominated sorting is better to solve the issue of non-convexity that cannot be resolved in the weighted summation single objective type of optimization. Furthermore, in this thesis we have proposed a novel meta-heuristic searching optimization in the framework of multi-objective particle swarm optimization. This is done by providing a more diversity awareness in the searching by using more than one criterion in the selection of solutions that will be added to the repository. This is done by using angle with crowding distance. We apply it in three protocols: original LAR, our tail based routing and scaled tail-based routing. The results have shown the superiority of tail-based routing when it is optimized using MO-PSO-angle in terms of domination with respect to network measures such as PDR, E2E delay and overhead. Lastly, we have performed statistical comparison with secondary metrics to show the overall competitive performance with respect to them, namely, number of non-dominated solutions and hyper-volume. Applying for PDR gives the result of 96% at 10 nodes and gives for E2E delay the percentage of 313%.

Acknowledgment

This research is funded by University Kebangsaan Malaysia (UKM), UKM Grant Code: GGP-2019-023.

References

[1] A. Abbasi, “applied sciences Protocol for VANETs in City Environment,” 2018, doi: 10.3390/app8050687.

[2] A. T. A. Naser Abdali and R. C. Muniyandi, “Optimized model for energy aware location aided routing protocol in MANET,” Int. J. Appl. Eng. Res., vol. 12, no. 14, pp. 4631–4637, 2017.

[3] M. A. Jubair et al., “Bat optimized link state routing protocol for energy-aware mobile ad-hoc networks,” Symmetry (Basel)., vol. 11, no. 11, 2019, doi: 10.3390/sym11111409.

[4] S. Benkerdagh and C. Duvallet, “Cluster-based emergency message dissemination strategy for VANET using V2V communication,” Int. J. Commun. Syst., vol. 32, no. 5, pp. 1–24, 2019, doi: 10.1002/dac.3897.

[5] F. Taha AL-Dhief, R. Chandren Muniyandi, and N. Sabri, “Performance Evaluation of LAR and OLSR Routing Protocols in Forest Fire Detection using Mobile Ad-Hoc Network,” Indian J. Sci. Technol., vol. 9, no. 48, 2016, doi: 10.17485/ijst/2016/v9i48/99556.

[6] G. Zhang, M. Wu, W. Duan, and X. Huang, “Genetic Algorithm Based QoS Perception Routing Protocol for VANETs,” Wirel. Commun. Mob. Comput., vol. 2018, 2018, doi: 10.1155/2018/3897857.

[7] S. Habib, S. Saleem, and K. M. Saqib, “Review on MANET routing protocols and challenges,” Proceeding - 2013 IEEE Student Conf. Res. Dev. SCOReD 2013, no. December, pp. 529–533, 2013, doi: 10.1109/SCOReD.2013.7002647.

[8] N. Harrag, A. Refoufi, and A. Harrag, "New NSGA-II-based OLSR self-organized routing protocol for mobile ad hoc networks," J. Ambient Intell. Humaniz. Comput., vol. 10, no. 4, pp. 1339–1359, 2019, doi: 10.1007/s12652-018-0947-4.

[9] G. Algorithm, "Genetic Algorithm 4.1," doi: 10.1007/978-3-319-93025-1.

[10] D. Wang, D. Tan, and L. Liu, "Particle swarm optimization algorithm : an overview," Soft Comput., vol. 22, no. 2, pp. 387–408, 2018, doi: 10.1007/s00500-016-2474-6.

[11] L. M. R. Rere, M. I. Fanany, and A. M. Arymurthy, "Simulated Annealing Algorithm for Deep Learning," Procedia Comput. Sci., vol. 72, pp. 137–144, 2015, doi: 10.1016/j.procs.2015.12.114.

[12] Q. Zhu, X. Tang, Y. Li, and M. O. Yeboah, "An improved differential-based harmony search algorithm with linear dynamic domain," Knowledge-Based Syst., vol. 187, no. xxxx, p. 104809, 2020, doi: 10.1016/j.knosys.2019.06.017.

[13] H. Shah, R. Ghazali, and N. M. Nawi, "Hybrid ant bee colony algorithm for volcano temperature prediction," Commun. Comput. Inf. Sci., vol. 281 CCIS, pp. 453–465, 2012, doi: 10.1007/978-3-642-28962-0_43.

[14] R. Masadeh, B. A. Mahafzah, and A. Sharieh, "Sea Lion Optimization algorithm," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 5, pp. 388–395, 2019, doi: 10.14569/ijacsa.2019.0100548.

[15] Hui Li and Qingfu Zhang, "Multiobjective Optimization Problems With Complicated Pareto Sets, MOEA/D and NSGA-II," IEEE Trans. Evol. Comput., vol. 13, no. 2, pp. 284–302, 2009, doi: 10.1109/TEVC.2008.925798.

[16] N. Elkhani, R. C. Muniyandi, and G. Zhang, "Multi-objective binary PSO with kernel P system on GPU," Int. J. Comput. Commun. Control, vol. 13, no. 3, pp. 323–336, 2018, doi: 10.15837/ijccc.2018.3.3282.

[17] C. A. Kerrche, F. Ahmad, M. Elhoseny, A. Adnane, Z. Ahmad, and B. Nour, Emerging Technologies for Connected Internet of Vehicles and Intelligent Transportation System Networks, vol. 242. Springer International Publishing, 2020.

[18] C. J. Joshua, R. Duraisamy, and V. Varadarajan, "A Reputation based Weighted Clustering Protocol in VANET: A Multi-objective Firefly Approach," Mob. Networks Appl., vol. 24, no. 4, pp. 1199–1209, 2019, doi: 10.1007/s11036-019-01257-z.

[19] T. O. Fahad and A. A. Ali, "Multiobjective Optimized Routing Protocol for VANETs," Adv. Fuzzy Syst., vol. 2018, 2018, doi: 10.1155/2018/7210253.

[20] S. Yasir, H. Adnan, A. Farhan, K. Fahad, M. Muazzam, and N. Tabassum, "CAMONET : Moth-Flame Optimization (MFO) Based Clustering Algorithm for VANETs," IEEE Access, vol. PP, no. c, p. 1, 2018, doi: 10.1109/ACCESS.2018.2868118.

[21] A. A. Khan et al., "A Hybrid-Fuzzy Logic Guided Genetic Algorithm (H-FLGA) Approach for Resource Optimization in 5G VANETs," IEEE Trans. Veh. Technol., vol. 68, no. 7, pp. 6964–6974, 2019, doi: 10.1109/TVT.2019.2915194.

[22] F. Aadil, K. B. Bajwa, S. Khan, and N. M. Chaudary, "CACONET : Ant Colony Optimization (ACO) Based Clustering Algorithm for VANET," pp. 1–21, 2016, doi: 10.1371/journal.pone.0154080.

[23] N. M. Al-Kharasani, Z. A. Zulkarnain, S. Subramaniam, and Z. M. Hanapi, "An efficient framework model for optimizing routing performance in vanets," Sensors (Switzerland), vol. 18, no. 2, 2018, doi: 10.3390/s18020597.

[24] N. Harrag, A. Refoufi, and A. Harrag, "New NSGA-II-based OLSR self-organized routing protocol for mobile ad hoc networks," J. Ambient Intell. Humaniz. Comput., vol. 0, no. 0, p. 0, 2018, doi: 10.1007/s12652-018-0947-4.

[25] C. E. Sciences, M. Bhagyavathi, and V. Saritha, "LeapFrog and Particle Swarm Optimization based Multipath Routing for VANETs," vol. 9, no. 31, pp. 1525–1533, 2016.

[26] J. Toutouh, "Intelligent OLSR Routing Protocol Optimization for VANETs," no. May, 2012, doi: 10.1109/TVT.2012.2188552.

Enhancement of Fundus Images for Diagnosing Diabetic Retinopathy using B-Spline

Tayba Bashir[1], Khurshid Asghar[2], Mubbashar Saddique[3], Shafiq Hussain[4], Inam Ul Haq[5]
Department of Computer Science, University of Sahiwal, Sahiwal, 57000, Pakistan[1, 4]
Department of Computer Science, University of Okara, Okara, 56300, Pakistan[2, 3, 5]

***Abstract*—Medical images, such as CT scan, MRI, X-ray, mammography and fundus are commonly used in medical diagnosis process and helpful to improve diagnose of disease in a better way and reduces the chances of ambiguous perceptions. Medical images are mostly available in low contrast, brightness and noisy form due to camera/ radio waves intrinsic properties while capturing, which disrupts the diagnosis process using medical images. Enhancement of these images can improve the diagnosis process. The proposed enhancement technique of fundus images is based on the B-spline interpolation, in which intensity transformation curve is based on the control points of the curve. Messidore and Drive datasets of Diabetic Retinopathy (DR) are used to evaluate the proposed enhancement technique. Results shows that the fundus images have reasonable visual and quantitative enhancement when performed comparison with recent techniques. Results are of evidence that the proposed approach has substantial outcome and preserves important information of fundus images by lowering noise.**

Keywords*—*B-spline; medical images enhancement; fundus images; diabetic retinopathy; interpolation

I. Introduction

Contrast enhancement of images is a wide area in digital image processing. The contrast in images make it better to visually analyze and understand the shape of the object clearly. The contrast of images depends on environmental effects like weak lenses of capturing device, low light due to fog and clouds, unprofessional photographer. The images captured under these circumstances have contrast alteration, high noise, and color vanishing [1].

Intensity enhancement makes the objects and background separate to attain the hidden detail [2]. Digital images have many applications in every field of life like medical, security, forensic, education, satellites, recognition, and pattern detection. Contrast enhancement techniques have two types, direct enhancement in which we use fuzzy logic, adaptive neighborhood contrast enhancement and segmentation. While the indirect contrast enhancement is based on histogram changes of image like histogram equalization, contrast limited adaptive histogram equalization and recursive mean separate histogram equalization [3]. Digital image processing is popular for two areas. One is the images information for human use and diagnosis. This is helpful to treat a patient with early diagnosis. The second part is based on processing, storing, altering for machine learning like bio-sensing and identifying.

Medical images enhancement is one the biggest need of the human body tissues to diagnose, monitoring and interpretation of diseases. Most of the medical images are in gray scan like CT scan, X-ray and MRI. The x-rays are frequently used for image elevated compactness like bone. The Computed tomography and Magnetic Resonance imaging used for low density tissues [4]. Enhancement in the medical field is useful to identify diseases like tumors, liver distortion, heart breakage, bone fracture, muscular pain and tissues pull etc. In medical images, a general problem is to capture images in motion form for real time assessment. The blur and deformed images are hard to understand and assessed correctly. The edge detection and organ identification are the most common part of preprocessing technique. There are so many techniques that help to enhance edges and organs in medical images. Medical images have darkness in objects and borders of these objects have noise, which is a basic factor to create discontinuities in images [5]. We have multiple techniques to enhance intensity of medical images like Image Negative, Power Law, Logarithmic Transformation, Histogram Equalization (HE), Bezier Curve for Contrast Enhancement (BCCE), Bezier curve using Fourier Transformation etc. There are many techniques that are studied to enhance image quality and accuracy without loss of information. To increase the low intensity of image the geometric Multiscale differential operators' technique was introduced [6] and it helped to improve edges and corners with returning the best mammographic result.

The medical image enhancement is basically depending on the Intensity transformation. The Breast MRI result observed using the contrast enhancements using fuzzy type1 and fuzzy type2 provides brightness of image. The colorful medical images are needed to enhance with better pixel results. The interactive color image segmentation is performed with the combination of Bezier curve, which give a balanced intensity of images. Medical images have many side effects one of the majors is radiation which is harmful for the human body. Hence, image enhancement enhances blur image and reduces noise of thermal images. Thus, breast cancer images are enhanced by applying reducing radiation and advanced electric equipment [7]. Discrete wavelet transformation is used for image compression, de-noising, segmentation etc. The discrete wavelet provides great results with coefficient domain [8]. The cubic spline shows better results with thresholding method and curve fitting technique. The piecewise polynomial is converted into small segments and enhances image [9].

The proposed technique is B-spline based intensity transformation of dark images. The results generated by proposed method of image enhancement are compared both visually and quantitatively using entropy, peak signal to noise ratio, contrast to noise ratio and structure similarity index image enhancement evaluation measures [2].

A. *Diabetic Retinopathy (DR)*

Diabetes has many serious obstacles but diabetic retinopathy such severe impediment which requires consideration at a very beginning level [10]. The diabetic retinopathy is increasing eye disease that intensifies with time source of permanent harm to retina and eventually turns into sightlessness. The recent studies show that 75% of patients have elevated glucose levels for twenty years are identified with diabetic retinopathy. Blindness is a major cause in the United States of America suffers adults due to diabetics [11]. The blindness risk in diabetic patients is 20% more than normal persons. The identification of retinopathy in patients is a big issue these days. India also suffers to deal with this problem as it is on sixth number in world countries who have diabetic patients [12]. It is the need of time to improve the diagnostic accuracy in ophthalmology. It is difficult to identify the retinoic mellitus by small observation. It is a basic step to identify the internal blood leakage, d-shape of small veins of eye for a doctor. If caught at an early stage a patient is recovered early or may be treated with a lens operation. Identification of diabetic retinopathy at an early stage of disease rescues it from loss of sight and reoperation. Most of the images are captured by oscilloscopes and it is not clear due to environmental effects, camera lens or may be the noise that affects images badly. Due to all these reasons, a new algorithm is trying to develop that enhances image without loss of data and returns great results to deal with retinopathy disease easily at an early stage. A comparison of state-of-the-art image enhancement techniques is shown in Table I.

State-of-the art image enhancement techniques like histogram equalization (HE), adaptive histogram equalization (AHE), contrast limited adaptive histogram equalization (CLAHE), exposure-based sub image histogram equalization (ESIHE) in histogram family and others like interpolation, modulation, classification and segmentation are commonly used to enhance medical images. The outcomes are not satisfied for multiple images due to enhanced contrast with increasing noise. Some techniques are over enhanced contrast ratios due to modification of unnecessary areas; some enhance contrast at the rate of loss of image information. Similarly, enhancement of medical images is a difficult task with keeping all the information and removing noise with increasing contrast of necessary areas and edges of image. The focus of this research is to develop a robust method which enhances DR images.

Rest of the paper is organized as follows: Section II describes the proposed methodology. Implementation and results are presented in Section III. Section IV is about conclusion and future work.

TABLE I. STATE OF THE ART COMPARISON

Reference	**Approach**	**Pros**	**Cons**
Zhao et al. [4]	Luminance and gradient modulation	Reduced dynamic range of luminance level using gradient computation. Enhancement is done at local level instead of globally.	Not clear the image edges detail.
Asghar et al. [2]	Bezier Curve using Fourier Transformation	Automatic reorganization of the type of input image. Enhanced images by modifying frequency domain.	Bezier curve have statics behavior, this may cause loss of information.
Akila et al. [3]	Indirect contrast enhancement	Contrast enhancement using Histogram and its types. Better results for low contrast images.	Detail of images can’t be viewed clearly.
Du-Yih Tsai et al. [8]	Wavelet coefficient mapping functions	Convert image non-liner coefficient into discrete wavelet coefficient. Using fast Fourier transformation for low resolution image enhancement.	Improvement in graph cut with new edge weight.
Yu-Ping et al. [6]	Multiscale differential operators.	High frequency features like edges are characterized Improved classification of chromosomes and qualitative contrasts.	Increase memory requirement due to completion of multiple process.
Thierry Blu et al. [13]	Fractional spline wavelet transformation	Its return positivity and compact support Whitened noise and enhanced low contrast area.	It works with alpha elements only.
Umar Talha et al. [14]	Hermit based interpolation	Under sampled reconstruction of image angel’s help to reduce radiations. Filtered back projection returns visible artifacts in the output image and has 94% accuracy.	Spatial filtering-based post processing is needed.
Najid et al. [15]	Classification and segmentation	Breast region segmented into small areas for exact diagnosis with smooth interpolation.	Loss of some information due to inappropriate detail.
Lehmann et al. [16]	B-spline using Fourier transformation	Interpolation performs using down sampling to get improved results. Extract luminance and represent naturalness.	The dark images need to reconstruct before implementation.
Xueyang et al. [17]	Fusion methods to enhance luminance	Blended of multiple enhancement approaches return great results. It can also use as post processing for any type enhancement.	Haze removal is required for better outcomes.

II. Proposed Methodology

To enhance medical images B-spline interpolation-based technique is proposed in this paper. It's based on three steps. In the first step the RGB are converted into YCbCr color space and extract 'Y' channel. In the second step apply B-spline transformation on the extracted Y channel. In the third step merge the enhanced Y channel with Cb and Cr and get resulted enhanced output image.

A. B-Spline

The proposed image enhancement technique is based on B-spline curve which is used to enhance medical images dynamically with securing all information by intensity transformation using. The B-spline curve is based on control points; value of a curve is 0 to 255. The P1 and P2 points are changed according to the distance (D) is measured to enhance an image contract. B-spline is looking like Bezier curve accepts its dynamic nature as shown in Fig. 1.

The blending function use control points to make curve. The n + 1 control points $P0, P1, \dots, Pn$ and a knot vector T , the B-spline curve of degree p defined by these control points. Polynomial b-spline has similar properties with Bezier curve. The B-spline interpolation has interval of point $[a, b]$. The degree of curve is denoted as p, so curve ≤ p. The knot vector of B-spline curve is denoted $T = \{t_0, \dots, t_m\}$ with value of parameter taken as $t_j \leq t_i + 1, i = 0, \dots, m - 1$. Control points are taken as bo ... $b_{n.}$ B-spline basic function is defined as:

$$P(u) = \sum_{i=0}^{n} p_i N_{i,D(u)} \tag{1}$$

The number of control points in B-spline are determined by D-1. The basic function defines itself repeatedly. B-spline function conditions are shown in

$$N_{i,1}(u) = 1 \Rightarrow t_i \leq u \angle t_{i+1} \tag{2}$$

$$= 0, otherwise$$

The polynomial function of n^{th} degree is described by equation (3).

$$N_{i,d}(u) = \frac{(u-t_i)N_{i,D-1}(u)}{t_{i+D-1}-t_i}, \frac{(t_{i+d}-u)N_{i+1,D-1}(u)}{t_{i+D,}-t_{i+1}} \tag{3}$$

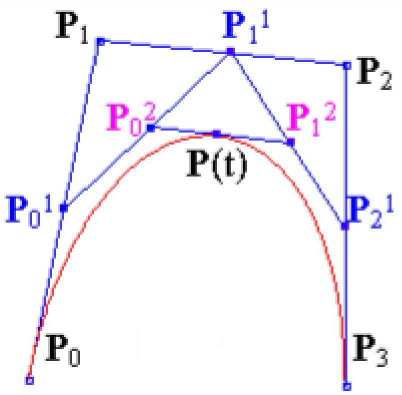


Fig. 1. B-Spline Curve.

The B-spline is similar to Bezier curve, but it has to pass out all the points known as interpolation; therefore, it yields better intensity transformation curve.

B. Algorithm

1) Select the input image

2) Covert to YCbCr color space

3) Calculate histogram of Y channel of image to categorize the image as dark, bright, low contrast, high contrast and backlight.

4) Calculate movement distance (D) using equation (4)

$D = (2^{(N-1)}) \times Tc$, (4) where N is the total number of pixels with frequencies greater than zero and αTc is calculated) with contrast ratio formula given in equation (2):

$$Tc = \frac{N}{(Lmax-Lmin+1)} \tag{5}$$

where, $Lmax$ and $Lmin$ are pixels with maximum and minimum luminance values respectively.

5) Apply equation (1) and calculate new position of control points $P1$ and $P2$, according to the calculated movement distance D, the end points $P0$ and $P3$ remain fixed at (0, 0) and (255, 255), respectively. Calculation of middle control points for all possible types of images are illustrated in equations (6), (7) and (8), respectively.

a) for over dark images:

$$P1 = P0 + \left(\frac{D}{2}, 0\right) \text{and } P2 = P3 - \left(\frac{D}{2}, 0\right) \tag{6}$$

b) for over bright and back light images:

$$P1 = P0 + \left(\frac{D}{2}, 0\right) \text{and } P2 = P3 - \left(\frac{D}{2}, D\right) \tag{7}$$

c) for over low contrast images:

$$P1 = P0 + \left(\frac{D}{2}, 0\right) \text{and } P2 = P3 + (1, 1) \tag{8}$$

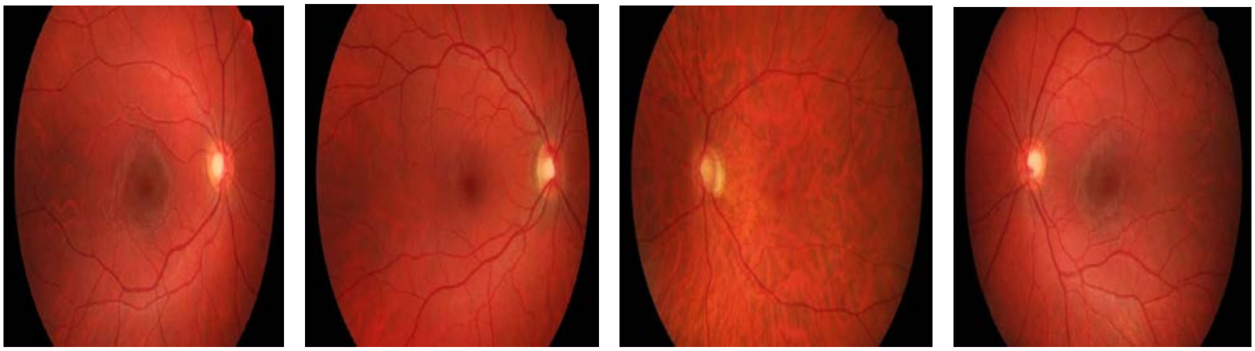
6) Perform intensity transformation using B-spline curve produced in step 5 on Y channel of the images to compute new histograms for mapping of new luminance values.

7) Finally, transformed Y is concatenated with Cb and Cr to get output image.

This procedure provides the enhanced fundus images without loss of information.

C. Dataset Description

The algorithm results tested on fundus images selected from two publicly available dataset i.e., DRIVE and MESSIDOR. The details are given under: Messidor is a fundus images dataset of 1200 lossless images. These images take at 45° with FOV camera in different directions [18]. The other online available dataset named as Drive consists of 40 lossless images in colored format with clinical expert remarks [12]. These images are categories into two parts: training and testing. Image resolution is taken as 768 × 584 pixels with a field of view of 45°. The four DR images (see Fig. 2) are taken to show the results.



IMG 1 IMG 2 IMG 3 IMG 4

Fig. 2. Sample DR Images for Enhancement.

III. Implementation and Results

The proposed technique of medical image enhancement based on b-spline interpolation is implemented on a dataset of diabetic retinopathy. Special digital fundus cameras are used to obtain retinal images [19]. Diabetes mellitus has become a serious problem for health and study shows that it is the main source of deaths in diabetic's patients [11]. The quality of the fundus images is affected with various factors like, environmental conditions in which it is captured, brightness and angle of image, the lens of camera and the weather condition. Therefore, diabetic mellitus results are not good for segmentation and classification due to low brightness and contrast of captured images [19].

There are a few approaches already proposed to enhance the retinopathy images, like histogram equalization, Bezier curve, Hermite spline, etc. In this paper we made a comparison of our technique with a proposed technique named as diabetic retinopathy enhancement technique for fundus images [20]. To evaluate the sustainability of proposed techniques the experiments observed with histogram-oriented methods and statistical analysis [21]. The evaluation measures structure similarity index, entropy, peak signal to noise ratio and contrast to noise ratio are performed. The experiments are performed with MATLAB R2018a on core i5 64bit operating system.

A. *Histogram*

A histogram is the rectangular frequency of data items based on numerical form with equal size. It helps to perform statistical analysis. The dependent variables are along the vertical axis. While the independents are on horizontal axis. The result is shown in the form of a bar line in the graph [22]. The bar lines raised according to the data points given to them. In the given table the histogram of real and enhanced images is shown clearly.

The comparison of histograms of images showed that the frequencies of original images are high in starting points. On the other hand, the frequencies of enhanced images are medium in length and balanced in range [23]. The original and enhanced image histogram is shown in Fig. 2(c) and (d) respectively. It is cleared from the figure that the histogram equalization of enhanced images is balanced and return good results because the frequencies of images are distributed.

B. *Curve Lines*

A straight line in one direction is known as a line. The line that has more than one point and different direction is known as the curve line. The B-spline curve is important for contrast enhancement of images while the control points decide the movement of curve line [4]. The curve line of enhanced images is shown in Fig. 3(e). The original image has straight curve, the upward curve known as enhanced with brightness while the downward curve decreased the brightness of image.

C. *Performance Evaluation*

It's the statistical and mathematical analysis of enhancement technique to figure out the exact calculation of image. The difference of original image value and enhanced image value is known as enhancement. The evaluation is compared with the hybrid proposed fundus image enhancement using histogram equalization, contrast limited adaptive histogram equalization and exposure-based sub image histogram equalization (HE). The function of adaptive histogram equalization (AHE) is to compute different images histogram and reallocate intensities to enhance contrast of image [24]. Contrast limited adaptive histogram equalization (CLAHE) is good for low contrast images; it enhances its low contrast images using adjacent neighbor pixels to improve contrast. The contrast of the image is based on clip and slope of histogram height directly. If height is increased the contrasts also increase. It overcomes noise amplification as well [25]. Exposure based sub image histogram equalization (ESIHE) is used to enhance contrast and height of histogram. It gives the highest value of histogram height then already defined threshold values [26]. We performed comparison between HE, AHE, CLASHE, ESIHE techniques and proposed B-spline technique results. Please see quantitative results in Tables I, II, III and IV against each evaluation measure.

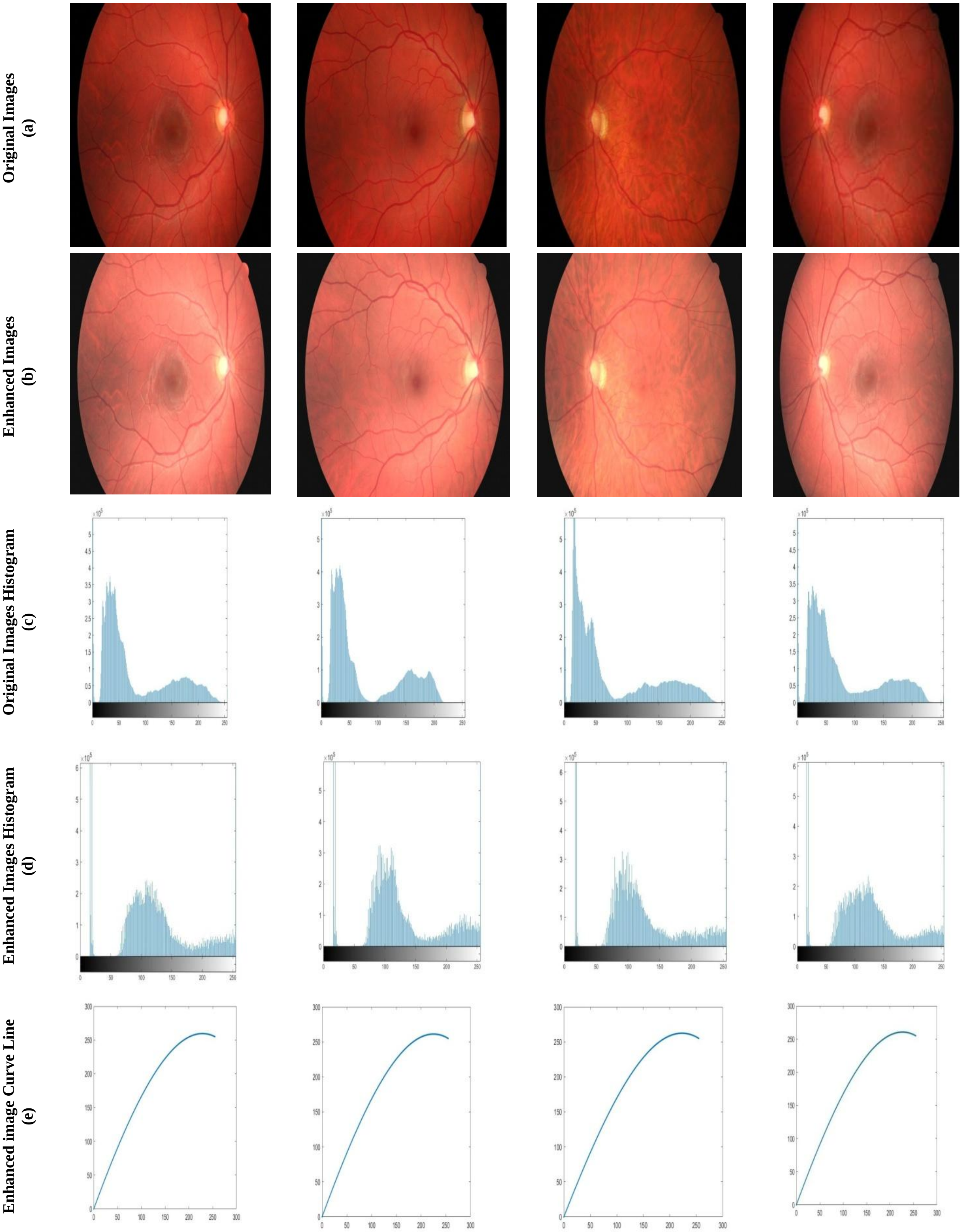


Fig. 3. (a) Sample DR Original Images (b) Enhanced Images (c) Hisgrogram of Orignial Images (d) Hisgrogram of Enhanced Iamges (e) Curve Line of Enhaced Iamges.

D. Evaluation Measures

Following evaluation measures are used for quantitative evaluation of the proposed approach.

1) Structure Similarity Index Measure (SSIM): It is the measure that identifies the best results of original and enhanced images as described in equation (9).

$$SSIM(\mathrm{x,y}) = \frac{(2\mu_x\,\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2\,\mu_y^2+c_1)(\sigma_x^2\sigma_y^2+c_2)} \tag{9}$$

The maximum similarity index value is 1. Image which has value near to 1 is called highest SSIM.

The SSIM value is enhanced if it is near to 1. After performing the experiments as shown in Table II, it is observed that the outcome value of HE, AHE and CLAHE not considerably significant to structure similarity index measure. But ESHIE nearly performed with enhancement of contrast in terms of SSIM. While the proposed technique is excellent to increase the contrast of an image. Structure similarity index of the dataset is starting from 0.9913 to 0.9917. The contrast of image enhances the best level compared with other techniques.

2) Entropy: Entropy [12] provides the average of content information of an image. Its large value considers good results. But in some cases where loss of image information occurs it’s not measured correctly [27].

$$E(P) = \sum_{k-0}^{L-1} P_k \times log\ P_k \tag{10}$$

Entropy value of the given images is calculated as shown in Table III and found that the HE, AHE, EHIE values of entropy are less then original values. The CLAHE values are high with amplification of noise.

3) Peak Signal to Noise Ratio (PSNR): To calculate the worth of reconstructed image peak values of signal are identified. The more value of peak signals the stronger image was returned. Before measuring PSNR value the mean square error is identified.

$$MSE = \frac{\sum_{i=1}^{M}\sum_{(j=1)}^{N}[X(i,j)-Y(i,j)]^2}{(M\times N)} \tag{11}$$

$$SNR = \frac{10\,log^{10}(peakvalue)^2}{MSE} \tag{12}$$

The value of PSNR with all techniques is calculated in Table IV. The CLAHE is showing very low PSNR due to high noise ratio. The HE, AHE and ESHIE techniques provide enhancement of image contrast ratio. The proposed technique provides drastically changed results of PSNR because it’s high contrast of image and low noise. The starting PSNR value is 60.24 to 60.45 on the chosen dataset. This result returns a very high contrast enhancement of images.

4) Contrast to Noise Ratio: To measure the value of noise and contrast in an output is known as the contrast to noise ratio. Now a days, it is widely used to measure medical images enhancement.

$$CNR = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{13}$$

CNR is another parameter to measure noise and contrast of image. The CLAHE provides the better result in histogram family that showing enhancement of images, proposed technique also returns good results on selected images. Table V shows the CNR comparison of different contrast techniques.

The enhancement returns better result of contrast. Results fluctuate according to the image CNR. The CNR of the image is changed from 6.64 to 13.09. It means the standard deviation and expectation value of images are enhanced greatly.

TABLE II. COMPARISON USING EVALUATION MEASURE SSIM

SSIM	TECHNIQUES	IMG1	IMG2	IMG3	IMG4
	HE	0.6705	0.6599	0.6356	0.6546
	AHE	0.7904	0.7832	0.8041	0.7888
	CLAHE	0.1498	0.1495	0.1497	0.1502
	ESHIE	0.9391	0.932	0.9367	0.9338
	Proposed	**0.9919**	**0.9913**	**0.9914**	**0.9917**

TABLE III. COMPARISON USING EVALUATION MEASURE ENTROPY

Entropy	Techniques	IMG1	IMG2	IMG3	IMG4
	Original	6.8979	6.6775	6.8096	6.9268
	HE	5.6178	5.6363	5.5438	5.6148
	AHE	6.7072	6.5338	6.5377	6.7421
	CLAHE	7.1793	7.0444	7.1434	7.1858
	ESHIE	6.7896	6.5811	6.7133	6.8135
	Proposed	**6.6831**	**6.6400**	**6.5900**	**6.7300**

TABLE IV. COMPARISON USING EVALUATION MEASURE PSNR

PSNR	Techniques	IMG1	IMG2	IMG3	IMG4
	HE	11.8856	11.3802	11.1596	11.6942
	AHE	21.3623	21.2975	21.8054	20.6336
	CLAHE	08.4919	08.8072	8.7615	08.7041
	ESHIE	28.0744	26.6361	26.9853	26.9576
	Proposed	**60.4500**	**60.2400**	**60.2500**	**60.3800**

TABLE V. COMPARISON USING EVALUATION MEASURE CNR

CNR	Techniques	IMG1	IMG2	IMG3	IMG4
	Original	12.6900	11.790	6.6400	15.7900
	HE	12.77600	11.8390	13.3737	15.8677
	AHE	17.3575	20.2954	17.7085	18.3116
	CLAHE	17.6047	22.5876	13.4085	20.1139
	ESHIE	12.0926	13.4324	14.2321	16.8950
	Proposed	**12.8900**	**12.0100**	**13.0900**	**17.4000**

IV. Conclusion and Future Work

Enhancement of medical images is an important task while diagnosis of diseases using images. In this paper the diabetic based retinopathy images are enhanced by applying B-spline. There are many famous methods to enhance diabetic retinopathy images in the, but they are not robust with the results due to noise, loss of information. The proposed technique is used to enhance the contrast of the images automatically by using B-spline, which passes through every point of curve and provides the best and amplified enhancement of contrast. The B-spline curve is dynamic in nature and works like a Bezier curve and returns enhanced images without loss of any information. The curve is computed based on control points the resultant image luminance and contrast is then transformed according to the distance (D) measured by control points. Messidore and Drive dataset of fundus images are used to evaluate the proposed techniques. By performing experiments, the results are visually and statistically enhanced the contrast of the fundus images, while the noise is also removed. In future we will work on enhancement of CT scan and MRI images.

References

[1] Y. Chang, C. Jung, P. Ke et al., “Automatic contrast-limited adaptive histogram equalization with dual gamma correction,” IEEE Access, vol. 6, pp. 11782-11792, 2018.

[2] K. Asghar, G. Gilanie, M. Saddique et al., “Automatic enhancement of digital images using cubic Bézier curve and Fourier transformation,” Malaysian Journal of Computer Science, vol. 30, no. 4, pp. 300-310, 2017.

[3] K. Akila, L. Jayashree, and A. Vasuki, “Mammographic image enhancement using indirect contrast enhancement techniques–a comparative study,” Procedia Computer Science, vol. 47, pp. 255-261, 2015.

[4] [4] C. Zhao, Z. Wang, H. Li et al., “A new approach for medical image enhancement based on luminance-level modulation and gradient modulation,” Biomedical Signal Processing and Control, vol. 48, pp. 189-196, 2019.

[5] R. Firoz, M. S. Ali, M. N. U. Khan et al., “Medical image enhancement using morphological transformation,” Journal of Data Analysis and Information Processing, vol. 4, no. 01, pp. 1, 2016.

[6] Y.-P. Wang, Q. Wu, K. R. Castleman et al., “Chromosome image enhancement using multiscale differential operators,” IEEE transactions on medical imaging, vol. 22, no. 5, pp. 685-693, 2003.

[7] A. A. Wahab, M. I. M. Salim, J. Yunus et al., “Comparative Evaluation of Medical Thermal Image Enhancement Techniques for Breast Cancer Detection,” Journal of Engineering and Technological Sciences, vol. 50, no. 1, pp. 40-52, 2018.

[8] D.-Y. Tsai, and Y. Lee, "A method of medical image enhancement using wavelet-coefficient mapping functions." pp. 1091-1094.

[9] C. S. Tutika, C. Vallapaneni, B. KP et al., “Cubic spline interpolation segmenting over conventional segmentation procedures: application and advantages,” arXiv preprint arXiv:1803.04621, 2018.

[10] S. Dutta, B. C. Manideep, S. M. Basha et al., “Classification of diabetic retinopathy images by using deep learning models,” International Journal of Grid and Distributed Computing, vol. 11, no. 1, pp. 89-106, 2018.

[11] T. Walter, J.-C. Klein, P. Massin et al., “A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina,” IEEE transactions on medical imaging, vol. 21, no. 10, pp. 1236-1243, 2002.

[12] M. D. Abràmoff, Y. Lou, A. Erginay et al., “Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning,” Investigative ophthalmology & visual science, vol. 57, no. 13, pp. 5200-5206, 2016.

[13] T. Blu, and M. Unser, "The fractional spline wavelet transform: definition end implementation." pp. 512-515.

[14] S. U. Talha, T. Mairaj, W. Khan et al., "Interpolation based enhancement of sparse-view projection data for low dose FBP reconstruction." pp. 1-6.

[15] N. Al-Najdawi, M. Biltawi, and S. Tedmori, “Mammogram image visual enhancement, mass segmentation and classification,” Applied Soft Computing, vol. 35, pp. 175-185, 2015.

[16] T. M. Lehmann, C. Gonner, and K. Spitzer, “Addendum: B-spline interpolation in medical image processing,” IEEE transactions on medical imaging, vol. 20, no. 7, pp. 660-665, 2001.

[17] X. Fu, D. Zeng, Y. Huang et al., “A fusion-based enhancing method for weakly illuminated images,” Signal Processing, vol. 129, pp. 82-96, 2016.

[18] E. Decencière, X. Zhang, G. Cazuguel et al., “Feedback on a publicly distributed image database: the Messidor database,” Image Analysis & Stereology, vol. 33, no. 3, pp. 231-234, 2014.

[19] J. I. Orlando, E. Prokofyeva, M. del Fresno et al., “An ensemble deep learning-based approach for red lesion detection in fundus images,” Computer methods and programs in biomedicine, vol. 153, pp. 115-127, 2018.

[20] I. Qureshi, J. Ma, and Q. Abbas, “Recent Development on Detection Methods for the Diagnosis of Diabetic Retinopathy,” Symmetry, vol. 11, no. 6, pp. 749, 2019.

[21] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez et al., "Deep retinal image understanding." pp. 140-148.

[22] S.-W. Lee, and J. K. Paik, "Image interpolation using adaptive fast B-spline filtering." pp. 177-180.

[23] C. Y. Wong, G. Jiang, M. A. Rahman et al., “Histogram equalization and optimal profile compression-based approach for colour image enhancement,” Journal of Visual Communication and Image Representation, vol. 38, pp. 802-813, 2016.

[24] O. Patel, Y. P. Maravi, and S. Sharma, “A comparative study of histogram equalization based image enhancement techniques for brightness preservation and contrast enhancement,” arXiv preprint arXiv:1311.4033, 2013.

[25] B. Singh, and S. Patel, “Efficient medical image enhancement using CLAHE enhancement and wavelet fusion,” International Journal of Computer Applications, vol. 167, no. 5, pp. 0975-8887, 2017.

[26] I. Qureshi, J. Ma, and K. Shaheed, “A hybrid proposed fundus image enhancement framework for diabetic retinopathy,” Algorithms, vol. 12, no. 1, pp. 14, 2019.

[27] A. Karci, “Fractional order entropy: New perspectives,” Optik, vol. 127, no. 20, pp. 9172-9177, 2016

Developing Skills of Cloud Computing to Promote Knowledge in Saudi Arabian Students

Ahmed Sadik[1], Mohammed Albahiri[2]*
King Khalid University
Guraiger, Abha, Saudi Arabia

***Abstract*—The present study aims to develop the skills of the cloud computing applications and the knowledge economy among the university students by designing a participatory electronic learning environment. A sample was chosen from the students of the "General Diploma" in the Faculty of Education, King Khalid University. This sample was divided into two groups; experimental group that comprised of 15 students trained through the participatory e-learning environment; whereas, the control group comprised of 17 students, who were trained through the Blackboard Learning Management System. Skills for cloud computing applications and a knowledge economy skills scale were developed. Kolmogorov-Smirnov Mann was used for identifying the normality test of variables. Whitney test and Spearman correlation test were used to analyze the results, which indicated that the design of a participatory e-learning environment based on the theory of communication contributed to improve the skills level of cloud computing applications and knowledge economy skills. The results showed that participatory e-learning environment based on the theory of communication significantly contributes towards improving the skills level of cloud computing applications and knowledge economy skills among the students from Saudi Arabian universities. Moreover, future studies need to focus on blueprint in the context of the educational system of Saudi Arabia.**

Keywords—Cloud computing applications; e-learning environment; higher education; knowledge economy

I. Introduction

Cloud computing has become an important technical trend that can reshape the IT operations (information Technology) and IT market operations. The students use a variety of devices, including desktops, laptops, smartphones, digital access devices, storage space, and online application development platforms through the services provided by cloud computing providers. The advantages of cloud computing include cost savings, high availability, and ease of access [1]. Author in [2] recommended the need for educational institutions to quickly turn to the use of cloud computing in the educational process as it does not constitute a cost or additional physical burden on the educational institution. According to [3], Google Drive is one of the cloud computing services that help to overcome the problems of collective learning. The problems of collective learning include the adoption of a learner in a group over others, the lack of participation of some members of the group, and the lack of commitment of some members for setting goals.

The knowledge economy revolves around knowledge acquisition and is used to improve life in all areas using the human mind and the use of scientific research. Skills of the knowledge economy are defined as a set of behaviors and activities that enable the learner to deal accurately and skillfully with knowledge to employ them effectively in the fields of science and life. The most important of these skills include; critical thinking skills, effective handling, problem solving, decision making, and creativity and innovation [4]. The benefit of the knowledge economy depends on how it becomes a learning economy using modern technology and techniques to communicate with others to spread ideas and innovation, rather than gaining global knowledge. Learners in the learning economy can create wealth that is equal with their ability to learn and engage in innovation and innovation with others [5].

The leader of the communication school confirmed that the process of learning takes place in different ways including; modern information and communication technologies. These technologies include; computers, multimedia-based software, websites, e-mail, email lists Blogs, and Wiki Virtual social networks. The communicative theory is like constructional theory that emphasizes social learning, allows learners to communicate, and interacts with each other or while learning. This theory emphasizes the role of participatory technology to acquire knowledge, skills, and digital values [6]. The design of the participatory e-learning environment considering the "communicative" theory of the above learning strategies combines the formal learning provided by the teacher so that all students benefit from their experience. The formal curriculum serves as the foundation for all students and provides participatory learning spaces through Web tools (2.0) (Blogging, wiki, Facebook, etc.). This helps the students to interact and share the information they have searched for, under the guidance of the teacher, along with directing and preparing the output of these participations.

Cloud computing provides useful support in the field of education, along with its significant use in infrastructure, communication, software applications, data storage, and platform system [7]. According to Arpaci [8], adopting cloud computing in education is capable of enhancing knowledge management. Further Li [9] summarized that cloud computing is beneficial in higher education as it enhances the level of education modernization, reduces costs, and helps in achieving sharing of educational resources. Huang [10] stated that benefits of cloud computing are observed through ease of access to educational content, training facilities for scientific innovation, collaboration and knowledge building, and providing support to students and teachers for easy facilitation

*Corresponding Author

and exchange of knowledge. The students are offered to share the teaching and learning materials, scientific and research articles through Google Docs or DropBox [11].

It is important to integrate technology in general, and cloud computing in universities, particularly. It is known that constructivist strategies for cognitive learning determines the efficiency of a classroom setting [12]. Certain educational environments are introduced by cloud computing based on teaching and learning practices that take place in the framework of social interaction and cooperation to build knowledge. There is need to focus on the role of cloud computing to enhance teaching and learning practices in Saudi Arabia; although, studies have shown the significance of cloud computing in higher education system in other regions.

In the similar context, this study aims to develop the skills of students toward cloud computing skills and their knowledge economy enrolled in the General Diploma in the Faculty of Education, King Khalid University. The aim has been fulfilled through the development of a participatory electronic learning environment based on the theory of "communication Connectivism". The study is significant as it would help in proposing training programs based on the design of an electronic environment that can benefit teachers in various scientific disciplines to develop their skills in the field of cloud computing. It would also help in designing platform-based electronic environment (Pbworks) that can benefit specialists in different subjects to develop their skills for the services of cloud computing and other various applications. Moreover, it would also instruct the curriculum developers towards the use of technological innovations in teaching and learning.

II. Material and Methods

A. Study Design

The study has conducted an experiment for determining the effectiveness of the proposed cloud-based learning environment in order to improve practices and learning performance in a higher education on product design. A quasi-experimental study was conducted on a product design course in a computer-aided classroom at a Saudi Arabian University for examining the effectiveness of the proposed cloud-based learning environment.

B. Study Sample

The study investigates the effectiveness of developing an electronic environment based on the theory of communication to develop the skills of cloud computing and the knowledge economy of students. The recruited participants were divided in two groups; one was experimental and the other was a control. The experimental group was trained on the proposed program based on the theory of communication through the platform Pbworks; while, the control group used the learning management system environment Blackboard available at the University. The tools of the study were applied in a tribal manner to ensure the equivalence of the two groups as shown in Table I. Table II shows that the value of the test Kolmogorov-Smirnov (0.16), (0.23), (0.18), was less than (0.05). This confirms the non-normal distribution of the data.

TABLE I. Degree of Equivalence between Students of the Two Study Groups

Group	**Tool**	**Number**	**Variance***	**Calculated value of (P)**
Experimental	Observation card	15	6.96	1.29
Control		17	5.37	
Experimental	Knowledge Economy Scale	15	8.98	1.34
Control		17	6.69	

TABLE II. Kolmogorov-Smirnov Test to Show Distribution of Data

Group	**Numb er**	**Tool**	**Val ue**	**Functi on level**	**Statistic al significa nce**	**Distribut ion**
Experime ntal	15	Observat ion card.	0.16	0.04*	function	"Abnorm al."
Control	17					
Experime ntal	15	Knowled ge-based	0.23	0.00*	function	"Abnorm al."
Control	17					

C. Study Materials

1) Designing an e-learning environment: Previous studies were examined to develop an e-learning environment based on the theory of communication [13,14]. The Siemens model considered in the present study comprises of five stages as illustrated in Table III (Siemens, 2005).

TABLE III. The Five Stages

The First Stage: The Field Scope				
Planning		The Second Stage: Construction Creation		
	Analysis			
		Design		
			Development	
				Delivery
• Target Category • Budget • methods of delivery • Strategy used. • Formal and informal learning.	• Learners' learning range. • Available technology. • Students • Nature of content. • Support required	• Learning Goals. • Selection of technological media. • Strengthening interaction • A variety of shapes, layouts, external appearances)	• Identify required skills. • Identify content experts. • Determine the evolution of the incident through the schedule.	• Play content. • Run and handle links that do not work.
Experimentation (during Phase I and II)			The Fourth Stage: Top Calendar Meta -Evaluation	
The Third Stage: User experience and experience User experience and piloting				
The Fifth Stage: Formative and final evaluation (for the first, second and third stages) Formative and summary evaluation (stages 1,2,3)				

The First Stage - The field phase consists of two processes; the planning process includes the following:

- Identify the target group: They are students of the King Khalid University.
- Content Identification: The content is in the “Computer in Education” course for students of the general diploma for the second semester (1438).
- Determining the budget for instructional design: The researcher used the e-learning (system blackboard) available at King Khalid University. It is available free of charge to students and teachers for networking.
- Identify formal and informal learning methods for educational content: The study of the course “Computer in Education” through the Learning Management System “Blackboard”. The platform Pbworks was also used for informal learning.
- Determining the general strategy followed: The researcher defined the strategy of learning through sharing and production in the educational task among students, as shown in Table IV.
- The researcher also identified the general strategy after integrating the activities into the formal learning as shown in Table V.
- Determining the delivery and delivery of instructional content: Learned content was delivered by students formally or informally through the World Wide Web.

D. Development Stage

At this stage, some computer programs were used to produce an e-learning environment based on the theory of communication. The most prominent of these programs and sites include; website Pbworks, website Appmakr, and IrfanView that is a special program for designing image and keeping its quality unchanged. The programs also include graphics program photoshop, location Google drive, production of a guide for the first group and second group.

1) Delivery: The main objective of this stage was to ensure that the educational content was running in e-learning management system and on the external sites that will be used by students in the formal and non-formal educational system. At the third stage, the content of the e-learning environment based on the theory of communication ensures that all the contents of the electronic environment were used by the student effectively, and referred to the e-environment environment for diploma students.

2) Evaluation: The measurement tools, namely: observation card and knowledge economy scale were applied electronically after studying all the educational contents of the students of the two study groups. The analysis process included the identification of target group. They had skills of using computers and the Internet.

3) Construction creation: The construction phase includes the design process and its objectives of learning were determined according to the formal and non-formal use of computers in education. Behavioral objectives for each lesson were determined according to the “Bloom Digital” classification.

The researcher developed a model for writing the content scenario of “Computer in Education” as shown in Table VI.

TABLE IV. DETERMINING THE GENERAL STRATEGY

Mission	Activity	Activity Execution Environment	Activity	Evaluation of activity	Decision
Writing a document in “Google”	Students will participate in writing the attached document by the researcher and upload it in an environment Pbworks	www.drive.google.com	Two days	The researcher evaluates the document prepared by the students	Students and teachers are discussed to make sure that the “document” attached by the researcher is correct and grades are given to the students who participated in the writing

TABLE V. GENERAL STRATEGY FOR INTEGRATION

Educational event	Learning		Teacher Role	The role of the learner
	Official	Informal		
Introducing students to the e-learning environment in the light of communication theory.	Official		The teacher introduces the students to the course (using the computer in education) and how to participate in informal learning environments (Pbworks). Teacher trains students to enter the site by username & password	Students enter the site elearning.kku.edu.sa For training on how to access the site. Students learn the contents of the course (use of computers in education). Login on the site www.pbworks.com In order to learn how to enter them.
Creating a website online.		Informal	The teacher guides students to how to build a personal site through a site Pbworks. The teacher provides a guide to how to build a website.	Create an e-site online. Each student displays the name of the site he designed in the blog Pbworks. Allow students to comment. Giving a degree to each student who designed a website online.

TABLE VI. CONTENT SCENARIO OF “COMPUTER IN EDUCATION”

Learning objective	Educational content of the goal	Multimedia					shape	Relay and sailing
		Text:	Audio	Photocopy	graphic	Video		

E. Preparation of Measuring Instruments

The note card was prepared to measure the behavioral performance of the Diploma students in the Faculty of Education, King Khalid University concerning skills of cloud computing in different educational fields. The main dimensions of the card were identified after studying the researches and studies that dealt with this aspect. These dimensions have been illustrated in Table VII.

After completing the preparation of the card, the researcher presented the card to a group of specialists in the field of educational technology, curriculum, and teaching methods and psychology. Their opinions indicated the suitable items of the card for study sample, with an amendment in the wording of some paragraphs in the second and fourth dimension. The observation card was applied to a sample of seven students after observing the opinions of arbitrators to determine the correctness of the language of the skills on the card, in terms of design.

F. Knowledge Economy Skills Scale

The knowledge economy skills scale has been prepared to provide the skills of knowledge economy to the students. The dimensions of the scale include; collaborative and collective work, innovation and creativity, problem solving and decision making, critical thinking, and application of technology. The scale included five dimensions as shown in Table VIII.

TABLE VII. DIMENSIONS OF THE CARD

Sr	Dimension	Statements
1	The first: Special skills in dealing with documents.	6
2	The second: Presentation skills.	6
3	The third: Special skills for spreadsheets.	8
4	The fourth: Special skills for creating electronic tests Online.	6
5	Fifth: Skills for creating interactive websites Online.	5
6	VI: Special skills for creating interactive video Online.	6
7	Seventh: Skills for creating interactive e-courses Online.	8
Total	4	45

TABLE VIII. THE NUMBER OF DIMENSIONS AND ITEMS OF THE SCALE IN ITS PRIMARY FORM

Sr	Dimension	Statements
1	Cooperative and collective action	6
2	Innovation and creativity	8
3	Problem solving and decision making	5
4	Critical Thinking	7
5	Application of technology	10
Total	5	36

III. RESULTS AND DISCUSSION

The study has reviewed the design models of the e-learning environments and the studies concerned with the theory of communication. There was a statistically significant difference at the level (0.05) between middle-grade experimental group (trained by electronic learning environment based on the theory of “connectivity”) and control group (trained by e-learning management system Blackboard).

The statistical analysis was performed to test validity using the Mann-Whitney test to compare two independent samples. Table IX shows the results of applying the test to indicate the differences between the two grades of the two groups concerned with skills of cloud computing. Table IX shows that the value of test was calculated for observing card for cloud computing skills. This shows that there was a statistically significant difference between the intermediate grades of the students’ grades in the post-application favoring the higher-grade average.

The ability of the experimental group to use cloud computing skills in a computer course was higher and statistically significant, as compared to the control group students in this course. This means that experimental group students have benefited from a participatory learning environment based on the theory of communication better than students who have been trained in the usual way of using the Blackboard environment.

There is a statistically significant difference at the level (0.05) between middle-grade experimental group and control group. Statistical analysis was performed using the Mann-Whitney test to compare two independent samples. Table X shows the results of applying the test to indicate the differences between the intermediate grades of the two groups concerned with the knowledge economy skills. Table X shows that that there was a statistically significant difference between the intermediate grades of the students in the post-application favoring the higher-grade average. This was in the favor of the experimental group.

The Mann-Whitney test was used to compare two independent samples, to find out difference in the sub-skills of the knowledge economy. Table XI shows the results of applying the test to indicate the differences between the intermediate grades of the two groups concerned with the knowledge economy skills.

Table XII shows the impact of the participatory e-learning environment among the students, which was equal to 75%. The percentage of the impact of the e-learning environment in the development of knowledge economy skills was 74%.

The correlation coefficient matrix (Spearman) was found between dimensions of the scale and the total score as shown in Table XIII. The correlation coefficient of the first dimension in the scale equal 0.75 and the correlation coefficient of the second dimension in the scale equal 0.76. The coefficient of the third dimension in the scale equals 0.46 and the fourth-dimension correlation coefficient by the scale as 0.58. Whereas, the scale equal 0.68, where all values are functionally and statistically acceptable.

The way students view and deal with the content of a participatory learning environment increase their motivation to learn and have positive attitudes toward learning through collaborative environment. These results are consistent with the studies conducted by [15-20]. The learning environment, designed according to the "communicative" theory is concerned with the needs of learners, which help them to work in organized partnership and carry out the tasks in an organized manner, along with providing continuous feedbacks. The interest of the study in checking the information obtained from the various websites has helped the students to develop their critical thinking skills.

TABLE IX. VALUES "U" AND ITS STATISTICAL SIGNIFICANCE AMONG THE MIDDLE RANKING STUDENTS IN THE APPLICATION DIMENSION NOTE CARD

Group	Implementation	N	FS-3 Average grade	Total grade	Values (U) Calculated	Values (Z) Calculated	Semantics
Demos	next	15	25	475	0.0 *	-4.82 *	Function
Control		17	9	153			

*Values U Table 75 (0.05, 15, 17)

TABLE X. VALUES OF "U" AND ITS STATISTICAL SIGNIFICANCE AMONG MIDDLE RANKING STUDENTS TO KNOW THE KNOWLEDGE ECONOMY SKILLS SCALE

Group	Implementation	N	FS-3 Average grade	Total grade	Values (U) Calculated	Values (Z) Calculated	Semantics
Demos	next	15	24.93	374	1.0*	-4.78 *	Function
Control		17	9.06	154			

*Values U Table 75 (0.05, 15, 17)

TABLE XI. VALUES OF "U" AND ITS STATISTICAL SIGNIFICANCE AMONG MIDDLE RANKING STUDENTS CONCERNED WITH THE KNOWLEDGE ECONOMY SKILLS SCALE

Group	Skill	N	FS-3 Average grade	Total grade	Values (U) Calculated	Values (Z) Calculated	Semantics
Demos	Cooperative and collective action	15	23.67	355	20,00	4.09	Function
Control		17	10.18	173			
Demos	Innovation and creativity	15	18.87	283	92.00	1.36	Not function
Control		17	14.41	245			
Demos	Problem solving and decision making	15	23.00	345	30.00	*3.72	Function
Control		17	10.76	183			
Demos	Critical Thinking	15	19.67	295	80.00	1.82	Not function
Control		17	13.71	233			
Demos	Application of technology	15	24.43	366.50	8.50*	4.51	Function
Control		17	9.50	161.50			

*Values U Table 75 (0.05, 15, 17)

TABLE XII. SCIENTIFIC AND PRACTICAL IMPORTANCE OF THE RESULTS OF THE STUDY

Independent variable:	Dependent variable:	$\eta^2= z^2/ n-1$	Impact
E – learning environment	Cloud computing and AI	0.75	large
	Knowledge-based	0.74	large

TABLE XIII. CORRELATIONAL ANALYSIS

Dimension	Cooperative and collective action	Innovation and creativity	Problem solving and decision making	Critical Thinking	Application of technology
Cooperative and collective action	1				
Innovation and creativity	0.43	1			
Problem solving and decision making	0.27	0.06	1		
Critical Thinking	0.24	0.48	0.04	1	
Application of technology	0.45	0.75	0.21	0.41	1
Scale as a whole	0.75	0.76	0.46	0.58	0.68

IV. Conclusion

The present study has developed the skills of the cloud computing applications and the knowledge economy among the university students by designing a participatory electronic learning environment. The results have shown that the design of a participatory e-learning environment based on the theory of communication contributed to improve the skills level of cloud computing applications and knowledge economy skills among the students from Saudi Arabian universities. Based on the study findings, it is recommended that teachers need to pay attention towards training of students during and before service regarding the use of modern technologies in the field of education. There is a need to train teachers for employing modern theories related to technology to support the teaching and learning initiatives in the universities such as social theory communicative.

The study has highlighted the relationship between cloud computing and social constructivism theory in the context of the educational system of Saudi Arabia. The study findings would be of great help for all the stakeholders in understanding this perspective. Moreover, the study findings highlight the need of focusing on blueprint for further research.

Acknowledgments

The author is very thankful to all the associated personnel in any reference that contributed in/for the purpose of this research.

References

[1] M. G. Avram, “Advantages and Challenges of Adopting Cloud Computing from an Enterprise Perspective,” Procedia Technology, vol. 12, pp. 529–534, 2014.

[2] R. Misevičienė and G. Budnikas, “Apskaitos sistemų projektavimas,” Aug. 2012.

[3] Albadani. R. M and Mostafa. A. F,“The Effect of different Electronic support systems through cloud computing on developing 3rd year students” Computer knowledge. Life Science Journal, vol. 14, pp.42-54, 2017.

[4] G. H. Al Sulim, “Evaluating Professional Skills of the Faculty Members at Al Imam Muhammad Bin Saud University by Female Graduate Students of the College of Social Sciences in Light of Total Quality Standards,” SSRN Electronic Journal, 2013.

[5] Knowledge of Islamic Jurisprudence among Secondary School Students in Kuwait,” Journal of Educational and Psychological Studies, vol. 10, no. 1, pp. 106–119, Jan. 2016.

[6] K. Hafez, “A Review of: ‘Arab Mass Media: Newspapers, Radio, and Television in Arab Politics/The Making of Arab News,’” Political Communication, vol. 24, no. 1, pp. 96–98, Jan. 2007.

[7] F. Zafar, A. Khan, S. U. R. Malik, M. Ahmed, A. Anjum, M. I. Khan, N. Javed, M. Alam, and F. Jamil, “A survey of cloud computing data integrity schemes: Design challenges, taxonomy and future trends,” Computers & Security, vol. 65, pp. 29–49, Mar. 2017. https://doi.org/10.1016/j.cose.2016.10.006.

[8] I. Arpaci, “Antecedents and consequences of cloud computing adoption in education to achieve knowledge management,” Computers in Human Behavior, vol. 70, pp. 382–390, May 2017. https://doi.org/10.1016/j.chb.2017.01.024.

[9] T. Li, “Construction and Implementation of Network Teaching Platform for Design Art Education Based on Cloud Technology,” Educational Sciences: Theory & Practice, 2018.

[10] Y.-M. Huang, “Exploring the intention to use cloud services in collaboration contexts among Taiwan’s private vocational students,” Information Development, vol. 33, no. 1, pp. 29–42, Jul. 2016. https://doi.org/10.1177/0266666916635223.

[11] H. Al-Samarraie and N. Saeed, “A systematic review of cloud computing tools for collaborative learning: Opportunities and challenges to the blended-learning environment,” Computers & Education, vol. 124, pp. 77–91, Sep. 2018. https://doi.org/10.1016/j.compedu.2018.05.016.

[12] C. Powell, Katherine, J. K. Cody, “Cognitive and social constructivism: developing tools for an effective classroom,” Education, vol. 130, no, 2, p. 241, 2009 “Master’s Degree Program for Applied Informatics in Education Majoring in Instructional Design and Distance Learning,” Proceedings of the 5th International Conference on Computer Supported Education, 2013.

[13] “Master’s Degree Program for Applied Informatics in Education Majoring in Instructional Design and Distance Learning,” Proceedings of the 5th International Conference on Computer Supported Education, 2013.

[14] C. Tan, “‘To Develop Every Student’: Towards Quality-Oriented Education,” Learning from Shanghai, pp. 79–87, Oct. 2012.

[15] Ahmed. A. T, “The impact of the design of an electronic learning environment in the light of the communicative theory on the development of achievement and the skills of personal knowledge management among students of educational techlogy”. Master Thesis, Faculty of Specific Education, Tanta University.

[16] J. G. S. Goldie, “Connectivism: A knowledge learning theory for the digital age?,” Medical Teacher, vol. 38, no. 10, pp. 1064–1069, Apr. 2016.

[17] M. M. Hassan, A. N. Qureshi, A. Moreno, and M. Tukiainen, “Participatory Refinement of Participatory Outcomes: Students Iterating over the Design of an Interactive Mobile Learning Application,” 2017 International Conference on Learning and Teaching in Computing and Engineering (LaTICE), Apr. 2017.

[18] The Effect of Connectivism Theory - Based Online Collaborative Learning on Academic Self - Efficacy and Mastery Motivation of Instructional Technology Private Diploma Students,”Arabix journal of education and psychology, no. 62, pp. 129–162, Jun. 2015.

[19] S. Darrow, “Connectivism learning theory: instructional tools for college courses. Master’s Degree”, Western Connecticut state university.

[20] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

Image Detection Model for Construction Worker Safety Conditions using Faster R-CNN

Madihah Mohd Saudi[1], Aiman Hakim Ma'arof[2], Azuan Ahmad[3], Ahmad Shakir Mohd Saudi[4], Mohd Hanafi Ali[5]
Anvar Narzullaev[6], Mohd Ifwat Mohd Ghazali[7]
Cyber Security and Systems (CSS) Research Unit, Islamic Science Institute (ISI)
Universiti Sains Islam Malaysia (USIM), Nilai, Malaysia[1, 3]
Faculty of Science & Technology (FST), Universiti Sains Islam Malaysia (USIM), Nilai, Malaysia[2, 6, 7]
Institute of Medical Science Technology, Universiti Kuala Lumpur (UNIKL), Kuala Lumpur, Malaysia[4]
Consultation, Research and Development Department[5]
National Institute of Occupational Safety and Health (NIOSH), Bangi, Malaysia[5]

***Abstract*—Many accidents occur on construction sites leading to injury and death. According to the Occupational Safety Health Administration (OSHA), falls, electrocutions, being struck-by-objects and being caught in or between an object were the four main causes of worker deaths on construction sites. Many factors contribute to the increase in accidents, and personal protective equipment (PPE) is one of the defense mechanisms used to mitigate them. Thus, this paper presents an image detection model about workers' safety conditions based on PPE compliance by using the Faster Region-based Convolutional Neural Networks (R-CNN) algorithm. This experiment was conducted using Tensorflow involving 1,129 images from the MIT Places Database (from Scene Recognition) as a training dataset, and 333 anonymous dataset images from real construction sites for evaluation purposes. The experimental results showed 276 of the images being detected as safe, and an average accuracy rate of 70%. The strength of this paper is based on the image detection of the three PPE combinations, involving hardhats, vests and boots in the case of construction workers. In future, the threshold and image sharpness (low resolution) will be two main characteristics of further refinement in order to improve the accuracy rate.**

Keywords—PPE; OSH; accident; construction site; image detection; faster R-CNN

I. Introduction

One of the most dangerous fields to work is the construction industry. The Occupational Safety and Health Administration (OSHA) has outlined safety measures and precautions in the form of a legislative framework for the construction industry. Based on works by [1-4], any worker, especially those in the construction industry, is exposed and vulnerable to accidents that could lead to non-permanent disability (NPD), permanent disability (PD) or death. As stated in work by [5], ignorance of safety procedures and protection such as wearing PPE, failing to understand written safety rules, and many migrant workers, are among the factors that lead to accidents. On top of that, the uniqueness of the industry and of construction site conditions, also play a big role in the cause of accident or death. Moreover, workers on construction sites could help to reduce the risk of accidents by informing their supervisor or employer of any risks that they have spotted so that appropriate control measures could be introduced to prevent such accidents. Generally, safety performance is measured based on lagging indicators such as Incident Rate (IR), Accident Rate (AR) and Experience Modification Rate (EMR). While works by [6-10], described the examples of safety measures available with different lagging indicators across the world. In addition, Abas and colleagues wrote a comprehensive paper on the factors that affects safety performance on construction projects [11]. These authors identified safety factors which are beneficial when it comes to reducing accident and compensation costs and to increasing productivity, employee awareness attitudes, and project on time completion. As proposed by [12,13], employers should evaluate their employees' knowledge and awareness regarding PPE and other equipment at the construction site, so that proper training and control measures could be implemented to reduce the accident or death risk.

Consequently, this paper presents an image detection model based on safe and dangerous condition facing workers on construction sites in terms of their compliance when it comes to wearing PPE. Existing works for construction sites on safety detection were more focusing on one PPE only such as hardhats, vest or boots. As for our paper, we proposed detection safety conditions based on three combinations of PPEs in the form of hardhats, vests and boots. Furthermore, for us to classify the worker as operating in a safe or dangerous manner, we used the Faster R-CNN algorithm. The PPE considered takes the form of hardhats, boots and vests. These are basic PPE that should be worn all time by construction workers.

This paper is organized as follows. Section II explains related work, Section III presents the method used in this research, Section IV consists of the findings and their evaluation, and Section V concludes the paper and makes suggestions for future work.

II. Related Works

According to work by [14], falling objects among construction workers is ranked as the one of the highest incident that occurred at construction sites. Hence, we need a solution to lower the death risk among such workers. There is a small number of existing works related to the construction system and to quality of project management such as [15,16, 23]. Work by [15] tended to focus on project progress monitoring, and communication between employees, while work by [16] proposed a quality management project for construction projects. Work by [23] proposed optimisation modelling for repetitive works on construction site by using an Unmanned Aerial Vehicle (UAV), and work by [26] detected workers on construction sites by using UAV and RFID concepts. Different scopes of work using UAVs were summarised by Ham and colleagues [27] such as for progress monitoring, building inspection, building measurement, surveying, safety inspection, structural damage assessment and geo-hazard investigation. As far as safety inspection concerned, existing works tend to focus on one piece of PPE only in the form of the hardhat. There are few existing studies related to image detection about construction sites, as summarised in Table I.

TABLE I. SUMMARISATION OF EXISTING WORKS RELATED TO CONSTRUCTION SITES

Author	Detection Method	Dataset	Challenges	Strength
[17]	Used: ResNet-152 and Faster R-CNN.	Training: ImageNet & MS COCO 2014 dataset. Testing: 3,241 images	Limited to features of construction worker's body only	Considering varying poses of the images
[18]	Used: Deep CNN and Faster R-CNN.	Training: 81,000 Testing: 19,000 images. Both were self-collected from 25 different construction projects.	Dedicated to existing construction workers.	Produce highly accurate results based on image size with average of more than 95%
[19]	Used: Google Inception v3	Training: 1208 images Testing: 27 images	It needs bigger image dataset.	Produced accuracy rate of 90%.
[20]	Used: Enhanced Faster R-CNN	Training: 9,500 images Testing: 1500 images	Full coverage of image sources.	Produced accuracy rate of more than 95%.
[21]	Used: Histogram of oriented gradients (HOG)	Training: 100 images (hardhat) & 1,800 dataset (people condition) Testing: Hundreds of self-collected images.	Limited to hardhat and worker images in standing position.	Achieved overall 94.3% precision.
[22]	Used: HOG and Circle Hough Transform (CHT).	Training: 954 images. Testing: 200 images.	Improvement for image detection.	Detection based on colours.
This research	Used: Faster R-CNN.	Training: MIT Places Database 1129 images Testing: 333 images (263 MIT, 65 self- collected, 5 google)	Improvement of accuracy rate based on picture size refinement.	Detection safety conditions based on 3 combinations of PPEs in the form of hardhats, vests and boots.

Based on Table I and on our summarised works, we have identified that the most common challenges for researchers in the future would be to have more PPE detection on construction sites or at the workplace, more dataset resources, the detection of object blockage endangering the target image, and different image positions and backgrounds. From the existing works, the best recommendations for image detection algorithm performance is the use of the Faster R-CNN algorithm. Hence, our research has tackled these challenges and used the Faster R-CNN algorithm as our detection algorithm.

III. METHOD

For this research, the setup, software and hardware used for the experiment as displayed in Table II, and the research processes are as depicted in Fig. 1.

TABLE II. EXPERIMENT SETUP

Hardware/ software	Description
Lenovo Legion i7-7700HQ @ 2.80 GHz	Computer specification.
Microsoft Windows 10 Home	Operating system that is used by the computer to run the project
NVIDIA® GeForce® GTX 1050 Ti	Graphic card specification to run Tensorflow software.
16 GB DDR4-2666 (1333 MHz)	RAM and central processing unit specification.
Tensorflow 1.15	Open source software used for training and testing the images. It consists of the Faster R-CNN algorithm. Tensorflow ran inside Anaconda for the usage of Tensorflow-GPU, which is faster than Tensorflow-CPU.
Anaconda	Virtual environment for Python code
LabelImg	Tool written in Python code for graphical image labelling, and for image training and testing.

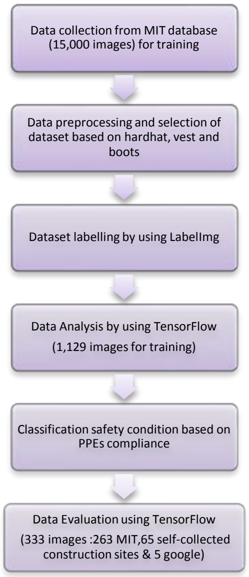

Data collection from MIT database (15,000 images) for training
Data preprocessing and selection of dataset based on hardhat, vest and boots
Dataset labelling by using LabelImg
Data Analysis by using TensorFlow
(1,129 images for training)
Classification safety condition based on PPEs compliance
Data Evaluation using TensorFlow
(333 images :263 MIT,65 self-collected construction sites & 5 google)

Fig. 1. Overall Research Processes.

For this research, the images were collected from the MIT Database [24]. From fifteen thousand construction images, 1,129 images were selected as the training dataset based on PPE components, which inclusive of hardhats, vests, and boots. These images were then labelled using LabelImg python scripting (see Fig. 2) and further analysed, trained and classified using Tensorflow. LabelImg is written in Python in order to label the images, together with Qt graphical interface. The annotations were saved as an XML file in PASCAL VOC. During the image analysis, we trained and classified the images by using the Faster R-CNN Inception v2 COCO model. This model uses fast R-CNN with shared convolutional feature layers, and a unified model composed of RPN (region proposal network) as depicted in Fig. 3.

The strength of Faster R-CNN is based on its ability to reuse the CNN results for the regional proposal process. Hence, only one CNN needs to be trained, and regional proposals can be made almost cost-free computationally [25]. Once the image has been inserted, the Faster R-CNN produces the classifications and bounding box co-ordinates of the specified classes in the images. In our research, this algorithm helps us to identify and to assign the safety condition based on PPE compliance. The safety condition was decided upon, as being either safe or unsafe (dangerous) based on worker compliance in terms of wearing a hardhat, vest and boots at the construction site as depicted in Fig. 4. Once this safety classification was completed, the evaluation was carried out with the aid of 333 images. For safety conditions, the formulation was as follows:

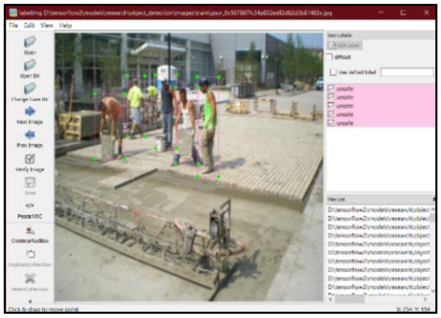


Fig. 2. Examples of Dataset Labelling of the Safety Conditions.

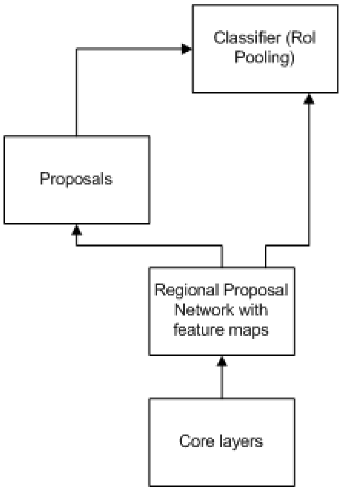


Fig. 3. Faster R-CNN Design.

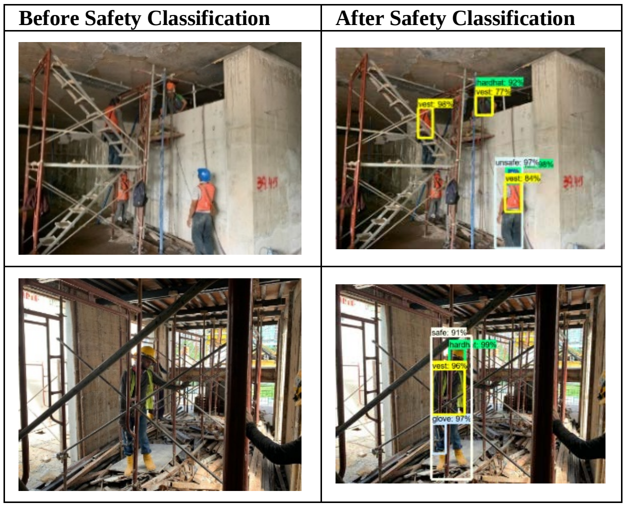


Fig. 4. Safety Classification Condition.

For PPE, there are three main variables, which are hardhat, vest and boots. The formula is as follows:

Let α_i be a hardhat I, and $\alpha = \bigcap_{i=1}^{n} \alpha_i$, β_j be a vest j, and $\beta = \bigcup_{i=1}^{m} \beta_i$, γ_k be as boots $\gamma = \bigcap_{i=1}^{p} \gamma_i$

Let PPE be the PPE classification and T be the target image. S is the safety model and it can be defined as the following function:

$$f\,(\text{PPE,T}) = \text{S} \tag{1}$$

where, $PPE(\alpha,\beta,\gamma) = \alpha \cap \beta \cap \gamma$

$$f\ (PPEi, Tj) = Sij \tag{2}$$

where, PPE represents the PPE classification, T represents the target image and S is the safety model.

$$PPE\ (\alpha, \beta, \gamma) = \ \alpha \cap \beta \cap \gamma \tag{3}$$

$$\alpha = \ \alpha_1 \ \cup \ \alpha_2$$

$$\beta = \ \beta_1 \ \cup \ \beta_2$$

$$\gamma = \gamma_1 \ \cup \gamma_2$$

$$\begin{bmatrix} PPE_i & \alpha\,\beta & \gamma \\ \vdots & \ddots & \vdots \\ PPE_n & \cdots & \delta_n \end{bmatrix}$$

where, α_1, α_2 : with hardhat, without hardhat

β_1 , β_2 : with vest, without vest

$\gamma_1 \ \cup \gamma_2$: with boots, without boots

The evaluation of this research is based on accuracy as follows.

$$Accuracy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \times 100 \tag{4}$$

where

Tp = True positive (number of worker correctly classified as safe).

Fp = False positive (number of worker incorrectly classified as unsafe).

Tn = True negative (number of worker correctly classified as unsafe).

Fn = False negative (number of workers incorrectly detected as safe.

The findings based on these formulations are explained in the next section.

IV. Findings

Based on the experiment conducted, 1,129 images related to PPE that included hardhat, vest and boots were trained in order to classify conditions as being safe or unsafe. Then, 333 anonymous self-collected images from construction sites were used for evaluation.

During the training, from 1,129 images, a total of 2,373 hardhats, 1,023 pairs of boots and 1,478 vests were detected. In terms of the evaluation of 263 images, a total of 156 hardhats, 49 boots, 73 vests and 123 safe conditions were detected with an overall accuracy of 70%. In the case of a further 70 images self-collected from construction sites, 53 cases were detected as being safe. Table III summarises the experimental results, while Fig. 5 shows examples of the evaluation image results. For this experiment, a total of 6 hours was used for image training, and the total accuracy loss was 0.5 or less, as displayed in Fig. 6.

Many factors contributed to the accuracy rate. Apart from the model itself, other factors such as the training dataset, input image resolution, and training configurations including batch size, input image resize, learning rate, and learning rate decay, also affected the accuracy rate [28]. In our case, the ability to detect safe conditions with an accuracy rate of 70% is considered as a good result in terms of real-time detection. It is hard and very subjective to make comparisons with any other existing works due to the different settings of the experiment for each existing work. However, the selection of the best detector algorithm and the best configuration are crucial in terms of image detection. These two factors contribute to the best balance of speed and accuracy. We chose the Faster R-CNN algorithm due to its better accuracy rate compared to that of other existing algorithms, as summarised in Table I. In addition, based on the experiment conducted, the formulations developed mean that the construction workers' compliance with wearing PPE can easily be identified and measured.

TABLE III. Experiment Results

PPE	Total Detected Images (Training)	Total Detected Images (Testing of 70 Images
Hardhat	2372	156
Vest	1478	73
Boot	1023	49
Safe	572	53

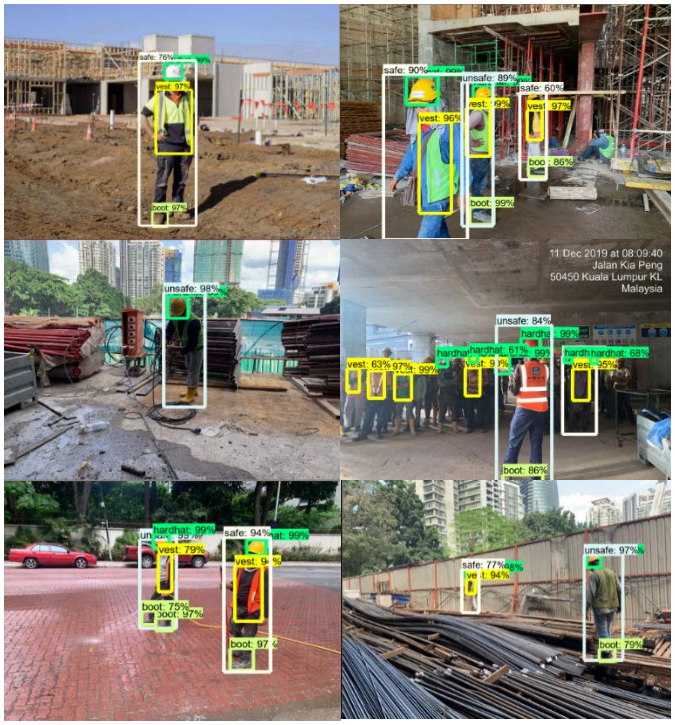


Fig. 5. Examples of the Tested Images.

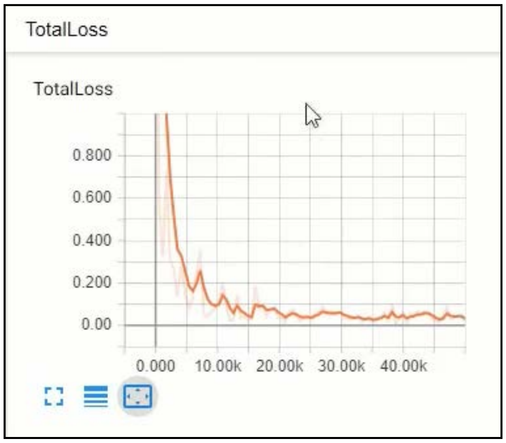


Fig. 6. Total Loss Accuracy Rate.

V. Conclusion

Based on the experiment conducted, there are a few considerations. These are the threshold value assigned during the data configuration settings, the momentum optimizer value, and the belief that image resize and image sharpness (low resolution) could be further adjusted or improved for better accuracy. All these elements are among the components of the Faster R-CNN algorithm. Nonetheless, this paper has successfully developed formulations and an image detection model relating to construction workers' compliance when it comes to wearing PPE in the workplace. This will help create a safer and healthier environment on construction sites. For future work, the threshold, momentum optimizer, image resize and image sharpness will be further refined and improved to obtain an improved accuracy rate.

ACKNOWLEDGMENT

The authors would like to express their gratitude to National Institute Of Occupational Safety And Health (NIOSH) Malaysia, UNIKL and Universiti Sains Islam Malaysia (USIM) for the support and facilities provided. This research paper project is under NIOSH grant: [03-16/03/DRONE(E)/2019/01] / [USIM/NIOSH-UNIKL/FST/052003/42119].

REFERENCES

[1] Nath, Nipun D., Amir H. Behzadan, and Stephanie G. Paal. "Deep learning for site safety: Real-time detection of personal protective equipment." Automation in Construction, 112, 103085, 2020.

[2] Wright, Tamara, Atin Adhikari, Jingjing Yin, Robert Vogel, Stacy Smallwood, and Gulzar Shah. "Issue of compliance with use of personal protective equipment among wastewater workers across the southeast region of the United States." International Journal of Environmental Research and Public Health ,16, no. 11, 2019.

[3] [3] Wong, Tom Ka Man, Siu Shing Man, and Alan Hoi Shou Chan. "Critical factors for the use or non-use of personal protective equipment amongst construction workers." Safety science, 126: 104663, 2020.

[4] M. Gheisari and B. Esmaeili, "Unmanned Aerial Systems (UAS) for Construction Safety Applications," in Construction Research Congress 2016: Old and New Construction Technologies Converge in Historic San Juan - Proceedings of the 2016 Construction Research Congress, CRC 2016, 2016, pp. 2642–2650.

[5] V. H. P. Vitharana, G. H. M. J. S. De Silva, and S. De Silva, "Health hazards, risk and safety practices in construction sites – a review study," Eng. J. Inst. Eng. Sri Lanka, vol. 48, no. 3, p. 35, Jul. 2015.

[6] Dyck D and Roithmayr T., "Great Safety Performance: an improvement process using leading indicators. - PubMed - NCBI," AAOHN J., vol. 52, no. 12. pp. 511–520, Dec-2004.

[7] J. Lin and A. Mills, "Measuring the occupational health and safety performance of construction companies in Australia," Facilities, vol. 19, pp. 131–139, Mar. 2001.

[8] Hasnora Jafri, Mohd Wijayanuddin Ali, Arshad Ahmad, and Mohd Zaki Kamsah, " Effective occupational health and safety performance measurements," in Int. Conf. of Chemical and Biochecmical Eng. (Sabah), 2005, pp. 702–708.

[9] S. Salminen, J. Saari, K. L. Saarela, and T. Rasanen, "Organizational factors influencing serious occupational accidents," Scand. J. Work. Environ. Heal., vol. 19, no. 5, pp. 352–357, 1993.

[10] Construction Industry Development Board (CIDB), "Safety and Health Assessment System in Construction(SHASSIC) CIS 10:2008," 2008.

[11] N. H. Abas, N. Yusuf, N. A. Suhaini, N. Kariya, H. Mohammad, and M. F. Hasmori, "Factors Affecting Safety Performance of Construction Projects: A Literature Review," in IOP Conf. Series: Materials Science and Engineering 713 (2020) 012036, 2020.

[12] T. D. Smith and D. M. Dejoy, "Safety climate, safety behaviors and line-of-duty injuries in the fire service," Int. J. Emerg. Serv., vol. 3, no. 1, pp. 49–64, Jan. 2014.

[13] Y. Li, Y. Ning, W. T. Chen, and D. W. M. Chan, "Critical Success Factors for Safety Management of High-Rise Building Construction Projects in China," 2018.

[14] Sanni-Anibire, Muizz O., Abubakar S. Mahmoud, Mohammad A. Hassanain, and Babatunde A. Salami. "A risk assessment approach for enhancing construction safety performance." Safety science, vol. 121, pp. 15-29, 2020.

[15] A. Khelifi and K. Hesham Hyari, "A Mobile Device Software to Improve Construction Sites Communications 'MoSIC,'" IJACSA) Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 11, 2016.

[16] P. Thanh Nguyen et al., "Construction Project Quality Management using Building Information Modeling 360 Field," IJACSA) Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 10, 2018.

[17] H. Son, H. Choi, H. Seong, and C. Kim, "Detection of construction workers under varying poses and changing background in image sequences via very deep residual networks," Autom. Constr., vol. 99, pp. 27–38, Mar. 2019.

[18] W. Fang, L. Ding, B. Zhong, P. E. D. Love, and H. Luo, "Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach," Adv. Eng. Informatics, vol. 37, pp. 139–149, Aug. 2018.

[19] D. Gil, G. Lee, and K. Jeon, "Classification of Images from Construction Sites Using a Deep-Learning Algorithm," in 35th International Symposium on Automation and Robotics in Construction (ISARC 2018), 2018, pp. 176–181.

[20] Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., Rose, T.M. and An, W., "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos," Autom. Constr., vol. 85, pp. 1–9, Jan. 2018.

[21] M.-W. Park, N. Elsafty, and Z. Zhu, "Hardhat-Wearing Detection for Enhancing On-Site Safety of Construction Workers," J. Constr. Eng. Manag., vol. 141, no. 9, p. 04015024, Sep. 2015.

[22] A. H. M. Rubaiyat et al., "Automatic Detection of Helmet Uses for Construction Safety," in 2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW), 2017, pp. 135–142.

[23] S. Bang and H. Kim, "Context-based information generation for managing UAV-acquired data using image captioning," Autom. Constr., vol. 112, p. 103116, Apr. 2020.

[24] "MIT Places Database for Scene Recognition." [Online]. Available: http://places.csail.mit.edu/index.html.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[26] M. Abbas, B. Eddine Mneymneh, and H. Khoury, "USE OF Unmanned Aerial Vehicles and Computer Vision in Construction Safety Inspections," in Proceedings of the Third Australasia and South-East Asia Structural Engineering and Construction Conference (ASEA-SEC-3), 2016.

[27] Ham, Youngjib, Kevin K. Han, Jacob J. Lin, and Mani Golparvar-Fard. "Visual monitoring of civil infrastructure systems via camera-equipped Unmanned Aerial Vehicles (UAVs): a review of related works." Visualization in Engineering 4, no. 1 (2016): 1.

[28] Jonathan Hui, "Object detection: speed and accuracy comparison (Faster R-CNN, R-FCN, SSD, FPN, RetinaNet and YOLOv3)," 2018.[Online]. Available: https://medium.com/@jonathan_hui/object-detection-speed-and-accuracy-comparison-faster-r-cnn-r-fcn-ssd-and-yolo-5425656ae359.

A Systematic Overview and Comparative Analysis of Service Discovery Protocols and Middlewares

Jawad Hussain Awan[1]*
Institute of Information and
Communication Technology
University of Sindh, Jamshoro, Pakistan

Usman Naseem[2], Nazar Waheed[4]
University of Technology
Sydney, Australia

Shah Khalid Khan[3]
RMIT University
Melbourne, Australia

***Abstract*—Context is major source of communication, where information is gathered easily from user context due to the progress of smart and context aware systems. Even, service directory also supports the systems to response the requests, sent through client. In this paper, the authors overviews context aware systems, their sensing capabilities in location or beyond location along with COIVA (a context aware system). Eight discovery protocols along with their functionalities (such as DEAPspace, DNS-SD, JXTA, RDP, LDAP, CORBA Trader, UDDI and Superstring) are discussed and compared to evaluate the performance and efficiency of system. In addition, six middleware (such as CAMPUS, CASF, SeCoMan, CoCaMAAL, BDCaM, and FlexRFID) are compared to evaluate factors (such as Architectural style, Context abstraction/Reasoning level, Context awareness level, Contextual adaptation approaches, Decision making, and Programming model). The authors further categorized them into sub categories discussed in Section 4 and named CoCaMAAL as better middleware as compared to others.**

Keywords*—*Pervasive computing; service discovery protocols; context-aware; middleware; privacy

I. Introduction

In Modern advancement in Pervasive computing brings a number of challenges like heterogeneity, scalability, privacy and more. It is possible because of code, user and device mobility. Hence, the research opportunities may support to discover new applications in environment with the help of these factors. The interaction between resource and user takes place in ad-hoc, permanent and transience nature of computing environment [1]. In this way, a number of applications are proposed having concept of location and context awareness. Context awareness not only enables the context aware systems, but also progresses the applications, such as context-aware recommendation, context-aware reminders, and intelligent call redirection [2], [3]. Human imagination persistent the invention and aggressive environment creates delay in pervasive computing. Hence, there is no any technical solution for human interface and complexity [4]. Thus, the main objective of pervasive computing research is to find and discover solution for above highlighted issues as invention may progress. Thus, discovery protocols are proposed to discover, advertise and register resources, services for clients or users. The main theme of this paper is to provide an overview and comparative analysis of service discovery protocols and middlewares. This paper also comprises of five sections. Section II discusses and compares the service discovery protocols. Section III discusses context-aware and its existing systems, while Section IV compares and discusses middlewares, while section V provides discussion regarding critical analysis of service discovery protocols and middlewares. Section VI concludes the research precisely.

II. Service Discovery Protocols

In this section, following service protocols are discussed and illustrated.

A. DEAPspace

The DEAPspace protocol is developed by Institute of Business Management (IBM) to focus small similar scale networks of Bluetooth. Specially, DEAPspace priorities to all nodes as being available in broadcast range while Konark rely on the TCP/IP rather than IP. DEAPspace also utilize a dynamic slotting of broadcast scheme, and pro-actively pushes the advertisements within the network. Hence, the resources knowledge is available among nodes of network.

DEAPspace protocol uses input or output schemes to define the services. The description of scheme is dependent of MIME and hierarchical in nature [5]. Though, each element may comprise of attributes. Like "Application → PostScript →version2", the PostScript element has attributes like color = value and ppm = value. The value for color may be Yes or No while ppm (Pages Per Minute) value may be a numeric value. For example, color=no and ppm=10. Furthermore, DEAPspace neither supports query relaxation nor expressions over the attributes.

B. DNS-SD (Domain Name System - Service Discovery)

DNS-SD is an Apple's Rendezvous technology and offers service discovery functionality for [6]. At initial level, these devices uses ZeroConf draft IETF standard to assign IP addresses from link-local range. Whenever, a node lacks IP addresses and creates conflict with peer nodes in network. Hence, each node assigned an IP address to avoid conflict among peers as well as uses DNS-SD in service allocation. Each node or device hosts a DNS server and clients to register or locate services on the network as easily discovered by other clients whenever requested utilize multicast (mDNS-SD) messages. Though, Apple provides well-known implementation of DNS-SD in Rendezvous technology, and DNS scheme is used to construct DNS-SD. New message is not defined in DNS scheme while scaled to internet. The scalability feature is gained at particular domain while DNS-SD does not provide service discovery to use cases, relax

*Corresponding Author

queries. Therefore, it relies on DNS defined messages. It is also notified from the research that DNS-SD is only suitable for IP based environments and requires suitability of other than IP supported environment.

C. JXTA (Juxtapose)

JXTA is Peer-to-Peer (P2P) technology and build by Sun Microsystems [7]. This technology is helpful to design P2P applications. Further, JXTA is composed of following protocols and their features illustrated in table no such as Peer Discovery Protocol (PDP), Custom Service Discovery Protocol (CSDP), Higher-level Service Discovery Protocol (HSDP), Peer Resolver Protocol (PRP), and Rendezvous Protocol (RP).

PDP is responsible to discover resource for an advertisement within peer group. The resources include peers, peer-groups, pipes and modules [8]. CSDP is implemented within non-world peer-groups, and then PDP works for bootstrapping purpose. If custom discovery protocol is unavailable then PDP is responsible for discovering of resources [9]. HSDP is utilized to process detailed discovery information. The information includes discovery of queries and advertisement and both are based on Extensible Markup Language (XML). In the definition of HSDP, reduction in the interaction and increase in scalability ratio may take place in peer groups. In few applications, PDP is allowed to discover resources as well [10]. PRP is responsible to route query and its response in peer group of JXTA. The PRP also forwards every method to a specified handler, that handler has semantic definition of message, the message is then sent to peer or a group peers within network [11]. The sending and receiving of message is the function of RP. RP also locates resources and limits the scalability also. Few nodes become rendezvous peers and propagate the messages to subscribed peers by them in network or peer group [12].

Table I compares five JXTA protocols such as PDP, CSDP, HSDP, PRP, and RP having Peer Group, Non-Peer Group, Detailed discovery information, Rendezvous Peers, Semantic definition, Scalability, Resource advertisement, and Resource discovery. PDP has four maximum features capability such as Peer Group, Non-Peer Group, Resource advertisement, and Resource discovery as compared to other protocols. Hence, PDP is favorable than others.

D. RDP (Remote Desktop Protocol)

RDP protocol is designed by Perkins and Harjono [13], used for fixed network mobile nodes. These nodes move from one network to another network like when an employee moves his/her laptop or handheld device from office (LAN network) to home (LAN network) or vice versa . The RDP uses a bootstrap approach to centralize a database, resides in the network. In this protocol, resource description uses URL and keywords as forms and queries are based on URNs (Unique Reference Numbers). URN consists of service information, optional resolution path, optional naming authority, and keywords. The initial section of URN describes the service information such as type. URN may be n^{21} or n^{2c}. n^{21} represents that one matched resource must be returned while n^{2c} represents all matched resources. DHCP provides the address of resource database and it is overridden by optional resolution path. Naming authority name the institution where mobile device discovers itself. Naming authority is also interpreting mechanism of scheme field. The designed protocol uses UDP (User Datagram Protocol) to deliver registration description and advertisements for queries. This protocol is also applicable for IP based environment like DNS-SD.

E. LDAP (Lightweight Directory Access Protocol)

LDAP is an easy version of the X.500 standard [14]. LDAP is a programming model to design discovery or registry service while it lacks the features of discovery protocols. LDAP is suitable for distributed environment and provides important features utilized in resource discovery. Such as duplication, exchange, and security. Though LDAP is mainly a directory consists of data types, objects (limited to a directory entry). LDAP is also utilized as a resource directory and discovery protocol. In this way, services are registered in directory and clients search or discover them for request of resource [15], [16]. For example, Java objects are stored or registered in the LDAP directory for a retrieval request. The schema is only source to describe that how objects can be stored in directory.

TABLE I. JXTA PROTOCOLS AND FUNCTIONALITY

Ref	**Protocol**	**Peer Group**	**Non-Peer Group**	**Detailed discovery information**	**Rendezvous Peers**	**Semantic definition**	**Scalability**	**Resource advertisement**	**Resource discovery**
[8]	**PDP**	√	√	×	×	×	×	√	√
[9]	**CSDP**	×	√	×	×	×	×	√	√
[10]	**HSDP**	√	×	√	×	×	√	×	×
[11]	**PRP**	×	×	×	×	√	×	×	×
[12]	**RP**	√	×	×	√	×	√	×	×

F. CORBA Trader

Offering and discovering a service is primary function of discovery protocols. Various institutes developed their own protocols for these services like Open Distributed Processing (ODP) trading function [17]. ODP has both features such as offering and discovering a service. Though, these services are known as exporting and importing capabilities. Generally, ODP is a service model but not implemented yet while CORBA trading function is implemented for importing and exporting capability. Like other protocols, CORBA trader is based on five components. Lookup, Link, Register, Proxy, Admin. The clients uses Lookup component to find services. These services are advertised or registered with the help of Register component in the Trader. Then, Link component is responsible to perform internal operation of Traders. While, admin component is also responsible to set trader policies and the legacy services are wrapped or hidden by Proxy component. CORBA trader is suitable for only static networks where nodes are known or defined while unsuitable for dynamic networks like MANETs (Mobile Ad hoc Networks). In this discovery model, services are chosen via granularity rather than query relaxation.

G. Universal Description, Discovery and Integration (UDDI)

UDDI is a general arrangement in which businesses are enabled to find offered services along with businesses. In this arrangement, when a new or suitable service is offered or found, then, that service is integrated into application. UDDI has four types of entities (Business entity, Business service, tModel), describe the UDDI information and presented in XML. Business entity is at upper level and contains the information of name, address, service of business. At least one or more services are registered in each business entity and combines similar web services offered by businesses [18]. Information in UDDI is described by four entity types, each of which is represented in XML. The business Entity is the top-level structure. It contains information such as the name of the business (or other entity, such as a department within an institution), its address and the type of service the business provides. Each business Entity contains one or more business Services. This entity type logically groups a set of related web services provided by the business. The business Service is a descriptive structure. It lacks technical information but bind with template structures, which contains technical information like how to invoke or interact or respond the web service. At last, tModel is existing structure outer side of the hierarchy. The structure of tModel utilizes and defines reusable components.

H. Superstring

Superstring is most efficient and appropriate for static and dynamic environments because it defines three components (two routing protocols and one API for resource or service discovery) in environments [19]. Superstring also scales the number of nodes or devices from computers to handheld or portable devices. The dynamic network deploys a routing protocol that progresses the network to decrease the query processing time for services or resources and adjusts to change in the network. While static network also deploys a routing protocol, that is responsible to scale the resources in wide area network or environment. The deployed protocol consists of various numbers of resources. Least dynamic nature increases the scalability ration in network.

The description of Superstring is efficient and hierarchical. Thus, the description contains a description model (This model has simple and easy expression language, query relaxation and reserved elements). A set of primitives are also defined by Superstring in the description language, responsible to allow queries and advertisements. Hence, queries and advertisements are issued to rapidly convey context-awareness to the applications existing in environment. Table II compares eight service discovery protocols such as DEAPspace, DNS-SD, JXTA, RDP, LDAP, CORBA Trader, UDDI, and Superstring. The authors identified that superstring is better service protocol fulfills the functionality and features, i.e. IP based Environment, Other than IP based Environment, Fixed network, Distributed Environment, DynamicNetworks, Scalability, Advertisement, and Query relaxation.

TABLE II. SERVICE DIRECTORY PROTOCOLS AND FUNCTIONALITY

Ref	Protocol	IP based Environment	Other than IP based Environment	Fixed network	Distributed Environment	Dynamic Networks	Scalability	Advertisement	Query relaxation
[5]	DEAPspace	×	×	×	×	×	×	×	×
[6]	DNS-SD	√	×	×	×	×	×	×	×
[7]	JXTA	×	×	×	×	×	×	×	×
[13]	RDP	√	×	√		×	×	×	×
[14]	LDAP	×	×	×	√	×	×	×	×
[17]	CORBA Trader	×	×	×	×	×	×	×	×
[18]	UDDI	×	×	×	×	×	×	×	×
[19]	Superstring	×	×	√	-	√	√	√	√

III. Context Aware Systems

In the research, context may be defined in different ways as per situation or circumstances. The initial definition of context was defined by Schmidt [20] as "A context describes a situation and the environment a device or user is in. A context is identified by a unique name. For each context, a set of features is relevant. For each relevant feature a range of values is determined by the context".

A researcher Dey [21] also defined context in semantic research that "Any information that can be used to characterize the current situation of an entity". Two other researchers Schilit and Theimer categorized the context into four categories such as Computing, user, physical and time context. They also defined as "Computing context: network connectivity, communication costs, communication bandwidth, nearby resources such as printers, displays, and workstations. User context: the user's profile, location, people nearby, the current social situation. Physical context: lighting, noise levels, traffic conditions, and temperature [22]. Time context: time of a day, week, month, and year" [23] [24].

Later on, Schilit, Adams and Want [25] defined context aware system as: "A system is context-aware if it uses context to provide relevant information or services to the user, where relevancy depends on the user's task".

Context aware system or context awareness is backbone of pervasive computing. These systems consists of various functions and some of them are as under [26] and shown in Fig. 1.

Sensing the Context such as location sensing (Indoor and Outdoor). Sensing Low-level Contexts beyond Location such as Time context, nearby objects, Network bandwidth, Orientation, and other low-level contexts. Sensing High-level Contexts such as user's current activity. Sensing Context Changes such as moving of a person from one location to another.

A. Sensing the Context

The basic function of the context aware system is to sense the context. The context may be location, time, user and computing context. Thus, this mechanism is mostly available in these systems to sense and then deliver to application as execute the task as per flow or function [27]. The location is important context to know the user movement from one location to other location. This context varies as user moves and it is easy nowadays to collect or gather user location information because he/she allows the devices to supply their location to applications [1]. User cooperation supports the application or system to be accurate and reliable. If the location sensing uses automatic technique then the system is independent of user and sense the context by applied mechanisms. Such as, an employee is entering in his/her office and press the fingerprint for authentication. His/her location is collected automatically from the system after pressing the fingerprint. The location can be sensed in two modes either indoor or outdoor.

GPS (General positioning System) is best choice for outdoor positioning [28]. The Government of US allowed the GPS signal at 10 to 20 meter range to achieve more accuracy then earlier [29]. Various applications such as automobile navigation systems in computing environment get benefit from the new policy because tiny and inexpensive devices lack the capability of GPS. Bulusu and researchers proposes a connectivity based localization technique [30]. In this proposed technique, the accuracy may achieve within the range of 3 meters (if the reference points are known). However, GPS is not appropriate in indoor applications because the GPS signals have low signal strength in indoor physical space [31]. Thus, the signals penetrated in the buildings and make unreliable and fluctuating readings due to multipath reflections. Moreover, it becomes a challenge to build an ideal indoor service that is inexpensive, scalable and vigorous with lofty update rate of spatial information [32]. That's the reason that most of indoor research projects have their own location tracking systems such as Active Badge, Cyber guide, Shopping Assistant and 3D-iD. Few of them uses IR (Infra Red) and others RF (Radio Frequency) [33] [34].

B. Sensing Low-Level Contexts beyond Location

Location is not only the context but time, nearby objects, network bandwidth, and orientation are also low-level contexts. The contextual information of time is not difficult to achieve. Most of the systems have capability of built-in clocks, few uses timestamp like Active badge. Time-of-day, day, week, month, year, season, time zone, and onward are different forms of contextual information [35]. Nearby objects are also sensed by the contextual systems. It is also easy to find nearby objects via projection from the existing dataset of systems' database because the computing environment records the location of people and objects which became part of that system. Such as Teleporting system and the context-aware Pager are good examples of sensing nearby objects.

Besides location, time and nearby objects, network bandwidth is also a chief component of computing and a significant computational context too[36]. The change in bandwidth is not easy method without the support of system. Hence, system support is necessary to get acknowledge when change in network bandwidth takes place. Few systems at user level like Odyssey system [37] uses API calls and others at kernel level like Congestion Manager uses up calls[38] to measure the bandwidth and notify the changes when occurs.

In addition, few low-level contexts can be embedded with system to measure or know physical contexts like intensity of light, vibration, sound, temperature in indoor environment. The researchers or project team may deploy bi-sensor or multi-sensor prototypes to sense more than one context at a time. TEA project has also deployed multi-sensor prototype in their research to measure multi contexts [39]. The research also notified [40] that user's cooperation is necessary while enhancing the sensing of contextual level in mobile devices because the addition of sensors reduces the user's mobility due to supplementary size and weight. Further, the research is going on to reduce the deployment of sensors within mobile devices in the computing environment as per user needs or specification [26].

C. *Sensing High-Level Contexts*

Like low level context, high level context sensing is also function of context aware systems. Sensing high level context is an emerging challenge for researchers. In this connection, three approaches have been proposed [41] to find user activity. Primary approach is to machine vision while secondary approach is to consult user's calendar and tertiary approach is using Artificial Intelligence (AI) [42] [43]. Primary approach is dependent on computer vision and image processing technology to sense complex social context. Secondary approach uses user's calendar that supports the systems to measure user activities. Tertiary approach with the help of AI recognizes the context by gathering multiple low level context sensors to get contextual information [44].

D. *Sensing Context Changes*

The context aware systems not only sense the low level and high level contexts but also sense the changes occurred in system. Various context-aware applications have feature to notify the contextual changes. Generally, a source monitor is defined to poll the present contextual information and then shares the changes occurred in context services. The services have the ability to publish-register-discover or notify interfaces. Once, the changes take place in context and then the information is discovered or notified by registered client or user from context service. Every context has its own properties the location of an employee changes in office building as per necessity while the location of scanner remains same. Thus, each context has different polling rates [45], [46].

Sometimes, the contextual information of context-aware system does not satisfy the clients or users because the system is tracking the location tracking device worn to client or user in indoor environment but that is kept on table or other place for a short period and forgot to wear. In such cases, it is inappropriate for applications.

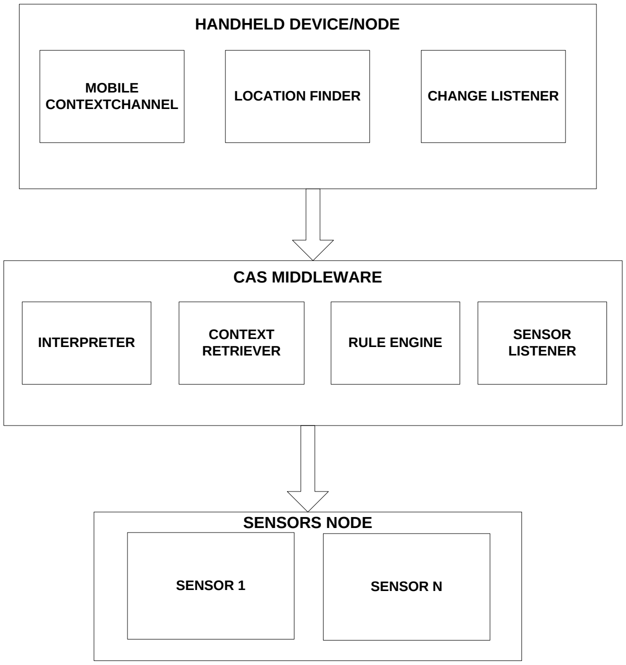


Fig. 1. Architecture of the Context-Aware System [34].

E. *Context-aware and Ontology-Powered Information Visualization Architecture (COIVA)*

COIVA is a context aware system that provides the infrastructure to initiate context-powered services via computing environment and smart devices. The semantic and information of contextual elements is shared in system to measure user needs and then adaptive services are offered where required. The COIVA architecture is comprised of two blocks such as COIVA core, and functional engine [47].

1) The COIVA core is a chief component of this architecture. This component is based on context model, which comprises of general and specific sub-models. General model is further categorized into four sub-components such as user, services, environments and devices. Ontology describes each sub-component to discover main elements of distinctive user-centered intelligent environments.

2) Functional engine is the second component of CIOVA. It comprised of three modules. Initial model is used to collect and unify context from intelligent systems. This model also plays a vital role in implementation and covert gathered information into contextual information. Intermediate model works as expert system to assume context and evade variations. The last module grips the meta-information (A piece of information that describes a contextual feature and its associations with additional information) [47].

IV. Middle Wares

Middlewares play an important role as middle man to control, monitor, supervise and maintain the systems or application smarter in context aware environment. Following are few modern Middleware system services [48].

A. *CAMPUS*

Context-Aware Middleware for Pervasive and Ubiquitous Service (CAMPUS) [49] is an important middleware and used at runtime for automated context-aware adaptation decision making. CAMPUS is dependent of three main technical approaches such as ontology, descriptive reasoning and compositional adaptation. Generally, the middle-wares depend on pre-defined decisions or policies in dynamic environment while CAMPUS has diverse feature to make decisions dynamically and varies with contextual changes [50][51]. The CAMPUS has three tiered architecture, comprises of programming layer, knowledge layer, and decision layer as shown in Fig. 2.

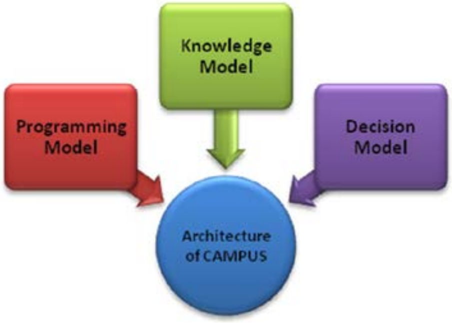


Fig. 2. Three Tiered Architecture of CAMPUS.

The decision layer is basic and an essential layer of CAMPUS. This layer deploys a multi-stage decision model. This model includes screening, selection and preprocessing, to prefer the finest tasklet substitute for a specified task. The automated decisions are taken at this layer, and then forwarded to the second layer named programming layer [52]. The programming layer receives adoptive decisions from decision layer and then reconfigures or constructs context-aware applications by adopting the forwarded instructions from decision layer[53]. The knowledge layer is responsible to represent knowledge semantics and comprises of three models or ontologies such as Service, Context and Tasklet. These ontologies are necessary for CAMPUS in making adoptive decisions. The represented semantics of knowledge are also utmost requirements, may include the required properties of context or existing or run-time context [54].

The CAMPUS initially developed in JAVA SE 1.6 with 1.5.1 Pallet (descriptive reasoning), and JESS 7.1 P2 (Logical Reasoning) [55]. This middleware provides dynamic adoptive decisions and integrate them with context-aware applications. Even though, the security or privacy is emerging issue of CAMPUS and other middleware or context awareness systems [56][57]. The future extension of CAMPUS may adopt security and advanced dynamic adaptation techniques or approaches for context-aware systems.

B. Context-Aware Services Framework (CASF)

CASF is a middleware and it is presented to provide a variety of context aware services. In addition, the architecture of CASF includes a service directory along with capabilities of composition as well. Because, it was noted that a number of context aware systems lacks these both capabilities. Thus, proposed and presented in CASF. This framework also supports automatic service discovery and integration. Hence, it is based on semantic services to achieve service integration, selection and gathering contextual information while CASF separates the services and contextual information of systems.

The main core of CASF architecture plays an essential role. Thus, it is called as Context Mediation Framework (CMF). It also includes three layers Such as Physical sensor, Public context and contextual service as illustrated in Fig. 3. Physical sensor layer is the basic contextual source of CASF, due to recognition nature. It recognizes the sensor data and categories as per service or system requirement. Public context layer includes two sub layers for contextual information and their complexity. First basic sub layer processes and provides only sensor data while the second one provides both sensor and contextual information together to other contextual information providers. Duse to semantic nature, all contextual information is gathered, processed and generated in this layer. With the help of ontology and OWL-S, the web services are constructed. Hence, interoperability and openness are achieved. The last layer of CMF is Context service layer. This layer consumes the contextual information. As, new context aware services are to be produced to fulfill user requests.

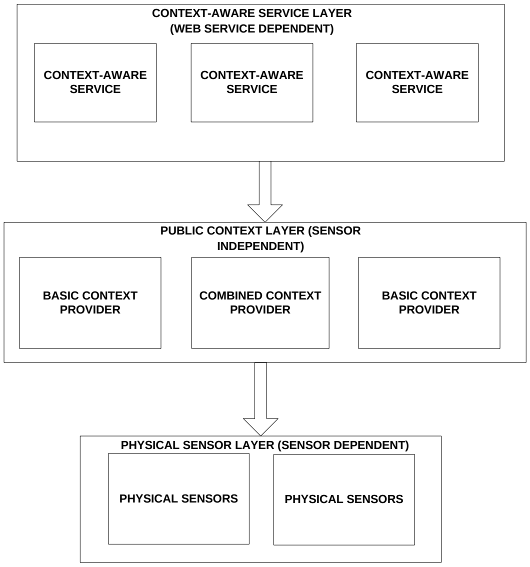


Fig. 3. Features of CASF.

The main uniqueness of CASF is the implementation of the idea of semantic web services. The contextual information is published on the basis of semantic web services. Besides that, this approach becomes reasonable to accomplish automatic discovery and assimilation for contextual data or information. In addition, advanced level protocols and ontologies have to be studied as contextual information is translated into web services. Like SOAP, communication protocols are to be specified or chosen to make communication as effective in between physical sensor and public context layer. However, CASF architecture also lacks the prototype of real environments for context aware services or systems.

C. Semantic Web-based Context Management (SeCoMan)

SeCoMan is proposed to offer a privacy solution in the development of context aware smart applications or services. While ontology is chief component in the development of SeCoMan as the description of entities be modeled as per requirement like obtaining functional knowledge, and specify the policies of context aware. The architecture of SeCoMan is categorized into three layers, including Context Management layer, Application layer, and Plug-in layer. Application layer is first and top level layer in which various applications reside to offer requested services of users. Context Management layer is the heart of the SeCoMan as well as allows the support to contextual applications. In addition, SeCoMan is divided into three different actors along with specific rights. Users, Application supervisor and framework supervisor are defined actors. However, applications having predefined queries are allowed to receive contextual information either indoor or outdoor as well as semantic ontology is also employed to define policies regarding privacy, location and authorization access.

Plug-in layer is third and last layer of SeCoMan framework. This layer is particularly focused on contextual locations and offers contextual information of SeCoMan framework. Plug in works as an independent source of contextual information. Generally, SeCoMan has limited solution to provide only contextual location. Hence, privacy is provided to users. Thus, they easily share their location to receive context aware services. The future of SeCoMan is under research to emerge cloud and distributed computing with SeCoMan architecture. This feature will enhance the capabilities of context aware systems.

D. CoCaMAAL (Cloud oriented Context aware Middleware in Ambient Assisted Living)

Forkan and other colleagues [58] presented and proposed the CoCaMAAL. The main objective of this research is to enhance the capability of biomedical sensors, because they lack processing power. This feature is necessary for AAL to achieve data aggregation and key monitoring. Besides this, cloud and distributed computing are emerged to perform computational tasks or needs. This middleware became an ideal, easy in data gathering and processing. Service Oriented Architecture (SOA) is soul of COCaMAAL because all the functions (such as modeling, adaptation, mapping, service distribution of context and more) are performed. The proposed middleware is hardware based architecture, also comprised of Body Sensor Network (BSN), supervision systems to fulfill user needs or requirements. Moreover, CoCaMAAL comprised of Context Aggregator Provider (CAP), Service Provider (SP), Context aware Middleware (CaM), and Data Visualization Approach (DVA). CAP is an intermediate tool that converts and abstracts high level of contextual information for AAL. In addition, contextual data is categorized either in pre-defined ontology or sensor based data. Besides, various contexts are emerged to execute and complete the required information. CAP is responsible for the execution of whole process. SPs are the general applications, which produce, generate and manipulate the services for user needs. CaM is used to identify assistive services for collected context and associated actions. CaM is also an essential tool of CoCaMAAL due to execution of key functions such as mapping, management, storage, and retrieval. DVA also plays an important role in utilization of user interfaces. The author developed the architectural prototype in Java language. The implementation and experiments have been done to measure response time, influence level while increase in context takes place. The results illustrated that CoCaMAAL is efficient and effective for AAL environment. The adaptation od computing technologies is also novelity in this research or prototype. However, Reliability, conflict among contextual systems and privacy are also under research and needs improvements in highlighted issues.

E. Big Data for Context Aware Monitoring (BDCaM)

BDCaM is an extended version of CoCaMAAL and proposed by [59]. BDCam is advanced middleware and used as supervising tool for Context-aware systems. This middleware demonstrates a supplementary feature such as personalized knowledge discovery as compared to CoCaMAAL. In this feature, the knowledge is learnt from collected data, which are anomalies of specific patient. The adoption technique and methodology are different than COCaMAAL [60]. Both are essential in decision making while contextual data is collected from context-aware systems. Correlations and Supervised learning are two approaches used to perform functions of BDCaM. Initially, Correlation takes place between the attributes of contextual information and values of threshold. Map Reduce Apriori Algorithm (MRAA) is applied to generate associations of patient-tailored [61]. The generated rules are used by supervised learning to manipulate collected contextual data.

The working architecture of BDCaM is also split into different distributed or cloud based components. Like AAL, CA, CP, CMS, SP and more. A prototype was also developed for health monitoring systems to measure the functionality of middleware. The results proved that the middleware has detection efficiency among anomalies. Security and privacy is also a challenge for context-aware systems [62]. The emerging of this middleware to other domains is also under research and not yet been suited or implemented.

F. FlexRFID

FlexRFID is modern and advanced middleware from discussed above architectures [63]. The aim of this architecture is to offer a policy-based solution in the development and implementation of context aware systems or applications and emerging diverse nodes. FlexRFID is multi-layered middleware, adopts Ponder as PSL (Policy Specification Language) as well as consists of Device Abstraction Layer (DAL), Business Event and Data Processing Layer (BEDPL), Business Rule Layer (BRL), and Application Abstraction Layer (AAL) layers. DAL abstracts the interactive activities among the devices, nodes, and communication medium. BEDPL offers Contextual Information Management (CIS) Such as aggregation, revolution and broadcasting). BRL copes with policy based operations), and at last, AAL allows communications amid applications and the FlexRFID.

According to literature, it is claimed [64] that FlexRFID offers filtration, grouping, integrity, removal of duplications. Hence, it can be said that FlexRFID is an enabled solution for synchronized communication among emerged technologies of middleware. In addition, Policy enforcement is also feature that differs FlexRFID with other middleware. Various policies such as abstract policy, system policy, ensure security, access control, and other customized services are defined for architecture. The authors have taken two scenarios for experimental work during their research, the results illustrated that response time and volume of policies are directly proportional as well as dependent to each other. If one increases and second one results an increase also. Recent version of FlexRFID only provides necessary privacy mechanisms to specify policies of access control. Application authentication, privacy at sensor nodes and tags, integration of FlexRFID with other distributed or cloud services are to be improved at advanced level for specific applications.

Table III compares six middlewares such as CAMPUS, CASF, SeCoMan, CoCaMAAL, BDCaM, and FlexRFID having Architectural style, Context abstraction/Reasoning, Context awareness level, Contextual adaptation approaches , decision making, and Programming model. CAMPUS is found to be an effective and dynamic middleware, which is suitable for Semantic, Sensing, Parameter adaptation, and Contextual Reconfiguration.

TABLE III. COMPARISON OF MIDDLEWARES

Ref	Middleware	Architectural style		Context abstraction/Reasoning				Context awareness level		Contextual adaptation approaches			Decision making	Programming model
		Layered	Distributed	Key value	Ontology	Markup	Rule	Medium	Semantic	Sensing	Parameter adaptation	Contextual Reconfiguration		
[65], [66]	CAMPUS	√	×	×	√	×	×	×	√	√	√	√	Dynamic	Semantic-based DL
[57] [67]	CASF	√	×	×	√	×	×	×	√	√	√	×		ontology and OWL-S
[68]	SeCoMan	√	×	×	√	×	×	×	√	√	√	×		Ontology / Semantic
[58]	CoCaMAAL	×	√	×	√	×	√	×	√	√	√	√	Dynamic	
[59], [69]	BDCaM	×	√	×	√	×	×	×	×	√	√	√	Dynamic	
[70], [71]	FlexRFID	√	√	×	×	√	√	√	×	√	×	×	Static	Policy based

V. DISCUSSIONS

The discussions section provides the comparative analysis of three major features and services of pervasive computing field. Table I compares five JXTA protocols such as PDP, CSDP, HSDP, PRP, and RP having Peer Group, Non-Peer Group, Detailed discovery information, Rendezvous Peers, Semantic definition, Scalability, Resource advertisement, and Resource discovery. PDP has four maximum features capability such as Peer Group, Non-Peer Group, Resource advertisement, and Resource discovery as compared to other protocols. Hence, PDP is favorable than others. Table II compares eight service discovery protocols such as DEAPspace, DNS-SD, JXTA, RDP, LDAP, CORBA Trader, UDDI, and Superstring. The authors identified that superstring is better service protocol fulfills the functionality and features, i.e. IP based Environment, Other than IP based Environment, Fixed network, Distributed Environment, DynamicNetworks, Scalability, Advertisement, and Query relaxation. Table III compares six middlewares such as CAMPUS, CASF, SeCoMan, CoCaMAAL, BDCaM, and FlexRFID having Architectural style, Context abstraction/Reasoning, Context awareness level, Contextual adaptation approaches, decision making, and Programming model. CAMPUS is found to be an effective and dynamic middleware, which is suitable for Semantic, Sensing, Parameter adaptation, and Contextual Reconfiguration.

VI. CONCLUSION

Latest development in smart systems, context aware systems and middleware made real time environments smarter and efficient for users. Emergences of sensors technology have grown user interaction with systems and technology. Likewise, new policies, rules and procedure are being defined to enhance the capabilities and features of smart system. Hence, physical objects are moved to smart objects. In addition, real time data and intelligent data combined to achieve more accuracy and efficiency in such systems, connected components or nodes can be identified via embedded systems. These systems may communicate to each other via distributed systems or infrastructure. According to Cisco IBSG, the IoT world will include more than 50 billion objects in 2020. In this paper, we presented service discovery protocols and functionality is compared in Section 2, Section 3 overviews context aware systems, their features and discusses few context-aware systems. In addition, middlewares are also discussed and compared to identify most suitable one in Section 4. And discussion is presented in Section 5.

REFERENCES

[1] J. H. Awan, S. A. Memon, N. A. Memon, R. Shah, Z. Bhutto, and R. A. Khan, “Conceptual Model for WWBAN (Wearable Wireless Body Area Network),” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 1, pp. 377–381, 2017.

[2] U. Naseem, S. K. Khan, M. Farasat, and F. Ali, “Abusive Language Detection: A Comprehensive Review,” Indian J. Sci. Technol., vol. 12, no. 45, pp. 1–13, 2019.

[3] U. Naseem, “Hybrid Words Representation for the classification of low quality text,” 2020.

[4] M. Rosic, S. Mladenovic, and L. Borojevic, “Information system user interface complexity,” in Symposium of the Austrian HCI and Usability Engineering Group, 2010, pp. 509–512.

[5] J. D. Cormie, A. K. Fischman, and A. H. Vermeulen, “Dynamic application instance discovery and state management within a distributed system.” Google Patents, 2018.

[6] R. Droms and T. P. Donahue, “Visibility control for domain name system service discovery.” Google Patents, 2018.

[7] M. Amoretti and F. Zanichelli, “P2P-PL: A pattern language to design efficient and robust peer-to-peer systems,” Peer-to-Peer Netw. Appl., vol. 11, no. 3, pp. 518–547, 2018.

[8] S. Mallik, R. Palanki, D. P. Malladi, and N. Bhushan, “Hybrid modes for peer discovery.” Google Patents, 2019.

[9] Y. Amishav, E. J. Barkie, O. Dubovsky, and B. L. Fletcher, “Location-based domain name system service discovery.” Google Patents, 2018.

[10] S. Abdellatif, O. Tibermacine, and A. Bachir, “Service Discovery in the Internet of Things: A Survey,” in International Symposium on Modelling and Implementation of Complex Systems, 2018, pp. 60–74.

[11] F. Battaglia and L. Lo Bello, “A novel JXTA-based architecture for implementing heterogenous Networks of Things,” Comput. Commun., vol. 116, pp. 35–62, 2018.

[12] S. Kadam, D. Prabhu, N. Rathi, P. Chaki, and G. S. Kasbekar, “Exploiting group structure in MAC protocol design for multichannel Ad Hoc cognitive radio networks,” IEEE Trans. Veh. Technol., vol. 68, no. 1, pp. 893–907, 2018.

[13] A. T. Fausak, O. Rombakh, C. D. Robison, C. A. Andrews, and others, “System and method for remote access to a personal computer as a service using a remote desktop protocol and windows hello support.” Google Patents, 2019.

[14] M. A. Thakur, S. G. Bari, R. Deshmukh, and S. Auty, “A Survey of Directory and Database Protocols for Data Extraction,” in Information and Communication Technology for Sustainable Development, Springer, 2020, pp. 315–325.

[15] S. K. Khan, M. Farasat, U. Naseem, and F. Ali, “Performance Evaluation of Next-Generation Wireless (5G) UAV Relay,” Wirel. Pers. Commun., pp. 1–16, 2020.

[16] S. K. Khan, M. Farasat, U. Naseem, and F. Ali, “Link-level Performance Modelling for Next-Generation UAV Relay with Millimetre-Wave Simultaneously in Access and Backhaul,” Indian J. Sci. Technol., vol. 12, no. 39, pp. 1–9, 2019.

[17] M. Zaryouli and M. Ezziyyani, “Dynamic Adaptation and Automatic Execution of Services According to Ubiquitous Computing,” in International Conference on Advanced Intelligent Systems for Sustainable Development, 2018, pp. 984–990.

[18] G. S. Kumar, A. M. Nancy, S. Soral, and A. Shrivastava, “Query Optimization in Universal Description Discovery and Integration for Effective Web Services Discovery,” J. Comput. Theor. Nanosci., vol. 15, no. 6–7, pp. 2420–2424, 2018.

[19] J. Dator, “Governing the futures: dream or survival societies?,” in Jim Dator: A Noticer in Time, Springer, 2019, pp. 395–408.

[20] A. Schmidt, K. A. Aidoo, A. Takaluoma, U. Tuomela, K. Van Laerhoven, and W. de Velde, “Advanced interaction in context,” in International Symposium on Handheld and Ubiquitous Computing, 1999, pp. 89–101.

[21] C. Shin, J.-H. Hong, and A. K. Dey, “Understanding and prediction of mobile application usage for smart phones,” in Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 2012, pp. 173–182.

[22] J. H. Awan, U. Naseem, and S. K. Khan, “A proposed framework for the security of Financial Systems,” Indian J. Sci. Technol., vol. 12, no. 21, pp. 1–8, 2019.

[23] B. N. Schilit, D. M. Hilbert, and J. Trevor, “Context-aware communication,” IEEE Wirel. Commun., vol. 9, no. 5, pp. 46–54, 2002.

[24] P. Kremer, C. Elshaug, E. Leslie, J. W. Toumbourou, G. C. Patton, and J. Williams, “Physical activity, leisure-time screen use and depression among children and young adolescents,” J. Sci. Med. Sport, vol. 17, no. 2, pp. 183–187, 2014.

[25] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, “Context aware computing for the internet of things: A survey,” IEEE Commun. Surv. tutorials, vol. 16, no. 1, pp. 414–454, 2014.

[26] A. Botta, W. De Donato, V. Persico, and A. Pescapé, “Integration of cloud computing and internet of things: a survey,” Futur. Gener. Comput. Syst., vol. 56, pp. 684–700, 2016.

[27] J. H. Awan, S. Memon, R. A. Khan, A. Q. Noonari, Z. Hussain, and M. Usman, “Security strategies to overcome cyber measures, factors and barriers,” Eng. Sci. Technol. Int. Res. J, vol. 1, no. 1, pp. 51–58, 2017.

[28] J. Awan, S. Memon, Z. Bhutto, and W. Awan, “Jamshoro City Based Location Service (JCLS),” J. Inf. Commun. Technol., vol. 10, no. 1, pp. 34–40, 2017.

[29] J. M. Curry, “A Web of Drones: A 2040 Strategy to Reduce the United States Dependance on Space Based Capabilities,” 2015.

[30] L. Yang, J. Cao, W. Zhu, and S. Tang, “Accurate and efficient object tracking based on passive RFID,” IEEE Trans. Mob. Comput., vol. 14, no. 11, pp. 2188–2200, 2015.

[31] J. H. Awan, A. Q. Noonari, R. A. Khan, K. H. Muhammadani, A. Shoro, and W. Ahmad, “Smart Phone Based Tracking System for the Billing Accuracy in Electricity Meters,” Eng. Sci. Technol. Int. Res. J., vol. 1, no. 3, pp. 33–38, 2017.

[32] S. Alletto, R. Cucchiara, G. Del Fiore, L. Mainetti, V. Mighali, L. Patrono, and G. Serra, “An Indoor Location-Aware System for an IoT-Based Smart Museum.,” IEEE Internet Things J., vol. 3, no. 2, pp. 244–253, 2016.

[33] A. Pandharipande and D. Caicedo, “Smart indoor lighting systems with luminaire-based sensing: A review of lighting control approaches,” Energy Build., vol. 104, pp. 369–377, 2015.

[34] M. R. Rahimi, J. Ren, C. H. Liu, A. V Vasilakos, and N. Venkatasubramanian, “Mobile cloud computing: A survey, state of art and future directions,” Mob. Networks Appl., vol. 19, no. 2, pp. 133–143, 2014.

[35] A. Etkin, C. Büchel, and J. J. Gross, “The neural bases of emotion regulation,” Nat. Rev. Neurosci., vol. 16, no. 11, p. 693, 2015.

[36] S. A. Memon and J. H. Awan, “Transformation towards Cyber Democracy: A study on Contemporary Policies, Practices and Adoption Challenges for Pakistan,” in Handbook of Cyber-Development, Cyber-Democracy and Cyber-Defense, 2017, pp. 1–20.

[37] Y. J. Lee, J. Trevathan, I. Atkinson, and W. Read, “An intelligent agent system for managing heterogeneous sensors in dispersed and disparate wireless sensor network,” Int. J. Sens. Networks, vol. 27, no. 3, pp. 149–162, 2018.

[38] A. Mankee, M. S. Safiullah, and H. Raza, “Intelligent air-interface congestion manager.” Google Patents, 2015.

[39] M. Umlauft, C. Raffelsberger, A. Kercek, A. Almer, T. Schnabel, P. Luley, and S. Ladstaetter, “A communication and multi-sensor solution to support dynamic generation of a situational picture,” in Information and Communication Technologies for Disaster Management (ICT-DM), 2016 3rd International Conference on, 2016, pp. 1–7.

[40] Ö. Yürür, C. H. Liu, Z. Sheng, V. C. M. Leung, W. Moreno, and K. K. Leung, “Context-awareness for mobile sensing: A survey and future directions,” IEEE Commun. Surv. Tutorials, vol. 18, no. 1, pp. 68–93, 2016.

[41] D. Ferreira, V. Kostakos, and A. K. Dey, “AWARE: mobile context instrumentation framework,” Front. ICT, vol. 2, p. 6, 2015.

[42] J. H. Awan, S. Memon, S. M. Pathan, M. Usman, R. A. Khan, S. Abbasi, A. Q. Noonari, and Z. Hussain, “A user friendly security framework for the protection of confidential information,” Int. J. Comput. Sci. Netw. Secur., vol. 17, no. 04, pp. 215–223, 2017.

[43] U. Naseem, I. Razzak, and I. A. Hameed, “Deep Context-Aware Embedding for Abusive and Hate Speech detection on Twitter,” Aust. J. Intell. Inf. Process. Syst., p. 69, 2019.

[44] L. Snidaro, J. García, and J. Llinas, “Context-based information fusion: a survey and discussion,” Inf. Fusion, vol. 25, pp. 16–31, 2015.

[45] M. S. Birrittella, M. Debbage, R. Huggahalli, J. Kunz, T. Lovett, T. Rimmer, K. D. Underwood, and R. C. Zak, “Intel{\textregistered} Omni-path architecture: Enabling scalable, high performance fabrics,” in High-Performance Interconnects (HOTI), 2015 IEEE 23rd Annual Symposium on, 2015, pp. 1–9.

[46] E. Ibarra, A. Antonopoulos, E. Kartsakli, and C. Verikoukis, “HEH-BMAC: Hybrid polling MAC protocol for WBANs operated by human energy harvesting,” Telecommun. Syst., vol. 58, no. 2, pp. 111–124, 2015.

[47] N. Nebhani, J.-C. Lapayre, and M. S. Bouhlel, “Ontology traceability for the adaptation of services in pervasive environment,” in Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on, 2016, pp. 4543–4548.

[48] G. Liang and J. Cao, “Social context-aware middleware: A survey,” Pervasive Mob. Comput., vol. 17, pp. 207–219, 2015.

[49] X. Li, M. Eckert, J.-F. Martinez, and G. Rubio, “Context aware middleware architectures: survey and challenges,” Sensors, vol. 15, no. 8, pp. 20570–20607, 2015.

[50] G.-Z. Liu, F.-R. Kuo, Y.-R. Shi, and Y.-W. Chen, “Dedicated design and usability of a context-aware ubiquitous learning environment for developing receptive language skills: a case study,” Int. J. Mob. Learn. Organ., vol. 9, no. 1, pp. 49–65, 2015.

[51] M. Satyanarayanan, “Pervasive computing: Vision and challenges,” IEEE Pers. Commun., vol. 8, no. 4, pp. 10–17, 2001.

[52] M. Conti, A. Passarella, and S. K. Das, “The Internet of People (IoP): A new wave in pervasive mobile computing,” Pervasive Mob. Comput., vol. 41, pp. 1–27, 2017.

[53] Y.-H. Feng, T.-H. Teng, and A.-H. Tan, “Modelling situation awareness for Context-aware Decision Support,” Expert Syst. Appl., vol. 36, no. 1, pp. 455–463, 2009.

[54] A. I. Maarala, X. Su, and J. Riekki, “Semantic reasoning for context-aware Internet of Things applications,” IEEE Internet Things J., vol. 4, no. 2, pp. 461–473, 2017.

[55] I. Elhart, M. Langheinrich, N. Memarovic, and E. Rubegni, “A good balance of costs and benefits: convincing a university administration to support the installation of an interactive multi-application display system on campus,” in Proceedings of the 5th ACM International Symposium on Pervasive Displays, 2016, pp. 197–203.

[56] J. H. Awan, S. Memon, M. Shah, and F. H. Awan, “Security of eGovernment Services and Challenges in Pakistan,” in SAI Computing, 2016, pp. 1082–1085.

[57] X. Li, M. Eckert, J. Martinez, and G. Rubio, “Context Aware Middleware Architectures: Survey and Challenges,” Sensors, vol. 15, pp. 20570–20607, 2015.

[58] A. Forkan, I. Khalil, and Z. Tari, “CoCaMAAL: A cloud-oriented context-aware middleware in ambient assisted living,” Futur. Gener. Comput. Syst., vol. 35, pp. 114–127, 2014.

[59] A. Forkan, I. Khalil, A. Ibaida, and Z. Tari, “Bdcam: Big data for context-aware monitoring-a personalized knowledge discovery framework for assisted healthcare,” IEEE Trans. cloud Comput., 2015.

[60] A. Forkan, I. Khalil, and Z. Tari, “CoCaMAAL : A cloud-oriented context- aware middleware in ambient assisted living,” Futur. Gener. Comput. Syst., vol. 35, pp. 114–127, 2014.

[61] G. Wu, H. Zhang, M. Qiu, Z. Ming, J. Li, and X. Qin, “A decentralized approach for mining event correlations in distributed system monitoring,” J. Parallel Distrib. Comput., vol. 73, no. 3, pp. 330–340, 2013.

[62] J. Awan and S. Memon, “Threats of Cyber Security and Challenges for Pakistan,” in 11th International Conference on Cyber Warfare and Security: ICCWS - 2016, Boston USA, 2016, p. 425.

[63] M. A. El Khaddar and M. Boulmalf, “Smartphone: the ultimate IoT and IoE device,” in Smartphones from an Applied Research Perspective, InTech, 2017.

[64] M. A. El Khaddar, M. Chraibi, H. Harroud, M. Boulmalf, M. Elkoutbi, and A. Maach, “A policy-based middleware for context-aware pervasive computing,” Int. J. Pervasive Comput. Commun., vol. 11, no. 1, pp. 43–68, 2015.

[65] E. J. Y. Wei and A. T. S. Chan, “CAMPUS: A middleware for automated context-aware adaptation decision making at run time,” Pervasive Mob. Comput., vol. 9, no. 1, pp. 35–56, 2013.

[66] K. Hisazumi, T. Nakanishi, T. Kitasuka, A. Fukuda, and others, “CAMPUS: A context-aware middleware,” in The 2nd CREST Workshop on Advanced Computing and Communicating Techniques for Wearable Information Playing, 2003.

[67] Á. Fides-Valero, M. Freddi, F. Furfari, and M.-R. Tazari, “The PERSONA framework for supporting context-awareness in open distributed systems,” in European Conference on Ambient Intelligence, 2008, pp. 91–108.

[68] A. H. Celdrán, F. J. G. Clemente, M. G. Pérez, and G. M. Pérez, “SeCoMan: A Semantic-Aware Policy Framework for Developing Privacy-Preserving and Context-Aware Smart Applications.,” IEEE Syst. J., vol. 10, no. 3, pp. 1111–1124, 2016.

[69] A. R. M. Forkan and W. Hu, “A context-aware, predictive and protective approach for wellness monitoring of cardiac patients,” in Computing in Cardiology Conference (CinC), 2016, 2016, pp. 369–372.

[70] M. E. Ajana, H. Harroud, M. Boulmalf, and H. Hamam, “FlexRFID: A flexible middleware for RFID applications development,” in Wireless and Optical Communications Networks, 2009. WOCN’09. IFIP International Conference on, 2009, pp. 1–5.

[71] A. Sengupta and S. Z. Schiller, “FlexRFID: A design, development and deployment framework for RFID-based business applications,” Inf. Syst. Front., vol. 12, no. 5, pp. 551–562, 2010.

Classification of Multiple Sclerosis Disease using Cumulative Histogram

Menna Safwat[1], Fahmi Khalifa[2], Hossam El-Din Moustafa[3]
Electronics and Communications Engineering Dept. Mansoura University Mansoura, Egypt[1, 2]
Electronics and Communications Engineering Dept. Faculty of Engineering, Mansoura University Mansoura, Egypt[3]

***Abstract*—Multiple sclerosis (MS) is a chronic disease that affects different body parts including the brain. Detection and classification of MS brain lesions is of immense importance to physicians for the administration of appropriate treatment. Thus, this study investigates an automated framework for the diagnoses and classification of MS lesions in brain using magnetic resonance imaging (MRI). First, the MRI images format converted from dicom images of each patient into TIF format as MS lesion appears in white matter (WM) obviously. This is followed by a brain tissue segmentation using a k-nearest neighbor classifier. Then, cumulative empirical distributions or cumulative histograms (CH) of the segmented lesions are estimated along with other texture/statistical features that work on the difference between the intensity of MS lesions and its surrounding tissues. Finally, these CDFs are fused with and the statistical features for the classification of MS using K mean classifiers. Experiments are conducted, using transverse T2-weighted MR brain scans from 20 patients that are highly sensitive in detecting MS plaques, with gold standard classification obtained by an experienced MS. By comparing the evaluated performance with statistical features, our proposed fusion scored the highest accuracy with 98% and a false-positive rate of 1%.**

Keywords—Cumulative Histogram (CH); Magnetic Resonance Image (MRI); Multiple Sclerosis (MS); White Matter (WM)

I. Introduction

Multiple sclerosis (MS) is an autoimmune inflammatory chronic disease of the central nervous that appears in the white matter (WM) [1] [2]. MS is an illness that can influence on the optic nerves in your eyes, brain and spinal cord. It is the reasons for defect in balance, muscle control, vision and other basic body functions [3]. In the detection of MS lesions Magnetic resonance imaging (MRI) became the most precise scanning. The accurate manual evaluation of each lesion in MR image would be a difficult, challenging task, subjective and low re-productivity. Automatic segmentation offers an attractive alternative to manual segmentation, a process that still takes up time with intra- and inter-expert variability. However, the progression of the MS lesions shows considerable variability and MS lesions present temporal changes in shape, location, and area between patients and even for the same patient. The research work included in this thesis aims at creating a robust technique for the automatic segmentation of MR brain damage. This makes the automatic segmentation of MS lesions a challenging problem, so in [4] they focused on differentiating between active MS and cold-spot lesion from brain MRI. MRI is a cornerstone in current diagnosis standard by enabling to show the distribution of WM lesions in space and time at high specificity and sensitivity [5]. The challenge was in identification of MS in MR Images since the lesions have different size, shape and also different locations with anatomical variability [6].

Based on a study presented by [7], techniques can be divided into deformation-based method and intensity-based method. In the intensity-based technique, pixels will be compared in different successive scans. While in deformation-based approaches, the non-rigid registration between successive scans will obtain the deformation features, It provides a discrete local displacement field that defines the deformation occurring between two images.

Texture feature analysis considered as an alternative quantitative method to classify contrast enhanced Multi sclerosis plaques [8], classify several sub-areas within lesions undergoing ‘active’ demyelination [9], differentiate white matter disease plaques from cerebral microangiopathies [10], and differentiate between relapsing and remitting Multi sclerosis plaques [11]. The common texture analysis approaches are statistical and Spectral approaches in classifying MS. Statistical texture analysis used statistical parameters to characterize texture features of the image. It is divided into first order (standard deviation, variance, mean, entropy, CH, skewness, ...) that provides a general evaluation of pixel distribution and is relatively clear, second order like runlength matrix (RLM) and Gray-level co-occurrence matrix (GLCM). While spectral analysis is the analysis of pixels pattern that make a unique texture and frequency distribution of an image. This approach has Fourier transform, Wavelet transform and Stockwell transform [12].

Texture is the number of operations performed to specifics spatial variations of gray-level pixels in MRI. Texture analysis has the possibility to support early detection of MS as it detects the slight difference in tissues. In our study we used texture analysis of MS lesion and normal tissue then we combined a group of texture feature to differentiate between tissues.

II. Material

This study used MRI data from 20 patients (five men and fifteen women) with mean age of 31±15 years. The data was collected from ELMOGY and ELRAKHAWY Radiology Medical Centers using gradient-echo T2 imaging using a 3T MRI scanner (Signa Explorer MR) with a phased-array torso surface coil. Approximately 70 axial slices were collected from each patient at slice thickness: 5 mm.

III. Research Method

Our framework is represented by a flowchart as shown in Fig. 1. The proposed analysis pipeline is considered as supervised approach that begins with registration of brain MR images. This is followed by a set of pre-processing steps using a bank of filters to remove the skull. Then a segmentation process is utilized using k-nearest neighbours (KNN) algorithm to segment the brain into three tissue parts: WM, the grey matter (GM) and MS lesion for each slice. Finally, a classification step is performed using a cumulative histogram (CH) feature to differentiate between MS and other tissues. Evaluation is performed using performance parameters like accuracy, sensitivity and specificity.

A. Preprocessing

The segmentation process is not easy to apply on MR images as images have a changing parameters, partial voluming, noise, interfering intensities, blurred edges, motion, echoes, gradients normal anatomical variations and susceptibility artifacts [13,14]. In the image we need to neglect all the unwanted parts that have same gray scale, we can get rid of noise as it appears mostly outside WM of the brain while the lesion is in this region [15]. As it shown in Fig. 2, the preprocessing stage is divided into two steps. First, remove artifacts from images that lead to inaccurate segmentation. However, from the image processing viewpoint, it is common to simplify all these problems [16]. Second, skull stripping is another important pre-processing step since fat, skull, skin and other non-brain tissues may cause misclassifications [17].

B. Segmentation using Knearst Neighbors

K-nearest neighbor segmentation [18][19][20] is a statistical pattern recognition technique, that works on distinguishing between different samples (e.g. WM,GM, and CSF) by comparing resemble values in a defined feature space with values of samples in a learning set. A new pixel is classified by comparing the KNN samples to the nearest pixels according to a closeness measure, usually the Euclidian distance [21]. Commonly, the class that repeated mostly in the K-learning samples is assigned to this pixel. As it is shown in Fig. 3, KNN segment the MR image into WM, GM and output that will be classified using statistical features.

C. Texture Analysis

In this study we extract texture features from all normal WM regions and MS plaques. Textural analysis has two categories that depending on matrix or vector used in features calculation. In our proposed framework, we utilized CH-based features and compare the output accuracy with other features as statistics features. We have tested different statistical classification techniques in classifying the texture features of MS plaques and normal tissues and the best performance techniques have been selected.

For calculating all statistical features, we have expressed a given image of size $m \times n$ as a function $g(a, b)$ with two quantities x and y, where $a = 0,1, \dots, m-1$ and $b = 0,1, \dots, n-1$. The function $g(a, b)$ can be any amount $j = 0,1, \dots, Z-1$. In the image, Z will be the total number of intensity levels. According to histogram-based features $G(j)$ is the intensity level histogram, so the number of pixels in whole image have this intensity value.

$$G(j) = \sum_{a=0}^{m-1}\sum_{b=0}^{n-1}\delta(g(a,b),j) : \delta(j,i) = \begin{cases} 1. i = j \\ 0. i \neq j \end{cases} \tag{1}$$

Histogram is presenting simply the image as statistical data. The approximate probability density of the existence of the intensity levels $P(j)$ is equals to the histogram $G(j)$ dividing by the total number of pixels in the image.

$$P(j) = \frac{G(j)}{mn} \tag{2}$$

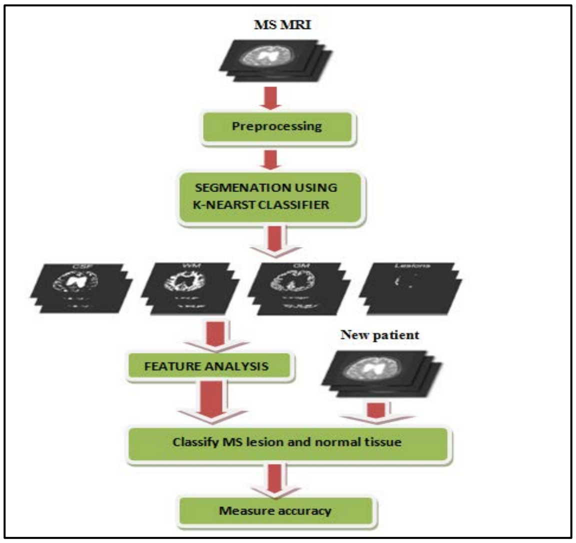


Fig. 1. Flowchart of Segmentation Process.

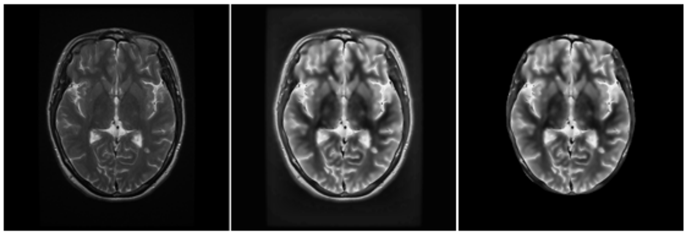

(a) (b) (c)

Fig. 2. Preprocessing Steps: (a) Original MRI Image, (b) Image after Filtration and (c) After Skull Off.

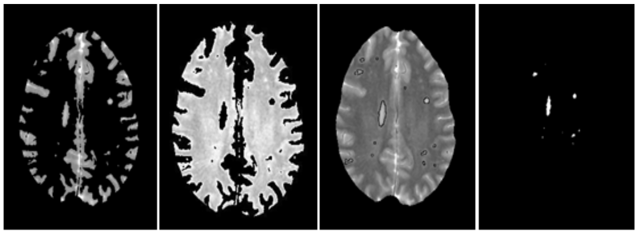

(a) (b) (c) (d)

Fig. 3. Segmentation using K-Nearest Neighbor. (a) Gray Matter (GM), (b) White Matter (WM),(c) Edge Detection using canny Edge, and (d) The Output of Segmentation Process.

The normalized histogram is obtained for the following texture features:

- Median: is the middle value of the normalized histogram vector that have even number of i, in formal way:

$$\mathrm{MED} = \frac{\left[\mathrm{P}\left(\frac{\mathrm{j}}{2}\right)+\mathrm{P}\left(\frac{\mathrm{j}+1}{2}\right)\right]}{2} \qquad (3)$$

- Standard Deviation: is a measure of the intensity levels dispersion from its mean, in formal way:

$$\mathrm{STD} = \sqrt{\sum_{\mathrm{j}=0}^{\mathrm{Z}-1}(\mathrm{j}-\mu)^2\mathrm{P}(\mathrm{j})} \qquad (4)$$

- Skewness: is a measure of the asymmetry of the probability distribution of gray level intensity about its mean, in formal way:

$$\mathrm{SKW} = \frac{1}{\sigma^3}\sum_{\mathrm{j}=0}^{\mathrm{Z}-1}(\mathrm{j}-\mu)^3\mathrm{P}(\mathrm{j}) \qquad (5)$$

- Cumulative histogram (CH): is the mapping that counts the cumulative number of pixel intensity values in all of the bins up to the current bin, this matrix consists of a normalized histogram of gray image intensity. We can calculate the cumulative sum by the following equation where Q represents the total number of intensity levels in the image:

$$\mathrm{CH} = \sum_{\mathrm{j}=1}^{\mathrm{Q}}\mathrm{P}(\mathrm{j}) \qquad (6)$$

- Entropy: is a measure of the information carried by the probability density P(j), in formal way:

$$\mathrm{ENT} = \sum_{\mathrm{j}-1}^{\mathrm{Z}-1} -\mathrm{P}(\mathrm{j})\log\mathrm{P}(\mathrm{j}) \qquad (7)$$

D. *K-means Classifier*

For in data mining and statistics K-means clustering is used for clustering analysis. The purpose of this algorithm is to analyse data into groups of clustering observations represented by variable K. The total number of specified clusters is represented by symbol K. Based on the provided features the data of each pixel is assigned to one of K groups. The objective of this algorithm is to minimize the value between the centroid of the cluster and the given result by repeatedly attaching the result to any cluster and ends when the measured distance is the smallest value.

IV. Results

The proposed framework has been texted on the collected data from the 20 subjects as described in Section II. Features were extracted from T2-W MR images using five features: CH, MED, STD, SKW and ENT. We obtain the evaluation of the performance by comparing the results of each classifier with the ground truth classification of the experienced neurologist. As demonstrated by our experiments, CH feature showed significant correlation with the detection of MS. It is worth mentioning that MS plaques have the highest gray level than normal WM as shown in Fig. 4.

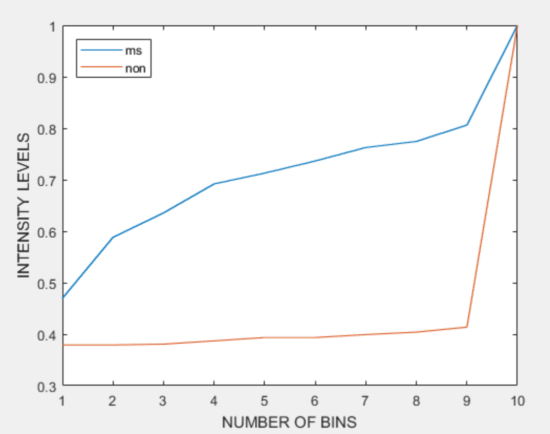


Fig. 4. Cumulative Intensity of Multiple Sclerosis Lesion (MS) and Normal Tissue.

For knowing more data about these features we obtained the evaluation of the performance, we have four basic parameters for each feature, false negatives (FN), false positives (FP), true negative(TN) and true positives (TP). The measurement parameters considered as the following:

- True Negative (TN) is the number of normal pixels that have been correctly classified.
- True Positive (TP) is the number of MS pixels that have been correctly classified.
- False Negative (FN) is the number of normal pixels that have been misclassified.
- False Positive (FP) is the number of MS pixels that have been misclassified.

We can perform these parameters in the output of CH and other classifiers as in Fig. 5 and 6. Where red, yellow, blue and green colors represents the FN, TP, TN and FP, respectively. Also in Fig. 7 shows the boxplot ranges for each classifier of MS and normal tissue.

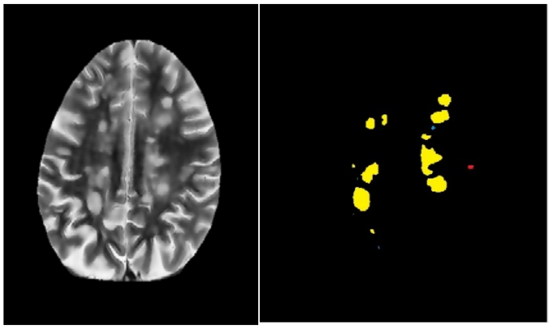
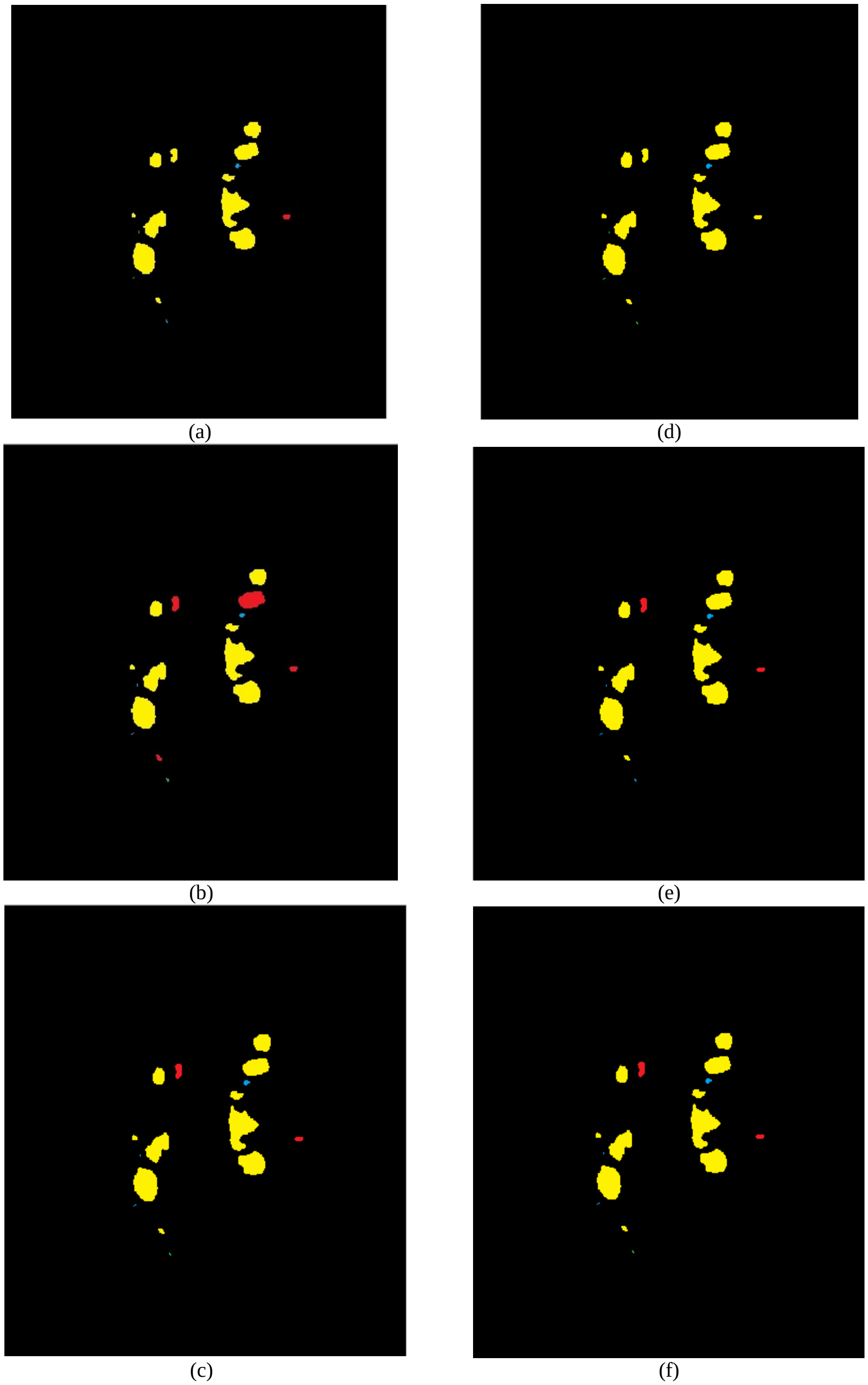


Fig. 5. Cumulative Histogram Classification Output using Color-Code Representations for FP (Green), FN (Red), TN (Blue), and TP (Yellow).



(a)

(d)

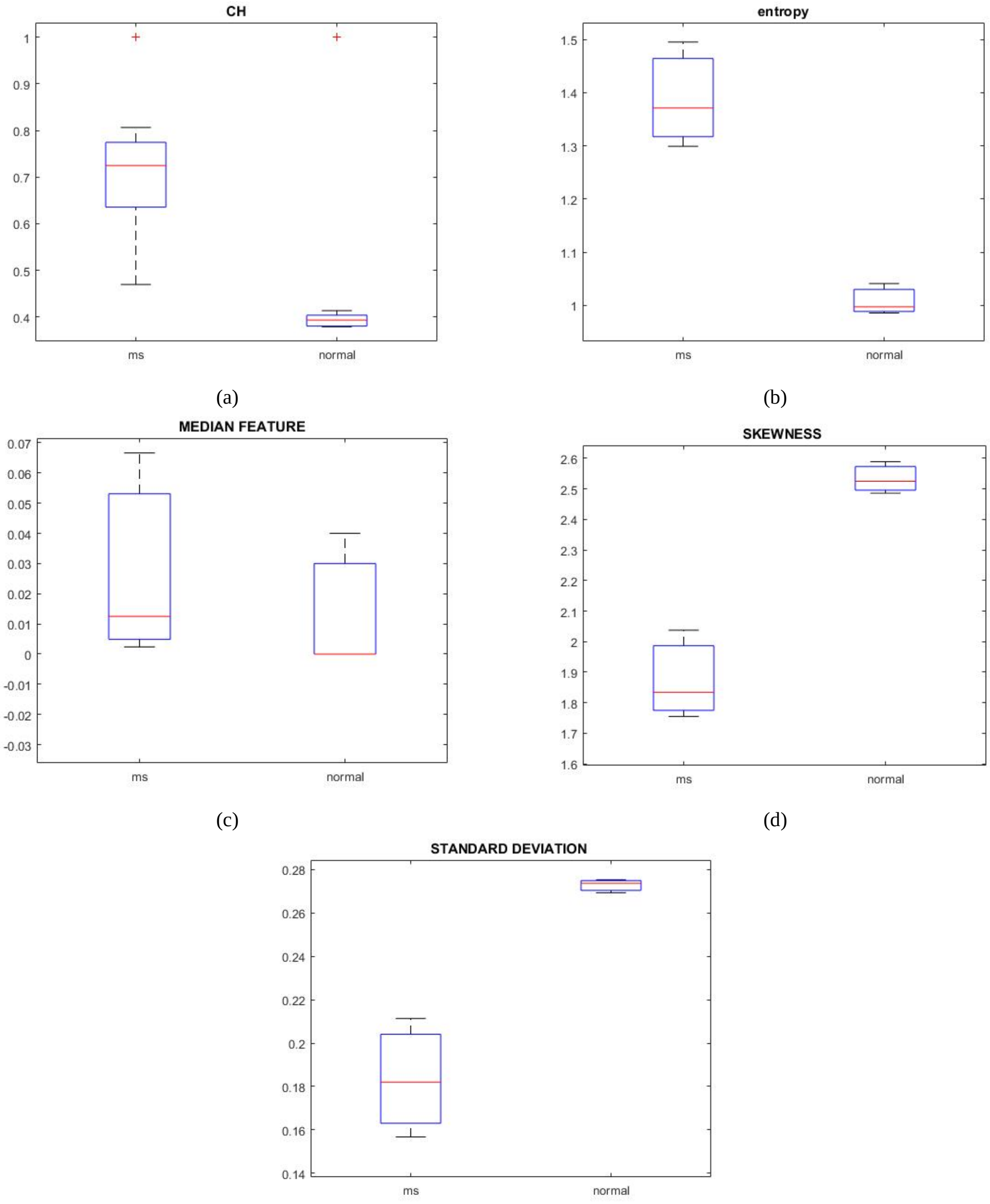
(b)

(e)

(c)

(f)

Fig. 6. The Classification Output of All Features: (a) CH Output, (b) SKW Output, (c) STD Output, (d) ENT Output, (e) MED Output, (f) Combination of All Features Output.



(a)

(b)

(c)

(d)

(e)

Fig. 7. The Box Plot of each Classification Feature for the Training Data: (a) Cumulative Histogram, (b) Entropy, (c) Median, (d) Skewness and (e) Standard Deviation.

We can use these parameters to evaluate:

1) Sensitivity (SE) which refers to the number of lesions that correctly classified, it is also called true positive rate (TPR). The ratio of TP plaques to the total number of MS plaques introduce the sensitivity of the test, in a more formal way:

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{8}$$

2) Specificity (SP) which refer to the ratio of TN plaques to the total number of normal plaques, it is also called true negative rate (TNR). In a more formal way:

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \tag{9}$$

3) Accuracy (AC) which refer to the summation of correct classification of MS and normal tissue to the total number of plaques. In a more formal way:

$$\mathrm{AC} = \frac{\mathrm{TPR+TNR}}{\mathrm{TPR+TNR+FPR+FNR}} \tag{10}$$

4) False Negative Rate (FNR) measures the normal plaques that misclassified as MS lesion divided by total number of MS plaques.

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{FN+TP}} \tag{11}$$

5) False Positive Rate (FPR) measures the multiple sclerosis plaques that misclassified as normal plaques divided by total number of normal plaques.

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \tag{12}$$

After several tests the performance of classification improved from 82-86% to 95-100% when Weiner and Adaptive filters have been added in preprocessing stage. The results of classification in two-dimension images using statistical features and CH are summarized in Table I, which show that the COMP (combination of all features) scored the highest accuracy and minimum FPR with 98% and 1% respectively, Also, CH gives the minimum FNR 1.78% with slight difference than the fused features. Thus combined features improve the performance of classification and overcome the high false positive rate in CH and increase the accuracy from 97 to 98%. Also in Fig. 7, it shows the difference between normal and abnormal tissues.

TABLE I. THE AVERAGE PERFORMANCE OF EACH FEATURE

feature	FPR	FNR	SE	SP	AC
CH	0.0317	0.0178	0.9821	0.9683	0.9752
STD	0.0508	0.0894	0.9106	0.9491	0.9298
SKW	0.0333	0.0751	0.9249	0.9667	0.9458
ENT	0.1526	0.0238	0.9762	0.8473	0.9118
MED	0.1351	0.0989	0.8649	0.9649	0.8866
COMP	**0.0100**	**0.0278**	**0.9721**	**0.9900**	**0.9810**

FPR: False Positive Rate, FNR: False Negative Rate, SE: Sensitivity, SP: Specificity, AC: Accuracy, CH: Cumulative Histogram, SKW: Skewness, STD: Standard Deviation, ENT: Entropy and MED: Median.

When comparing these results with previous work, we found that the proposed method has the minimum FPR and FNR than other work. Various researches have been accomplished to obtain the relation between the different gray levels and texture features [5]. In [22] texture analysis based gray level run length matrix (RLM) was performed on 110 Patients with classification accuracy of Multi Sclerosis 96.9%. Also in [23] they have been used MR Images of 30 MS patient where statistical texture feature analysis, autoregressive model, and wavelet-derived texture analysis were accomplished. The accuracy of classifying MS lesions and Normal Appearing WM was 96%-100%.

In [4] their work is divided into three models approach was based on logistic regression (LR) consists of ten texture parameters (Long Run Emphasis, Inverse Difference Moment, Difference Variance, Entropy, Run-Length Nonuniformity, Run percentage, Homogeneity, Low Gray-Level Run Emphasis, Long Run Low Gray-Level Emphasis, Short Run Low Gray-Level Emphasis), the enhanced lesions were classified correctly in twenty one patients with SE = 86% and SP = 84%. Also in [24] they used intensity subtraction and deformation field feature with a TPR of 74.30% and a FPR rate of 11.86% by obtaining a mean Dice similarity coefficient of 0.7. Finally in [25] they used Data Augmentation and AlexNet Transfer Learning model and their results are closed to our method with specificity 98.22% and sensitivity 98%.

V. CONCLUSION

In this paper, the proposed pipeline for the classification of MS lesions using a novel feature performed by combining statistical features with cumulative histogram (CH) as post processing for KNN segmentation. The result of CH feature classification showed that the MS areas were classified well from the other tissues although it has characteristics mostly similar to WM tissues and also by comparing these results with previous work we found that our results has the minimum FPR and highest accuracy. Also, the using of using adaptive filter and Weiner filter in preprocessing stage increased the accuracy of KNN segmentation. The results documented the potential of our framework to classify For qualifying the performance of this work and to improve the sensitivity with decreasing FPR, our future work will be dedicated to (i) conduct a supplement study on a larger number of data, (ii) utilize other segmentation algorithms that could improve the classification features performance, and (iii) extend our method to 3D imaging.

REFRENCES

[1] Compston, A. and Coles, A. (2008). Multiple sclerosis. The Lancet, 372(9648), pp.1502-1517.

[2] Luczynski, P., Laule, C., Hsiung, G., Moore, G. and Tremlett, H. (2019). Coexistence of Multiple Sclerosis and Alzheimer's disease: A review. Multiple Sclerosis and Related Disorders, 27, pp.232-238.

[3] O. Yu, Y. Mauss, G. Zollner, I. Namer and J. Chambron, "Distinct patterns of active and non-active plaques using texture analysis on brain NMR images in multiple sclerosis patients: preliminary results", Magnetic Resonance Imaging, vol. 17, no. 9, pp. 1261-1267, 1999.

[4] M. Salem et al., "A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis", NeuroImage: Clinical, vol. 17, pp. 607-615, 2018.

[5] X. Lladó et al., "Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches", Information Sciences, vol. 186, no. 1, pp. 164-185, 2012.

[6] Zhang, F., Wang, J. and Jiang, W. (2019). An integrative classification model for multiple sclerosis lesion detection in multimodal MRI. Statistics and Its Interface, 12(2), pp.193-202.

[7] Z. Karimaghaloo, M. Shah, S. Francis, D. Arnold, D. Collins and T. Arbel, "Automatic Detection of Gadolinium-Enhancing Multiple Sclerosis Lesions in Brain MRI Using Conditional Random Fields", IEEE Transactions on Medical Imaging, vol. 31, no. 6, pp. 1181-1194, 2012.

[8] S. Drabycz and J. Mitchell, "Texture quantification of medical images using a novel complex space-frequency transform", International Journal of Computer Assisted Radiology and Surgery, vol. 3, no. 5, pp. 465-475, 2008.

[9] J. Zhang, L. Tong, L. Wang and N. Li, (2008). Texture analysis of multiple sclerosis: a comparative study. Magnetic Resonance Imaging, 26(8), pp.1160-1166.

[10] P. Theocharakis et al., "Pattern recognition system for the discrimination of multiple sclerosis from cerebral microangiopathy lesions based on texture analysis of magnetic resonance images", Magnetic Resonance Imaging, vol. 27, no. 3, pp. 417-422, 2009.

[11] Y. Zhang, J. Wells, R. Buist, J. Peeling, V. Yong and J. Mitchell, "Active inflammation increases the heterogeneity of MRI texture in mice with relapsing experimental allergic encephalomyelitis", Magnetic Resonance Imaging, vol. 32, no. 2, pp. 168-174, 2014.

[12] Y. Zhang, "MRI Texture Analysis in Multiple Sclerosis", International Journal of Biomedical Imaging, vol. 2012, pp. 1-7, 2012.

[13] D. Sha and J. Sutton, "Towards automated enhancement, segmentation and classification of digital brain images using networks of networks", Information Sciences, vol. 138, no. 1-4, pp. 45-77, 2001.

[14] D. Malka et al., "Improved Diagnostic Process of Multiple Sclerosis Using Automated Detection and Selection Process in Magnetic Resonance Imaging", Applied Sciences, vol. 7, no. 8, p. 831, 2017.

[15] J. Sled, A. Zijdenbos and A. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data", IEEE Transactions on Medical Imaging, vol. 17, no. 1, pp. 87-97, 1998.

[16] S. Datta and P. Narayana, "Automated brain extraction from T2-weighted magnetic resonance images", Journal of Magnetic Resonance Imaging, vol. 33, no. 4, pp. 822-829, 2011.

[17] J. Friedman, F. Baskett and L. Shustek, "An Algorithm for Finding Nearest Neighbors", IEEE Transactions on Computers, vol. -24, no. 10, pp. 1000-1006, 1975.

[18] S. Warfield, "Fast k-NN classification for multichannel image data", Pattern Recognition Letters, vol. 17, no. 7, pp. 713-721, 1996.

[19] P. Anbeek, K. Vincken, M. van Osch, R. Bisschops and J. van der Grond, "Probabilistic segmentation of white matter lesions in MR imaging", NeuroImage, vol. 21, no. 3, pp. 1037-1044, 2004.

[20] M. Steenwijk, P. Pouwels, .M. Daams, J. van Dalen, M. Caan, E. Richard, F. Barkhof, and H. Vrenken, (2013). Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). NeuroImage: Clinical, 3, pp.462-469.

[21] C. Aldana Ramírez, N. Orozco Higuera and S. Barreto Melo, "Identification of multiple sclerosis brain lesions in magnetic resonance imaging using texture analysis", Revista Tecnura, vol. 18, p. 89, 2014.

[22] L Harrison, M. Raunio, K. Holli, T. Luukkaala, S. Savio, I. Elovaara, S. Soimakallio, H. Eskola, and P. Dastidar (2010). MRI Texture Analysis in Multiple Sclerosis: Toward a Clinical Analysis Protocol. Academic Radiology, 17(6), pp.696-707.

[23] N. Michoux, A. Guillet, D. Rommel, G. Mazzamuto, C. Sindic, and T. Duprez, (2015). Texture Analysis of T2-Weighted MR Images to Assess Acute Inflammation in Brain MS Lesions. PLOS ONE, 10(12).

[24] C. Loizou, E. Kyriacou, I. Seimenis, M. Pantziaris, S. Petroudi, M. Karaolis and C. Pattichis (2013). Brain white matter lesion classification in multiple sclerosis subjects for the prognosis of future disability. Intelligent Decision Technologies, 7(1), pp.3-10.

[25] Y. Zhang, V Govindaraj, C. Tang, W. Zhu and J. Sun (2019). High Performance Multiple Sclerosis Classification by Data Augmentation and AlexNet Transfer Learning Model. Journal of Medical Imaging and Health Informatics, 9(9), pp.2012-2021.

A Categorization of Relevant Sequence Alignment Algorithms with Respect to Data Structures

Hasna El Haji[1], Larbi Alaoui[2]
TIC-Lab, International University of Rabat
Rabat, Morocco

***Abstract*—Sequence Alignment is an active research subfield of bioinformatics. Today, sequence databases are rapidly and steadily increasing. Thus, to overcome this issue, many efficient algorithms have been developed depending on various data structures. The latter have demonstrated considerable efficacy in terms of run-time and memory consumption. In this paper, we briefly outline existing methods applied to the sequence alignment problem. Then we present a qualitative categorization of some remarkable algorithms based on their data structures. Specifically, we focus on research works published in the last two decades (i.e. the period from 2000 to 2020). We describe the employed data structures and expose some important algorithms using each. Then we show potential strengths and weaknesses among all these structures. This will guide biologists to decide which program is best suited for a given purpose, and it also intends to highlight weak points that deserve attention of bioinformaticians in future research.**

Keywords—Sequence alignment; data structures; bioinformatics

I. Introduction

"Sequence alignment" is a relevant subfield of bioinformatics that has attracted significant interest recently. It focuses on comparing two or more sequences to find homologies and visualize the effect of evolution across a family of genes [1].

Sequences can be divided mainly into two types: genomic and protein. Genomic sequences [2] are chains of nucleotides along a DNA/RNA macromolecule. They can be represented using the alphabet of the four letters of nitrogen bases: A, C, G and T. Protein sequences [2] are chains of twenty types of amino acids along a polypeptide. They can be represented using an alphabet of 20 letters (except B, J, O, U, X and Z) corresponding to the 20 existing amino acids.

Sequence alignment plays a crucial role in biology and medicine. Indeed, it is considered as the basis of many other tasks like phylogenetic analysis, evolution modeling, and prediction of gene expression. An assortment of algorithms has been applied to deal with sequence alignment. We can classify them according to two categories: exact and heuristic algorithms. Giving exact solutions, exact algorithms [3] are considered efficient but slow. Here we can cite Dynamic Programming (DP) developed by R. Bellman (1955). The principle of the DP consists of transforming a sequence S into a sequence Q, using three operations: substitution, insertion or deletion of a character. There is a cost to every operation and the aim is to find the sequence edited with a minimal cost. Considering two sequences S and Q, alignments are mapped to a matrix with entries representing optimal costs. Each cell is calculated based on its preceding cells. Concerning heuristic algorithms [4], they are designed for large databases and they give only approximate solutions to the problem. In fact, sequences are in exponential growth, thus, aligning a sequence against a database containing millions of sequences is not practicable. The aim of heuristic algorithms is to come up with fast strategies to rapidly identify relevant fractions of the cells in the DP matrix.

According to the number of compared sequences, "sequence alignment" is classified into two categories: pairwise alignment and multiple alignment. And according to the aligned regions, "sequence alignment" admits two major categories: global alignment category that aligns the entire given sequence and local alignment category that reveals similarity areas in long sequences. Fig. 1 gives a schematization of the Sequence Alignment Problem.

In the last twenty years a broad range of alignment techniques have been proposed [5]. Citing all these won't be conceivable inside the extent of this work. Moreover, to the best of our knowledge, only a few tools described in the literature have shown efficient results. Indeed, besides giving correct biological conclusions, they have managed to reduce the execution time from several days to a few minutes.

Through this paper, we intend to classify some recent alignment tools according to the widely used data structures. We typically give more interest to those having shown considerable improvement in terms of speed and memory consumption. We will highlight their strengths in sequence alignment tools, and we will also cite some of their drawbacks to reveal their limitations.

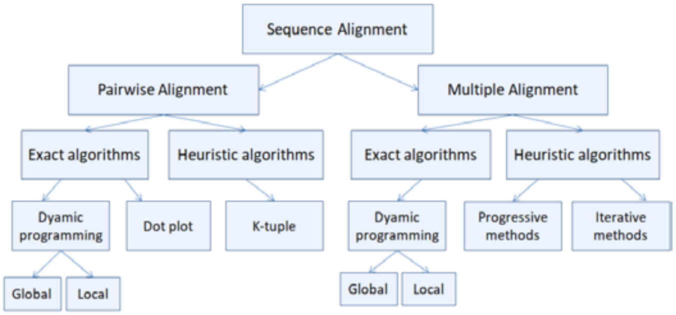


Fig. 1. A Schematization of the Sequence Alignment Problem.

This work is within the framework of the research project "Big Data Analytics - Methods and Applications (BDA-MA)". Author Hasna El Haji is financially supported by a PhD grant of International University of Rabat.

The paper is structured as follows. Most frequently employed methods for sequence alignment are briefly outlined in Section II. Then Section III introduces four of the commonly used data structures in literature and proposes a categorization of the most relevant algorithms based on these structures. Section IV discusses the utility of each structure in "Sequence Alignment". Finally, Section V concludes the paper and Section VI mentions a future work.

II. Related Work

A. Global Alignment

Regarding global alignment, the "Needleman-Wunsch algorithm" [6] is referred as a global alignment method based on dynamic programming. It was implemented in 1970. It gives a score to every possible alignment and tries to find all eventual alignments having the optimal score. It is executed in 4 stages:

1) Fixing the "similarity matrix" and the gap penalty,
2) Initializing the optimality matrix F,
3) Filling in the matrix F,
4) Giving a Traceback.

Let A be a sequence of length l, B be a sequence of length k, d be the gap penalty and $\Delta_{i,j}$ be the score of Match/Mismatch. Table I gives a pseudo-code of the Needleman-Wunsch Algorithm.

TABLE I. A Pseudo-Code of Needleman-Wunsch Algorithm

```
Input: two sequences to align
Initialization:
for i=0..l do:
    $F_{i,0} = d \times i$
for j=0..k do:
    $F_{0,j} = d \times j$
Recurrence relation:
For i=1..l do:
    For j=1..k do:
        $F_{i,j} = max \begin{cases} F_{i-1,j-1} + \Delta_{i,j} \\ F_{i-1,j} + d \\ C_{i,j-1} + d \end{cases}$
Output: Optimal alignment
```

B. Local Alignment

For local alignment, The "Smith-Waterman algorithm" [7] is a basic tool also based on "dynamic programming" but with extra choices to begin and end at wherever. Published in 1981, it maximizes the similarity measure by matching characters or inserting/deleting gaps in 4 steps:

1) Fixing the "similarity matrix" and the gap penalty,
2) Initializing the scoring matrix C,
3) Scoring,
4) Giving a Traceback.

Let A be a sequence of length l, B be a sequence of length k, $\Delta_{ins/del}$ be the gap penalty and $\Delta_{i,j}$ be the score of Match/Mismatch. Table II gives a pseudo-code of the Smith-Waterman Algorithm.

TABLE II. A Pseudo-Code of Smith-Waterman Algorithm

```
Input: two sequences to align
Initialization:
$C_{i,0} = 0$
$C_{0,j} = 0$
Recurrence relation:
For i=1..l do:
    For j=1..k do:
        $C_{i,j} = max \begin{cases} C_{i-1,j-1} + \Delta_{i,j} \\ C_{i-1,j} + \Delta_{ins/del} \\ C_{i,j-1} + \Delta_{ins/del} \\ 0 \end{cases}$
Output: Optimal alignment
```

C. Pairwise Sequence Alignment

"Pairwise sequence alignment" [8] is applied to examine the similarities of two sequences by finding the best matching alignment of them (the highest score). In other words, for a given sequence and a reference genome, the goal is to find positions in the reference where the sequence matches the best [9]. Three approaches generate the pairwise alignment: dot-plot analysis, "dynamic programming", and k-tulpe methods.

1) Dot-plot analysis (or Dot-matrix method): It is a qualitative and simple tool that compares two sequences to give the possible alignment [10]. Indeed, here are the steps of the method:

a) Two sequences A and B are listed in a matrix,

b) We start from the first character in B, we move over the matrix maintaining the first row and putting a dot in each column where there is a similarity between A and B,

c) The process continues until all possible comparisons between A and B are done. Such main diagonal dots refer to regions of similarity and isolated dots refer to random matches.

2) Dynamic programming: It can achieve global and local alignments. The global is most useful when the query sequences are similar and have the same length. The alignment calculation is generally done with the Needleman-Wunsch algorithm. The algorithm does not calculate the difference between two sequences but rather the similarity. Considering two sequences A and B, a two-dimensional array is filled row after row (starting from the last) and for each row, column after column (starting also from the last) respecting the recurrence relation (1):

$$F_{i,j} = max \begin{cases} F_{i-1,j-1} + \Delta_{i,j} \\ F_{i-1,j} + d \\ C_{i,j-1} + d \end{cases} \tag{1}$$

The local alignment is suitable to deal with dissimilar sequences with eventual similarities in their broader sequence context. An overall alignment would be insignificant. Calculation is generally done with the Smith-Waterman algorithm. The essential difference between Smith-Waterman and Needleman-Wunsch algorithms is that any cell of the initial comparison matrix in Smith-Waterman can be considered as a starting point for the calculation of scores and that any score that becomes less than zero stops the progression of the algorithm, then the score will be reset with the value 0. The calculation is based on the recurrence (2):

$$C_{i,j} = max\begin{cases} C_{i-1,j-1} + \Delta_{i,j} \\ C_{i-1,j} + \Delta_{ins/del} \\ C_{i,j-1} + \Delta_{ins/del} \\ \quad 0 \end{cases} \qquad (2)$$

3) K-tulpe methods (or word methods): They are heuristic methods faster than the original dynamic programming algorithms. They actually give only approximate solutions to the problem. K-tuple methods are implemented in the database search tools "FASTA" and "BLAST".

a) FASTA algorithm: An Algorithm for sequence comparison [11] based on the linked list structure: a query sequence is compared to all the strings in the database (DB). It is executed in six stages:

- Search for hot-spots (the largest common sub-sequences),
- Select the 10 best diagonal matches,
- Calculate the best scores for diagonal matches,
- Combine between good diagonal matches and indels,
- Calculate an alternative local alignment,
- Ordering of the results on the sequences of the DB

b) BLAST algorithm [12]*:* It is a tool based on a heuristic method that uses Smith-Waterman program. It looks for regions with strong similarity in alignments without spaces. It improves the speed of FASTA by looking for a smaller number of optimal hot spots. The substitution matrix is integrated from the first stage of hot spot selection.

D. Multiple Sequence Alignment

Multiple sequence alignment (MSA) [13] is a generalization of the "pairwise alignment"; it consists of comparing multiple related sequences. The aim is to deduce the presence of common ancestors between sequences.

Manually aligning more than three sequences can be difficult and time-consuming. Hence a variety of computational algorithms has been developed to accomplish this task. Most MSA algorithms use dynamic programming and heuristic methods (Progressive and Iterative).

1) Dynamic programming: DP is rarely used for more than three sequences because of its high running time and memory consumption. The same principle of DP in pairwise alignment can be applied here to multiple sequences. Unfortunately, the execution time grows considerably in an exponential way in comparison with the size of sequences, which is impractical. Nevertheless, a number of heuristic algorithms are used to accelerate computation. The most widely used heuristic methods today are the progressive and iterative techniques.

2) Progressive methods (tree methods): Invented in 1984, progressive alignment needs initial assumptions about the links between sequences to align, and uses those assumptions to build a guide tree to represent the links. The principle is as follows:

- Two most related sequences are aligned using dynamic programming methods,
- A third one is aligned to the first result,
- The process continues until a unique alignment of all the sequences persists.

The role of the "guide tree" is to choose a sequence to add to the alignment at each step.

The most popular progressive methods used at present are Clustal and T-coffee families.

a) Clustal family: Clustal (cluster analysis of the pairwise alignments) [14] are series of a widely used progressive programs; the original program was developed by Des Higgins in 1988 and was designed specifically to generate MSA on personal computers. The last standard version is ClustalΩ, it was updated in 2018 [32].

All versions of the Clustal family align sequences building progressively a multiple sequence alignment from successive pairwise alignments. This approach is executed in three steps:

- Provide a pairwise alignment,
- Construct a guide tree by a Neighbor -Joining m (developed in 1987 by Saitou and Nei),
- Use the tree to perform a multiple alignment.

b) T-coffee family: T-coffee [15] is a collection of multiple sequence alignment tools. It was originally published in 1998. T-coffee uses a new score function to evaluate the results. The method works through three steps:

- Create a library of all pairwise alignments and build a guide tree,
- Weight alignments by percentage of identical residues,
- Progressively build MSA using tree and weights.

3) Iterative methods: The major issue with progressive alignment is that errors in the initial alignments are transmitted to the whole MSA. Iterative methods [16] attempt to correct this problem by iteratively realigning subgroups of sequences; they start by making an initial global alignment of these subgroups and then revising the alignment to achieve a more efficient result. They can start from an initial MSA done with progressive alignment and then apply some modifications trying to improve it. Iteration is gainful in terms of coding, time complexity and memory requirements. The most widely used iterative methods are: MAFFT and MUSCLE.

a) MUSCLE: A method based on the guide tree construction technique. It produces a pairwise alignment for progressive alignment and for refinement. The progressive alignment employs a profile function called log-expectation. The refinement applies a tree-dependent restricted partition technique to reduce the execution time of the algorithm [17]. The method consists of three steps:

- Draft progressive: step of multiple alignment. It produces a first guide tree and a progressive alignment using k-mer distance (k-mer is a string part of size k) and log-expectation score.
- Improved progressive: step of building a second guide tree using the Kimura distance. It re-estimates the first tree and produces a new multiple alignment.
- Refinement: step of improving alignment. It refines multiple alignment using the tree-dependent restricted partitioning (deletes edges of guide tree, and re-form the alignment of separated trees).

b) MAFFT: Developed in 2002, the first version of MAFFT [18] was based on progressive alignment and clustering with the Fast Fourier Transform. It had been later provided to deal with large number of sequences and obtain more efficient results in accuracy. It is mainly executed in three stages:

- Detection of regions of similarity with Fast Fourier Transform,
- Application of the basic dynamic programming algorithm to select important strings,
- Build alignment.

III. A Categorization of the Most Efficient Algorithms based on their Data Structures

All the cited methods are the basis of the existing sequence alignment algorithms. To obtain fast and efficient solutions in memory, fundamental methodology in many of them is to build a data structure that occupies a reasonable space of memory. Data structures are one of the most important concepts in programming. They are a way of storing and managing data. Most of the sequence alignment algorithms are built upon such basic data structures like: suffix arrays, suffix trees, hash tables, graphs and others. In this section, we discuss four relevant data structures and we propose a structure-based categorization of the most important algorithms that have marked the last two decades.

A. Suffix Arrays and Suffix Trees

Given a sequence S, a suffix array is an ordered array of all suffixes of S, and a suffix tree is a representation of all suffixes in S in the form of a tree. The latter contains a leaf for each suffix, and each edge is labeled with a string of characters so that the path from the origin to each leaf gives the corresponding suffix.

In "Sequence Alignment", a suffix tree is an index data structure for analogous sequences. It stores all suffixes of an alignment of similar strings. In this category, we will cite 3 important algorithms:

In 2017, paper [19] proposed a multiple sequence alignment algorithm that combines between a suffix tree and a center-star strategy (MASC). The latter transforms an MSA problem into a pairwise alignment, and the suffix tree matches identical regions between two pairwise sequences. The algorithm can be executed in a linear time complexity $O(mn)$, where m is the amount of sequences and n is their average length. The method is also characterized with no loss in accuracy for highly similar sequences.

Earlier, in 2013, article [20] proposed a memory gainful edition of the suffix tree method: "Suffix Array of Alignment (SAA)". It deals with pattern research appropriately like the "Generalized Suffix Array (GSA)". The paper also presents a practical approach for building the SAA. Experiments have shown that the SAA is a relevant data structure for relatively identical strings. It only takes around one seventh of the memory provided by the GSA to process 11 strings.

Suffix trees aim to reduce memory space, they provide a linear space complexity and they allow a linear-time searching [21]. However they could require more than 20 bytes per character, rather, suffix arrays are more useful generally. And here we should cite reference [22] that shows in depth the weakness of suffix trees and their negative effect on the algorithm efficiency. The paper also proved that algorithms based on suffix tree could be replaced with equivalent algorithms based on suffix array, which is memory gainful.

Suffix trees are used for Read Alignment and Whole Genome Alignment while suffix arrays are more adequate to Prefix-suffix Overlaps Computation and Sequence Clustering.

B. Hash-Tables

A hash table is a kind of associative array storing pairs of keys and values. It represents a genome sequence as multiple lists of genomic positions. The concept of hash tables belongs to the heuristic method "BLAST". Indeed, all hash table tools adopt typically one principle.

Author in [23] has developed the first BLAST implementation algorithm such that the stage of "word mapping" is promoted by a hash table. The algorithm was later improved by other researchers involving the notion of parallelization and different other techniques to accelerate alignments. Author in [24] proposed in 2012 an accelerated short read aligner to approximate the original dynamic programming algorithm. It uses a hash table and applies the Needleman-Wunsch algorithm as an extension. The advantage of the hash table here is to reduce the amount of work by avoiding the generation of too many candidate regions.

Author in [25] proposed BFAST, a two-level indexing method. It applies the hash in indexing to decrease the research time. It was introduced to handle the alignment of short human genomic sequences; this makes it an efficient aligning method that is recommended to deal with each number of sequences and every reference genome.

The methods based on hash tables usually provide high sensitivity, but as a limitation, they occupy a lot of memory because the size of the array grows exponentially.

C. Tries

Tries are a kind of trees storing the entire suffixes of a sequence and enabling a quick matching of sequences. They derive from the middle letters of "retrieval". They are efficient in the storage of multiple sequences, and useful in accelerating the process of sequence searching [26].

We have to cite here the reference [27] that proposed in 2018 a fast trie based method for multiple alignment. It bypassed the classic enumeration of successive comparisons with all strings. It also provides an original algorithm combining a trie and an exact algorithm to find the edit distance between strings.

Author in [28] marked 2015 by developing two multiple sequence alignment tools. The first one employed tries to accelerate the alignment of highly similar sequences. The second worked in parallel with Hadoop to deal with big data. Tries worked as a dictionary that stores substrings and indexes. To collect each substring in a long sequence, tries reduce the running time by avoiding the individual substring research.

We also cite [12] that presented BLAT, a tool applying tries to find the regions matching with the query sequence. It demonstrated a considerable progress in accuracy and running time compared to the popular existing tools in the early 2000s.

The practical benefit of tries remains in time reduction. In fact, aligning a sequence to the same duplicates of a string is done only once. This is mainly due to the fact that duplicates fall on the same line in the tree. Though, a hash table should perform an alignment for each duplicate. Thus, tries stay considerably faster than hash tables in time execution.

However, given a reference sequence of length m, a trie can take $O(m^2)$ memory space, which makes it impractical to build a trie for long sequences.

D. Graphs

Recent researches have shown superior accuracy and speed in aligning sequences by using a "Variation Graph" instead of a reference genome. A "Variation Graph" is a directed graph where each edge spells a sequence. In fact, aligning sequences to graph is trying to determine the optimal corresponding path in the graph for every sequence.

In addition to variation graphs, a variety of graph data structures have been studied in the last decades, such as "De Bruijn" graphs, "ABruijn" graphs, String graphs, Partial Order graphs, "Wheeler graphs", etc.

Each of these structures had demonstrated a considerable progress in solving the sequence alignment problem (more details are given in [29]). However, they had registered some weaknesses: high execution time, overlook of arbitrary graphs, imprecise results. In what follows, we will expose three recent methods, based on variation graph, that have influenced more considerably the alignment accuracy.

Authr in [30] suggests PaSGAL, an algorithm to solve sequence-to-graph alignment using parallelism. It is considered as the first parallel algorithm provided for this propose. It generates improved results compared to previous tools on execution time term. Indeed, it allows a decrease from long durations to few hours.

The main idea of the algorithm is to accelerate the "Dynamic Programming" approach on 3 phases: the 1st and the 2nd phases compute the starting and ending cells of the alignment matrix, and the final phase performs a traceback. The latter aims to calculate the "base-to-base" alignment scores required for downstream biological analysis. The algorithm is highly recommended for pan-genomics and antibiotic resistance profiling.

In 2019, the authors of [31] studied two problems and proposed two solutions: a generalization of the "Shift-And algorithm" (designed for exact string matching) to graphs, and a generalization of "Myers’ bit vector alignment algorithm" to graphs. Both solutions are based on Needleman-Wunsch algorithm. The paper used a "bit-level parallelization" to estimate the distance between the query sequences and the graph. The method is supposed to fit with the mammalian genome.

The first author of [31] has recently improved his works, and came up with GraphAligner, an efficient sequence-to-graph alignment tool. Compared to existing graph methods, it is 12 times more effective in time complexity. It also takes in consideration long reads error correction and outperforms the current tools 3 times in error rate.

The major advantage of graphs in general is that each stage of the alignment can store and use results from previous alignments. More specifically, Variation graphs become today a reference for analyzing genetic variants.

IV. Discussion

We highlighted the weaknesses and strengths of the most relevant sequence alignment tools and precise their utility in Biology. We aim to facilitate the choice of appropriate tools for each biologist depending on his research intention.

Suffix trees are applicable to Read Alignment and Whole Genome Alignment. Suffix arrays might be also applicable to Read Alignment, but they are more useful in Prefix-suffix Overlaps Computation and Sequence Clustering. Compared to the discussed data structures, hash tables remain more performing in query time execution but they are memory consumers. As application in Bioinformatics, they are more suitable for storing sets of k-mers. Finally, Graphs are relevant in Read Mapping and they can fully represent population-wide variations.

V. Conclusion

This work is a small sample of two decades of scientific production in the field of sequence alignment. In fact, many studies have demonstrated considerable progress in storage and acceleration of the aligning process. The key idea of all the exposed methods is to create data structures to store calculated scores in the smallest possible space during alignment. That will provide a gain in both memory and runtime.

We started with a classification of the basic methods of sequence alignment. Then we introduced four of the commonly used data structures in literature (suffix arrays/trees, tries, hash-tables and graphs) and we proposed a categorization of the most relevant algorithms based on these. There are such other structures like Burrows-Wheeler transform (BWT), bloom filters, FM-index (combination between the properties of suffix array and the BWT), etc., but we tried to give more attention to the four cited data structures by means of their efficiency in giving the best alignment.

VI. Future Work

From the performance point of view, none of the cited algorithms is yet considered as ideal. A best solution should combine accuracy, speed and small memory space. Furthermore, sequence databases are rapidly and continuously growing, thus the development of high performing methods is still under research.

We believe that graphs, in full development, can be refined further. As a future work, we will give more interest to the sequence-to-graph alignment.

References

[1] B. Chowdhury and G. Garai, “A review on multiple sequence alignment from the perspective of genetic algorithm,” Genomics, pp. 419–431, 2017.

[2] J. C. WHISSTOCK and A. M. LESK, “Prediction of protein function from protein sequence and structure,” Q. Rev. Biophys., pp. 307–340, 2003.

[3] A. PHILLIPS, D. JANIES, and W. WHEELER, “Multiple sequence alignment in phylogenetic analysis,” Mol. Phylogenet. Evol., pp. 317–330, 2000.

[4] İ. Ö. Bucak, V. Uslan, and S. Member, “An analysis of Sequence Alignment : Heuristic Algorithms,” pp. 1824–1827, 2010.

[5] H. Ng, S. Liu, and W. Luk, “Reconfigurable Acceleration of Genetic Sequence Alignment : A Survey of Two Decades of Efforts,” in 27th International Conference on Field Programmable Logic and Applications (FPL), 2017.

[6] “S.B. Needleman and C.D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins, ”J. Mol. Biol., pp. 443-453, 1970.

[7] A. Khajeh-Saeed, S. Poole, and J. Blair Perot, “Acceleration of the Smith-Waterman algorithm using single and multiple graphics processors,” J. Comput. Phys., pp. 4247–4258, 2010.

[8] W. Haque, A. A. Aravind, and B. Reddy, “Pairwise sequence alignment algorithms: a survey,” in ISTA ’09: Proceedings of the 2009 conference on Information Science, Technology and Applications, pp. 96–103, 2009.

[9] K. Salikhov, “Efficient algorithms and data structures for indexing DNA sequence data,” Bioinformatics [q-bio.QM]. Université Paris-Est; Université Lomonossov (Moscou), 2017. English. ffNNT : 2017PESC1232ff. fftel-01762479f.

[10] N. GALTIER, M. GOUY, and C. GAUTIER, “SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny,” Bioinformatics, pp. 543–548, 1996.

[11] A. L. Delcher, “Fast algorithms for large-scale genome alignment and comparison,” Nucleic Acids Res., pp. 2478–2483, 2002.

[12] W. J. Kent, “BLAT — The BLAST -Like Alignment Tool,” Genome Res., pp. 656–664, 2002.

[13] D. W. Mount, “Using Iterative Methods for Global Multiple Sequence Alignment,” Cold Spring Harb Protoc, pp. 1–6, 2009.

[14] J. Daugelaite, A. O’ Driscoll, and R. D. Sleator, “An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics,” ISRN Biomath., pp. 1–14, 2013.

[15] C. Notredame, “Recent progress in multiple sequence alignment: A survey,” Pharmacogenomics, pp. 131–144, 2002.

[16] T. Rausch, A. K. Emde, D. Weese, A. Döring, C. Notredame, and K. Reinert, “Segment-based multiple sequence alignment,” Bioinformatics, pp. 187–192, 2008.

[17] R. C. Edgar, “MUSCLE: Multiple sequence alignment with high accuracy and high throughput,” Nucleic Acids Res., pp. 1792–1797, 2004.

[18] K. Katoh, K. Misawa, K. Kuma, and M. T, “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform,” Nucleic Acids Res., pp. 3059 -3066, 2002.

[19] W. Su, X. Liao, Y. Lu, Q. Zou, and S. Peng, “Multiple Sequence Alignment Based on a Suffix Tree and Center-Star Strategy: A Linear Method for Multiple Nucleotide Sequence Alignment on Spark Parallel Framework,” J. Comput. Biol., 2017.

[20] J. C. Na et al., “Suffix array of alignment: A practical index for similar data,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), pp. 243–254, 2013.

[21] H. Li and N. Homer, “A survey of sequence alignment algorithms for next-generation sequencing,” Briefings In Bioinformatics, pp. 473-483, 2010.

[22] M. I. Abouelhoda, S. Kurtz, and E. Ohlebusch, “Replacing suffix trees with enhanced suffix arrays,” J. Discret. Algorithms, pp. 53–86, 2004.

[23] P. Krishnamurthy et al., “Biosequence Similarity Search on the Mercury System,” J. VLSI Signal Process. Syst., pp. 101–121, 2007.

[24] Y. Chen, B. Schmidt, and D. L. Maksell, “An FPGA Aligner for Short Read Mapping,” in 22nd International Conference on Field Programmable Logic and Applications, pp. 511–514, 2012.

[25] N. Homer, B. Merriman, and S. F. Nelson, “BFAST: an alignment tool for large scale genome resequencing,” PLoS One, 2009.

[26] J. Wang et al., “Interactive and fuzzy search : a dynamic way to explore MEDLINE,” Bioinformatics, pp. 2321–2327, 2010.

[27] P. A. Yakovlev, “Fast Trie-Based Method for Multiple Pairwise Sequence Alignment,” Doklady Akademii Nauk, pp. 64–67, 2019.

[28] Q. Zou, Q. Hu, M. Guo, and G. Wang, “HAlign: Fast multiple similar DNA/RNA sequence alignment based on the centre star strategy,” Bioinformatics, pp. 2475–2481, 2015.

[29] B. Kehr, K. Trappe, M. Holtgrewe, and K. Reinert, “Genome alignment with graph data structures : a comparison,” BMC Bioinformatics, 2014.

[30] C. Jain, A. Dilthey, S. Misra, H. Zhang, and S. Aluru, “Accelerating sequence alignment to graphs,” Proc. - 2019 IEEE 33rd Int. Parallel Distrib. Process. Symp. IPDPS 2019, pp. 451–461, 2019.

[31] M. Rautiainen, V. Mäkinen, and T. Marschall, “Bit-parallel sequence-to-graph alignment,” Bioinformatics, pp. 3599–3607, 2019.

[32] F. Sievers and D. G. Higgins, “Clustal Omega for making accuratealignments of many protein sequences”, Protein Science, pp. 135-145, 2018.

Bangla Optical Character Recognition and Text-to-Speech Conversion using Raspberry Pi

Aditya Rajbongshi[1], Md. Ibadul Islam[2], Md. Mahbubur Rahman[4], Anup Majumder[5], Dr. Md. Ezharul Islam[6]
Department of Computer Science and Engineering
Jahangirnagar University, Savar, Dhaka, Bangladesh

Al Amin Biswas[3]
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh

***Abstract*—Optical Character Recognition (OCR) technology is very helpful for visually impaired or illiterate persons who are unable to read text documents but need to reach the content of the text documents. In this paper, a camera-based assistive device is used that can be applied for visually impaired or illiterate people to understand Bangla text documents by listening to the contents of the Bangla text images. This work mainly involves the extraction of the Bangla text from the Bangla text image and converts the extracted text to speech. This work has been fulfilled with Raspberry Pi and a camera module by applying the concepts of the Tesseract OCR engine, the Open Source Computer Vision, and the Google Speech Application Program Interface. This work can help people speaking Bangla language who are unable to read or have a significant loss of visual sight.**

***Keywords*—Optical character recognition; Bangla text; speech conversion; Raspberry Pi; camera module**

I. Introduction

Text and speech is the primary medium for communication among human beings. For accessing the text information, an individual needs visual sight. Besides the visual sights, an individual can also know the information using their listening capability. According to the World Health Organization, the amount of visually impaired people is 285 million and the amount of blind people is 39 million in the whole world [1]. More than 90 percent of the visually impaired people exist in developing countries [2] and on the other hand, according to UNESCO [3], 27.11% of adults are illiterate. All these facts have raised the importance of research to develop systems that can help visually impaired persons to overcome their limitations.

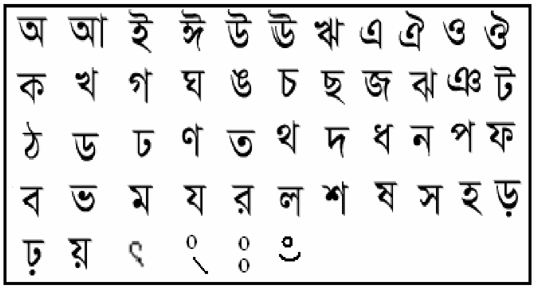
Raspberry Pi [4] is a single-board computer. For exploring computing and learning various programs such as python, scratch, etc. this device helps people a lot. It has the capability of calculating, playing music, gaming, and other functions that are done by a computer. The main advantages of this device are portability and low cost. For experimental and innovation activities this board is designed. Its two types of model differentiate each other based on the USB port. For the above-mentioned features, Raspberry Pi has become an essential and ideal tool for IoT and automation research.

In the recent era, OCR [5] has been used for converting images to text. It helps millions of people to know the information from scripts such as airline tickets, medical documents, mail, etc. in their perspective file. In the recent advancement of OCR technology and algorithms such as the Tesseract OCR engine [6], it can recognize a huge number of characters in various languages. The application of OCR touches every technological organization in the world. It also included the recognition of characters from handwriting scripts in various languages.

Bengali or Bangla is an Indo-Aryan language. As a primary or secondary language, around 210 million people speak in Bengali, among them around 100 million, and 85 million speakers are from Bangladesh and India, respectively [7]. Bangla OCR is different from other languages because of the basic structure of the Bengali script. Bengali letters as in Fig. 1 have different transformers and edges. Besides, a large number of characters are contained in the Bengali script. There are 57 characters in Bengali scripts among them 21 characters are a vowel and 36 characters are consonants. Because of curves in the character, sliding, and stroke characters researchers face various challenges.

In this paper, we propose a Bengali OCR based system. The image which is captured by the user using the camera included in the proposed device is analyzed in various phases of Tesseract (OCR Engine) methods. The text in the input image is extracted using the method of Open Source Computer Vision (OpenCV). For converting the text into a speech, GTTs (Google Text to Speech) library is used and it works offline. In Raspberry Pi board, a slot is used for connecting the headphone and the user needs to be connected with the headset. Thus, the user can hear the speech of the text through the headphone.

This work is ordered as follows: Section II outlines the existing relevant work. Then, Section III explains the working methodology with the system architecture that includes system hardware and software implementation. System evaluation is demonstrated in Section IV. Subsequently, the conclusion of the paper by mentioning some future works is appended in Section V.



অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ
ক খ গ ঘ ঙ চ ছ জ ঝ ঞ ট
ঠ ড ঢ ণ ত থ দ ধ ন প ফ
ব ভ ম য র ল শ ষ স হ ড়
ঢ় য় ৎ ং ঃ ঁ

Fig. 1. Bengali Letters.

II. Related Works

Very few research has been done that would help blind people to follow or recognize objects to pass their daily life smoothly. The aspects of this kind of problem are notably complex and this kind of work is exceptional to do. Many researchers have discovered several possible solutions for text to speech conversion using diverse methodologies.

R. Naveen et al. [8] introduced a method of a camera-based assistive framework to assist blind people to recognize faces, signs, obstacles, text objects and the feelings of persons and gave it as a required audio output through the earphone that has been used in their daily life using the raspberry pi. This work is mainly devised to blind navigation purposes but not shown any result-oriented data and didn’t mention the output accuracy rate or level.

L. Nagaraja et al. [9] proposed a camera-based assistive method that read texts from an image for blind peoples and printed notes and papers and converted it to speech. This work is a vision based recognition system based on raspberry pi and the images are extracted from OCR engines. Due to the resolution problem of the images, this work has not achieved 100% accuracy.

V. A. Devi and S. S. Baboo [10] developed an embedded system based OCR based framework that can read especially Tamil texts from an image using the raspberry pi. This work is only for Tamil text conversion from an image and it is only captured over a printed page.

G. K. Sagar and T. Shreekanth [11] have implemented an OCR system that converts text to speech using raspberry pi for helping blind people and people with poor vision. In this TTS system images are captured through a webcam and processed by the TTS and amplified by audio and give the output to the speaker. Classification is done by the Template Matching and Neural Networks.

S. A. James et al. [12] used the raspberry pi to implement an automatic book reader which is OCR based. Here, OCR is used to recognize the text of the reader and. By using the Adaboost learning model, calculation of text classification and adjacent character grouping is performed.

C. A. Todd et al. [13] introduced a tool which is known as audio haptic. This is mainly for visually impaired users. This work is based on web technologies with audio and haptic interaction to aid the visually impaired.

D. Velmurugan et al. [14] produced a smart book reader with a raspberry pi controller which is also OCR based. Using the various methods, the pixels of the image are transformed. The extracted text is then converted to speech which is listened to by headset or another device.

C. S. T. Thu and T. Zin [15] implemented OCR to recognize English characters (capital) and numbers using MATLAB. This work is done by two major steps like OCR system and TTS conversion. Here Neural Network is used for classification.

H. M. Htun, T. Zin, H. M. Tun [16] developed a TTS system that converts text to speech and numbers, words, and sentences for visually impaired and handicapped people. Here, the token is obtained by segmenting the input sentence, and each of the words is considered as POS tagging. Bigram Model is used for the POS tagger.

A. Goel et al. [17] implemented a system for English book readers in Raspberry Pi. Here python programming is used for text extraction from captured images and audio speech conversion.

H. Rithika and B. N. Santhoshi [18] proposed a model that aids the user to listen to the text image's content in the fancied language. It includes the text extraction from the input picture and transforming this to interpret speech. The speech-language depends on the user. This work is accomplished by employing a Raspberry Pi and a camera module and also with the help of the theories of the Tesseract OCR engine and GSAPI.

III. Proposed Methodology and System Architecture

The proposed system is divided into two sections as system hardware implementation and system software implementation. In our daily life, there are many Bangla scripts. The user who can’t read the Bangla scripts papers, he captures the image using a raspberry pi camera. After capturing the image, the image is processed and analyzed in raspberry pi using the tesseract (OCR engine) and OpenCV [19] methods. Then the text of the analyzed image is extracted. Using the GTTs library, this extracted text converts to speech offline. Finally, the user hears the speech that is written on the Bangla scripts. Fig. 2 shows the proposed system architecture.

A. System Hardware Implementation

The following parts are used as the constitution of the hardware system of the device: a Raspberry Pi Camera Module for image capturing, breadboard for push-button, to execute image recognition programs Raspberry pi board uses and to produce the desired output speech a headphone uses. The system hardware organization is shown in Fig. 3.

a) Raspberry Pi: Raspberry pi [4] developed by the Raspberry Pi organization for basic computer science at schools in developing countries, is a group of single-board computers. There are various versions of Raspberry Pi based on memory uses or USB support. Broadcom BCM2835 SoC is used as a first-generation Raspberry Pi. In this system, the Raspberry Pi 3 B model is used.

b) Pi camera: The Pi camera module has the excellency of taking pictures and recording videos in Raspberry Pi. In Raspberry Pi, there is a port to connect the Pi Camera. There are various versions of the Pi camera. Most of them can deliver a clear image. In this device, Raspberry Pi Camera V2 is used for capturing images. It has 5MP Resolution and supports 1080p30 video recording.

c) Memory (Storage): There is no hard disk or solid-state drive in the device. Instead of this, a micro SD card is used for booting the Linux kernel-based OS (operating system). In Raspberry pi, there is a slot for inserting a micro SD card. This card is also used for the data storage of capturing images. When the image is captured, it is stored on the card. The analyzing image is also stored in the memory card. The specification of the used hardware components is provided in Table I.

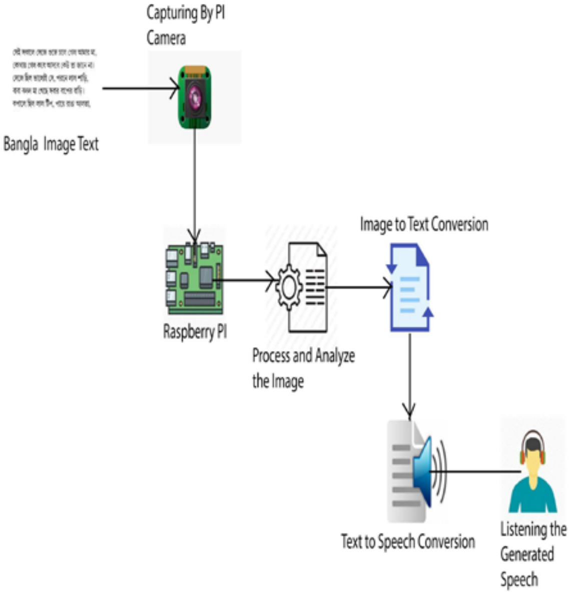

Capturing By PI Camera
Bangla Image Text
Raspberry PI
Process and Analyze the Image
Image to Text Conversion
Text to Speech Conversion
Listening the Generated Speech

Fig. 2. Proposed System Architecture.

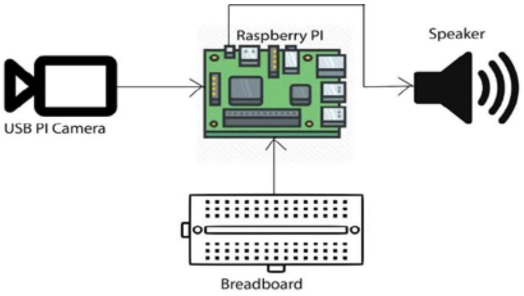


Fig. 3. System Hardware Organization.

TABLE I. SPECIFICATION OF HARDWARE COMPONENT

Component Name	Version/Storage
Raspberry Pi	3, B model
Pi Camera	V2 model
Breadboard	PMSA003
Memory	Samsung 16GB micro SD

d) Camera enabling: In Raspberry Pi, there are three ways to enable the Pi camera. One is to set manually and another one is to use the command line and also in python code. It's better to enable through the command line. Without enabling a Camera, it doesn't work. After enabling the camera, its configuration automatically adjusts to the Raspberry Pi.

e) Push-button setup in breadboard: A breadboard is a board of rectangular plastic having multiple tiny holes. In this project, a push-button is used to capture the image which is integrated with the breadboard. When the push button is hit, the image is captured. To set up the push button in the breadboard, the circuit diagram of Fig. 4 is used. Fig. 5 shows system setup after connecting all of the hardware components.

B. System Software Implementation

In a system software implementation, we have installed the Raspbian Operating System [20] in a memory card and inserted it in the raspberry pi. After installing Raspbian OS, we have also installed various tools, packages, and libraries. Table II shows the details of some software components.

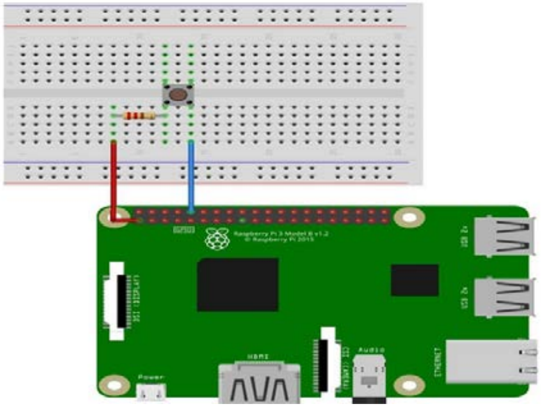


Fig. 4. Circuit Diagram for Push-Button Setup.

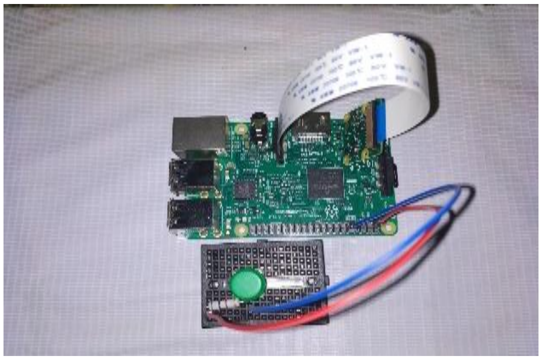


Fig. 5. System Setup after Connecting Hardware Components.

TABLE II. SPECIFICATION OF SOFTWARE COMPONENT

Software Component Name	Version
Raspbian OS	4.19
Python	3.6.0
Tesseract	3.05.00
OpenCV	CV2

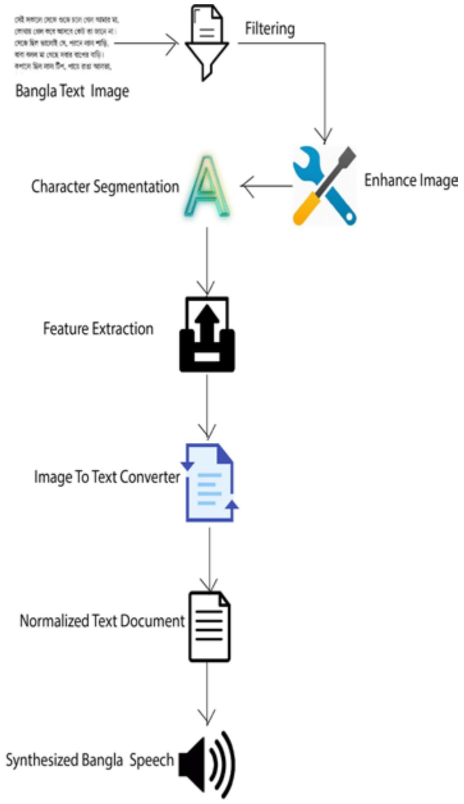


Fig. 6. System Operational Architecture.

The captured image from the Raspberry Pi camera is stored in a file as a JPEG format. The stored image is analyzed in python script through the following steps as sketched in Fig. 6.

The details of the above steps are described below:

a) Acquisition of image: In this step, a Bangla text image is captured using the Pi camera. Captured Bangla text image is sent to the prepossessing step where various unwanted noise is reduced.

b) Prepossessing of image: By applying relevant morphological transformation like dilation, back hat transformation, threshold, producing the necessary contours, discrete cosine transformations, and forming bounding box, the unwanted noise in the image is banished in the prepossessing of image. First of all, the captured image is rescaled to the relevant size and then converted into a grayscale image for further processing. The grayscale image is then compressed using the discrete cosine transformation. The compressed image is very helpful for further processing. In the compressed image, there exist various unwanted high-frequency components. Those components are omitted by setting the vertical and horizontal ratio. For decompression of the image, the inverse discrete cosine transform is applied. There are two operations like back top-hat transformation and dilation which are used in the image. Then the operation named black top-hot transformation is employed in the image. This operation helps to extract the object or elements that are smaller than the defined. After this operation, the dilation operation is applied for adding the pixel to the edges of the object of the image. The number of pixels depends on the shape and size of the present object. Now the thresholding algorithm is applied in the present image. Among all the thresholding algorithms, here adaptive thresholding is chosen. Then using specific functions of OpenCV, the contours of the image are generated. There are many bounding boxes of the objects or elements in the present image. For drawing those, the generated contours are used. For extracting every character of the present image, the drawn bounding boxes are used. Finally, by applying the OCR (Optical Character Recognition) engine, the full text of the present image is detected.

c) Extraction of text: From the input image, the recognized text is extracted in this step. This extraction is performed using the Tesseract OCR engine.

d) Text to speech converter: Applying the GTTs engine, the extracted text is converted to speech in this step. With the help of some predefined libraries of this engine, we performed the text to speech conversions. In the GTTs engine, there are online and offline systems. In our project, we have used an offline system for user portability.

e) Desired output speech: When speech is generated, the user can easily hear it through the headphones. As it is based on a Bangla text image, the user can easily hear the speech in Bangla language.

IV. System Evaluation

For the system analysis, there needs to be some Bangla scripts paper. A monitor is included for proper monitoring of recognition of text from the captured image in the system. The Fig. 7 shows the final view of our system.

For the result, an image is captured using the Raspberry Pi camera. It is considered as the input image. Image prepossessing, extraction of texts are handled by the various defined methods of the tesseract (OCR engine). The accuracy of the extraction of text in captured images is not 100% because of the mid resolution pi camera. After extraction of text, a user can easily hear the speech through a headphone that is connected to the earphone slot in Raspberry Pi. For the precise image, the accuracy of text extraction is satisfactory. Sometimes the accuracy of text extraction is 100% for some precise images. Fig. 8, 9, 10, and 11 show the input image, the Bangla text extraction from the input image, the texts displayed on the screen, and a user of hearing speech, respectively.

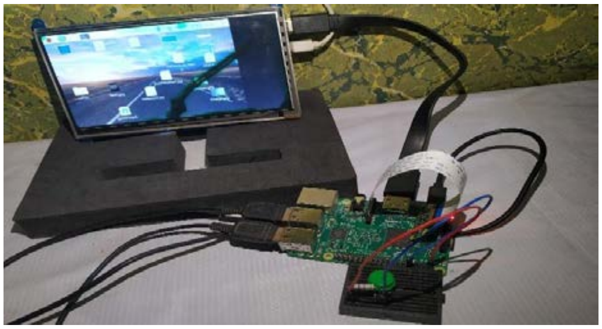
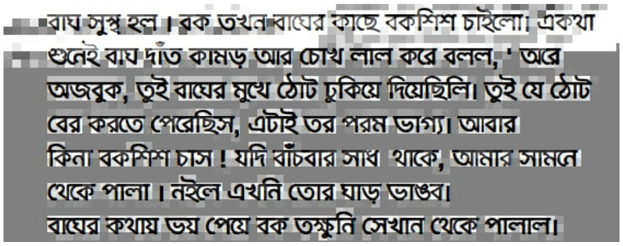
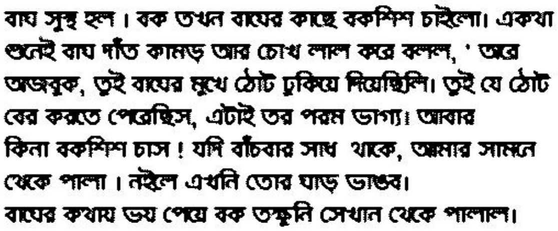


Fig. 7. The System Connecting with Monitor.



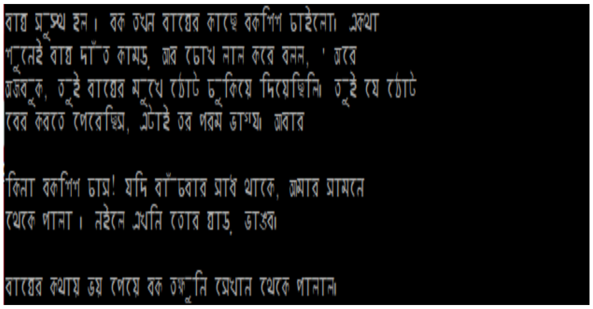
বাঘ সুস্থ হল। বক তখন বাঘের কাছে বকশিশ চাইলো। একথা
শুনেই বাঘ দাঁত কামড় আর চোখ লাল করে বলল, ' অরে
অজবুক, তুই বাঘের মুখে ঠোঁট ঢুকিয়ে দিয়েছিলি। তুই যে ঠোঁট
বের করতে পেরেছিস, এটাই তর পরম ভাগ্য। আবার
কিনা বকশিশ চাস ! যদি বাঁচবার সাধ থাকে, আমার সামনে
থেকে পালা। নইলে এখনি তোর ঘাড় ভাঙব।
বাঘের কথায় ভয় পেয়ে বক তক্ষুনি সেখান থেকে পালাল।

Fig. 8. Input Image.



বাঘ সুস্থ হল। বক তখন বাঘের কাছে বকশিশ চাইলো। একথা
শুনেই বাঘ দাঁত কামড় আর চোখ লাল করে বলল, ' অরে
অজবুক, তুই বাঘের মুখে ঠোঁট ঢুকিয়ে দিয়েছিলি। তুই যে ঠোঁট
বের করতে পেরেছিস, এটাই তর পরম ভাগ্য। আবার
কিনা বকশিশ চাস ! যদি বাঁচবার সাধ থাকে, আমার সামনে
থেকে পালা। নইলে এখনি তোর ঘাড় ভাঙব।
বাঘের কথায় ভয় পেয়ে বক তক্ষুনি সেখান থেকে পালাল।

Fig. 9. Extracted Text from the Input Image.



বাঘ সুস্থ হল। বক তখন বাঘের কাছে বকশিশ চাইলো। একথা
শুনেই বাঘ দাঁত কামড় আর চোখ লাল করে বলল, ' অরে
অজবুক, তুই বাঘের মুখে ঠোঁট ঢুকিয়ে দিয়েছিলি। তুই যে ঠোঁট
বের করতে পেরেছিস, এটাই তর পরম ভাগ্য। আবার

কিনা বকশিশ চাস! যদি বাঁচবার সাধ থাকে, আমার সামনে
থেকে পালা। নইলে এখনি তোর ঘাড় ভাঙব।

বাঘের কথায় ভয় পেয়ে বক তক্ষুনি সেখান থেকে পালাল।

Fig. 10. Extracted Text Displayed on the Monitor.

Fig. 11. Outcome Observation via Headphone.

V. Conclusion and Future Works

This research uses raspberry pi, pi camera, Tesseract OCR engine, etc. to help the people listening to the content of the Bangla text image who are visually impaired or illiterate. It can also be used by any person who wants to listen to the content of the image instead of reading the content of the image. We have achieved 97.4% accuracy for precise Bangla text images. For the middle range of a Pi camera, the quality of the captured image is not so good in low light. During the night the quality of the captured image is obscure. As a result, sometimes the accuracy of the extraction of text is not up to mark.

In the future, we would like to enhance this system by appending the higher resolution Pi camera to increase accuracy for text extraction and by eliminating noise from speech using advanced algorithms. Furthermore, we would like to improve the portability of the system by compacting the hardware design through design improvement and hardware upgrade. We would also like to extend our research to extract text on Bangla handwritten script.

References

[1] "Globaldata on Visual Impairment," Available Online: https://www.who.int/blindness/publications/globaldata/en/ [Last Access: 20 May 2020].

[2] "WHO |Sight savers International," Available Online: https://www.who.int/workforcealliance/members_partners/member_list/si/en/ [Last Access: 20 May 2020].

[3] "Bangladesh Literacy Rate," Available Online: https://countryeconomy.com/demography/literacy-rate/bangladesh [Last Access: 20 May 2020].

[4] J. Arthur, Raspberry Pi: The complete guide to Raspberry Pi for beginners, including projects, tips, tricks, and programming. CreateSpace Independent Publishing Platform, 2017.

[5] S. Mori, H. Nishida, and H. Yamada, Optical character recognition, John Wiley & Sons, Inc., 1999.

[6] C. Patel, A. Patel, and D. Patel, "Optical character recognition by open source OCR tool tesseract: A case study," International Journal of Computer Applications, vol. 55, no. 10, pp. 50-56, 2012.

[7] “Bengali Language,” Available Online: https://www.britannica.com/topic/Bengali-language [Last Access: 20 May 2020].

[8] R. Naveen, S. A. Sivakumar, M.Cholavendan, S. S. Kumar, A. S. K. Singh, S.Thilak, and C. Arunkumar, "Face and text recognition for the visually impaired persons based on Raspberry Pi," International Journal of Advanced Engineering and Research Development, vol. 5, no. 03, pp. 90-94, 2018.

[9] L. Nagaraja, R. S. Nagarjun, M. Nishanth, D. NIthin, and V. S. Murthy, "Vision based text recognition using Raspberry Pi," National Conference on Power Systems & Industrial Automation (NCPSIA 2015), 1-3, 2015.

[10] V. A. Devi and S. S. Baboo, "Embedded optical character recognition on Tamil text image using Raspberry Pi," International Journal of Computer Science Trends and Technology (IJCST), vol. 2, no. 4, 2014.

[11] G. K. Sagar, and T. Shreekanth, "Real time implementation of optical character recognition based TTS system using Raspberry Pi," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 7, pp. 149-156, 2017.

[12] S. A. James, S. Sanjana, M. Monisha, "OCR based automatic book reader for the visually impaired using Raspberry Pi," International Journal of Innovative Research in Computer and Communication Engineering, vol. 4, no. 7, pp. 1111-1118, 2016.

[13] C. A. Todd, A. Rounaq, U. K. Nur, and F. Boufarhat, "An audio haptic tool for visually impaired web users," Journal of Emerging Trends in Computing and Information Sciences, vol. 3, no. 8, pp. 1136-1145, 2012.

[14] D. Velmurugan, Srilakshmi, S. Umamaheswari, S. Parthsarathy, and K. R. Arun, "Hardware implementation of the smart reader for visually impaired people using Raspberry Pi", International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering, vol. 5, no. 3, pp. 2055-2063, 2016.

[15] C. S. T. Thu and T. Zin, "Implementation of text to speech conversion," International Journal of Engineering Research & Technology (IJERT), vol. 3, no. 3, pp. 911-915, 2014.

[16] H. M. Htun, T. Zin, H. M. Tun, "Text to speech conversion using different speech synthesis," International Journal of Scientific & Technology Research, vol. 4, no. 7, pp. 104-108, 2015.

[17] A. Goel, A. Sehrawat, A. Patil, P. Chougule, and S. Khatavkar, "Raspberry Pi based reader for blind people," International Research Journal of Engineering and Technology, vol. 5, no. 6, pp. 1639-1642, 2018.

[18] H. Rithika and B. N. Santhoshi, "Image text to speech conversion in the desired language by translating with Raspberry Pi," 2016 IEEE International Conference on Computational Intelligence and Computing Research, IEEE, 2016.

[19] S. Gollapudi, Learn computer vision using OpenCV, Apress, 2019.

[20] W. Harrington, Learning raspbian, Packt Publishing Ltd, 2015.

IoT Enabled Air Quality Monitoring for Health-Aware Commuting Recommendation in Smart Cities

Riaz UlAmin[1]*
Dept. of CS, University of Okara
Okara, Pakistan

Najeeb Ullah[3]
Dept. of Telecom Engineering
BUITEMS, Quetta, Pakistan

Abdul Sattar Malik[5]
Dept. of Electrical Engineering
BZU, Multan, Pakistan

Muhammad Akram[2]
Dept. of Computer Engineering
BUITEMS, Quetta, Pakistan

Muhammad Ashraf[4]
Dept. of Software Engineering
BUITEMS, Quetta, Pakistan

***Abstract*—The importance of air pollution control in smart cities has been realized by almost every department of society. Research community has been working in collaboration with industry to craft sensors for measuring different types of pollution levels in the environment. However, it is rarely possible to implant the sensors in all geographical areas. It is important to measure pollution levels in almost every part of the world with life and implement clean environment policies. However, in unplanned areas, the implementation of environmental policies faces problems because such areas lack in communication infrastructure and cost of huge amount of fixed or static sensors. This work envisions availability of sensor-equipped-VANET based system to monitor pollution levels in different areas of an unplanned city. This paper proposes an autonomous VANET system that can carry environmental sensors to collect data from an area at different intervals, process it to transform data into information and forward the information to node that has capability to collect all information and then send it to server machine for further process either using VANET or some reliable network connectivity. Based on the collected data, this research further contributes health aware commuting recommendation based on cost effective monitoring of air quality.**

***Keywords*—Un-planned Areas; Vehicular ad hoc Networks; Pollution Monitoring; AQI; Health Aware Commuting Recommendations**

I. Introduction

VANET are networks that can be characterized by their openness in nature, uninterrupted energy availability and high speed mobility. Research on vehicular communication is continuously emerging and evolving. It not only addresses safety of life aspects on the road but it also considers non safety related applications. Classifications of such applications can easily be found in literature [1]. One important application of VANET proposed in this paper is to monitor air pollution particularly in unplanned areas. Unplanned areas are defined as the areas that are attributed by several characteristics. Usually such areas are developed without proper town planning regulations. Thus such areas, in essence, lack in availability of different services and communication infrastructures. Lack of town planning and industrial zone planning not only poses threats to clean environment but also makes conventional air pollution monitoring systems harder to implement. Particularly, in developing countries, with growth in economy, unplanned land development is likely to take place. Growth in industrialization and unplanned land development poses several challenges to the environment. Continuous monitoring of different areas for clean environmental measurements is already a challenge, furthermore, lack of communication infrastructure for environmental data collection and higher cost of large scale fixed sensor network for environment monitoring system, extravagate the problem. On the other hand it is important to monitor such areas for air pollution. This work exploits VANET services as cost-effective means to monitor air pollution of such unplanned areas.

A. *Vehicular Ad -hoc Network (VANET) and Wireless Sensor Networks (WSN)*

VANET are a subclass of MANET that can be equipped with several types of sensors. Equipping vehicles with such sensors help to envision several applications that can be deployed over VANET. One such application introduced in this paper, serves monitoring an unplanned area for air pollution. Importance of air pollution monitoring has been widely realized in society. Electronics engineering research and industrial community has been successfully crafting wide range of sensors to measure different types of air pollutions. The deployment of theses sensors can greatly affect the monitoring system's performance. Majorly the deployment of sensors to for monitoring an area for air pollution can be executed in one of the following three approaches.

1) Static deployment of sensors, the information is periodically collected by these sensors and forwarded to head node using fixed network infrastructure (wired or wireless). The head nodes usually aggregate information, analyze and filters it, and forwards it to decision making node for further process. This approach has several challenges. These challenges include firstly the higher cost for deployment of sensors at large scale area, secondly, energy constrained air may also cause sensor node failures, and lastly but not the least in some cases the objects that needs to be sensed can be

*Corresponding Author

mobile in nature, thus static sensors could face limitation in coverage of object.

2) Mobile sensor based deployment. This approach considers to deploy sensors in explicitly mobile mode, no static/fixed sensors are assumed. These sensors can form groups/clusters. Information from mobile sensors is collected by cluster-heads, and then cluster-head after performing initial filtering forwards the data to some server using network based on cellular/ infrastructure-technology. There are several pros and cons of this deployment technique. This technique offers low cost monitoring of an area. It resolves the monitoring of target object mobility however, scale of its deployment and energy of mobile sensor nodes remains questionable.

3) The Hybrid Approach. This approach for deployment of sensors to monitor an area considers use of both static and mobile nodes. Conventional mobile sensor network following this approach of deployment also considers infrastructure based network availability for communication among sensor nodes and for communication required for data collection. This approach has several advantages over the previous two approaches, such as large scale area coverage, design dependent cost and target object mobility tracking, however, this approach is also not without drawbacks. This approach is complex to implement and if the design of network is not efficient, it may cause higher cost. Similarly, energy of sensor nodes remains a challenge.

II. Literature Review

There are several contributions in literature that are related to this research. In one of such contribution an online GPRS array for air pollution monitoring is proposed with system that consists of mobile sensing unit, a fixed Internet-Enabled Pollution Monitoring Server (Pollution-Server), sensor array (NO_2, CO & SO_2), GPS and GPRS module, which are integrated with each sensing unit. Sensed data from sensor array with location information are sent to the server through GPRS for further handling and investigation. Public can access the authorized air pollution information by customized Web App. The proposed system was tested in the city of Sharjah UAE [2]. The proposed system has limited scale and scope given the use of fixed infrastructure for collection of sensor data. Another similar work a vehicular pollution monitoring system designed based on internet of things (IoT) capable for monitoring pollutant on city road caused by vehicles emission is proposed in [3]. Sensors of different gases are installed at fixed position on city road. The proposed systems also guarantees the presence of wireless sensors for vehicle pollution system that specify in a straight forward accessibility of real time data through internet using IoT. However this system is installed at fixed position and is able for monitoring pollutant from vehicle emission only. A Mobile Air Quality Monitoring Network (MAQUMON), is proposed in which sensor nodes are mounted on cars. These sensor nodes are consisted of different hardware components including Gas sensors for measuring CO, NO2 and ozone concentration in Air. Data samples from sensor are taken differently for different scenarios. Samples are taken very frequently when car is in motion; while it is taken a few times an hour when a car is parked. These Samples are store and tagged with a time information and location. A Wi-Fi hotspot is used for collection of data, after collection of data it is being processed and published on the Sensor Map portal [4]. The research is closely related to our research in terms of use of VANET for collection of sensed data. However, this research further makes commuting recommendations based on the data collected through sensors. The concept of a Vehicular Wireless Sensor Network (VSN) is proposed in [5] in which sensors are deployed on cars for monitoring air quality. In this framework a is vehicle set with a gas sensor of CO2, a GPS module , a GSM module and ZigBee- based intra- vehicle wireless network. The concentration of CO2 in air is measured in an area of interest and the sensed data is reported through GSM short messages and collected by a server. The collected data is integrated with Google Maps for the demonstration of result. However the proposed framework uses an expensive GSM short messaging for reporting. Furthermore the recommendations on the basis of data are not in the scope of this research.

Authors in [6] proposed a low cost solar powered air quality monitoring system based on ZigBee wireless network for a fixed station. The proposed prototype was installed at fixed locations at school surrounding. The hardware prototype was consist of set of sensors array for monitoring Carbon monoxide (CO) ,Nitrogen dioxides (NO2) , relative humidity and temperature at school surrounding. The design was based on Arduino (UNO) and a ZigBee wireless node. The sensed date was communicated through ZigBee wireless node with low power consumption. The sensed data was monitored on desktop/ laptop through an application designed using lab view. However fixed air quality monitoring station has limitations. Because People's real life exposure to air pollutants is not achieved. At a fixed air quality monitoring stations only the pollutants in air at a fixed locations/area can be monitored, which is inefficient to capture the spatial variability in urban or industrial environment. For this purpose a mobile air quality monitoring system can be used to monitor the air quality which solves the entire existing problem associated with fixed air quality monitoring stations. Authors in another research [7] proposed low-cost air quality system for urban area monitoring through mobile sensing and low cost reliable sensors. A theoretical model is presented which involves two types of mobile sensing network, first is "Personal sensing network" and second is "public transport network". A Personal sensing prototype is designed which measure the concentration of Carbon monoxide (CO) in air. Research proposed two sub models. The first sub model is developed for measuring the concentration of CO in stationary places and second sub model is designed for measuring the concentration of CO in mobile situations. The sensed data with time and location information is stored to the flash memory and then uploaded to the local server and results are generated by graphs and heat map. In this paper theoretical model for, personal sensing network and public transport sensing network was proposed. But only one prototype is designed for personal sensing network, which can measure only the concentration of CO in air. However the proposed "Public transport sensing network" for future work, require more sensing devices for air quality monitoring.

A low cost, low power outdoor air pollution monitoring system called Gas Mobile was proposed in [8]. The research proposed prototype consists of Ozone (O3) gas sensor and an off-the-shelf smart phone. An USB interface is used to communicate smart phone with sensor module. Sensed data from the sensor module with time and location information, from build in GPS module is uploaded to sever using cellular network. Public can access the official data from modified web and mobile apps. For improving the sensor data quality two methods were proposed and implemented. The proposed prototype was mounted on bicycle (protected from wind) and the measurements were taken from several bicycles across the city. However in the proposed system only one gas sensor is used for measurement of O3 in atmosphere which is inefficient for measuring over-all air quality. The proposed system is low in data accuracy and reliability as the sensor are mounted on bicycles using handbags, etc. In [9] Authors Proposed Real time Air pollution monitoring system using Mobile phone. In this system the hardware prototype which consist of gas sensors, GPS, Bluetooth modem and a chip microcontroller is installed on the buses to collect the concentration of gases such as CO, N, smoke and temperature. The sensed data from sensors are sent to the control central unit that uploads this data to Internet through a Google Maps interfaced with Bluetooth module. However the proposed system is costly as the sensed data is simultaneously uploaded to the internet and interfaced with Google map. The proposed system fails when there is no availability of internet.

Researchers proposed air quality monitoring through mobile sensing in metropolitan areas [10]. In the proposed work an MSB (Mobile Sensing Box) was designed. The MSB consist of three main units Sensors, GPS module and a mobile phone. Sensors are used for Data collection, time and location information is carried by GPS module and the transmission of data is done by cellular mobile phone. The prototype is mounted on the car which travels around the city. The sensed data is then uploaded to cloud server. Public can accesses authorized air pollution information through web portal. However the proposed work is costly, because each sensed data is uploaded to cloud server using cellular modem.

III. System Design

Research presented in this paper considers implanting air monitoring sensors in vehicles. In theory it resembles the approach 2 presented in the previous section, however, it has advantage of no energy limitation of sensors and scale of deployment of sensor nodes. As the area being monitored scales with vehicle's mobility in an area. This approach also permits to collect and share data at cluster head without any fixed network infrastructure. The information collected at different sensor nodes is forwarded to central collection point at cluster head node using VANET communication protocol such as IEEE 802.11p. This approach reduces the cost of network used as IEEE 802.11p is free to use and an efficient communication among vehicle can be executed following this standard.

This hardware prototype shown in Fig. 1 consists of several components. Two gas sensors are connected to Arduino Mega 2560, which measure the concentration level of CO and NO2 in air. The Sensors are programmed in a way that they measure the concentration levels of gases with interval of every two seconds. A GPS module is connected to Arduino board which gives information about locations in form of Latitude and Longitude.

For taking time and date information, an RTC module is connected to Arduino board. A storing unit is connected to Arduino board which stores all the data in the form of string, including the data from sensors, GPS and RTC. A XBEE module is connected to Arduino board which transmits the stored data to a server. To minimize the cost, the proposed system considers using storing unit for storing and transmitting data via XBEE module to server. Once data is received at server, it further stores it to a separate database for records. Through designed queries the data can be filtered and sorted for days, months and years. The hardware prototype designed and implemented in this research is portable and can be connected to any vehicle at different positions. In this case, the hardware prototype was connected to vehicles powered by vehicle battery.

Vehicles can have their route spanned over a large area. These areas may be monitored for real time concentration levels of several gases in air. This hardware prototype can also be used as a personal sensing device with addition of on board external battery. The hardware prototype was tested to perform both in stationary and mobile scenarios.

Integration of VANET and wireless sensor networks (WSN) has already been presented in literature. However to best of knowledge, it is first contribution to monitor an unplanned area with the help of VANET based sensor network for air pollution. There are several types of air pollutions, however, this research considered measuring NO2 in the air by using this Air quality measurement module mounted in university buses measuring NO2 in different areas of the city. Fig. 2 shows the general methodology taken in this research.

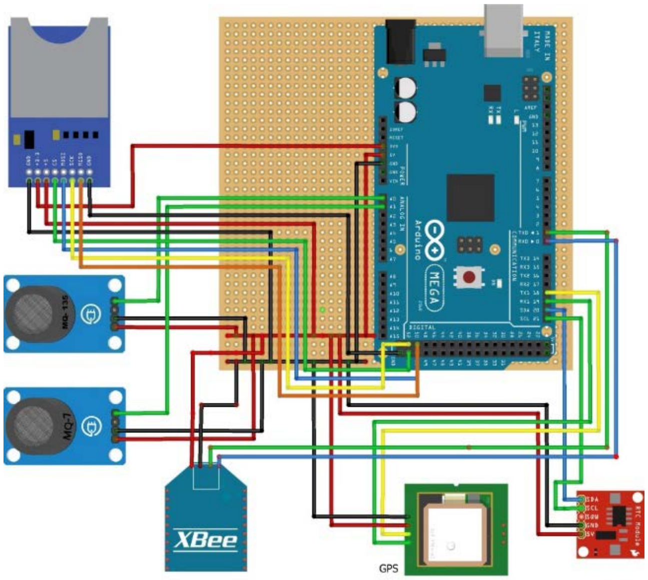


Fig. 1. Block Diagram of Hardware Prototype.

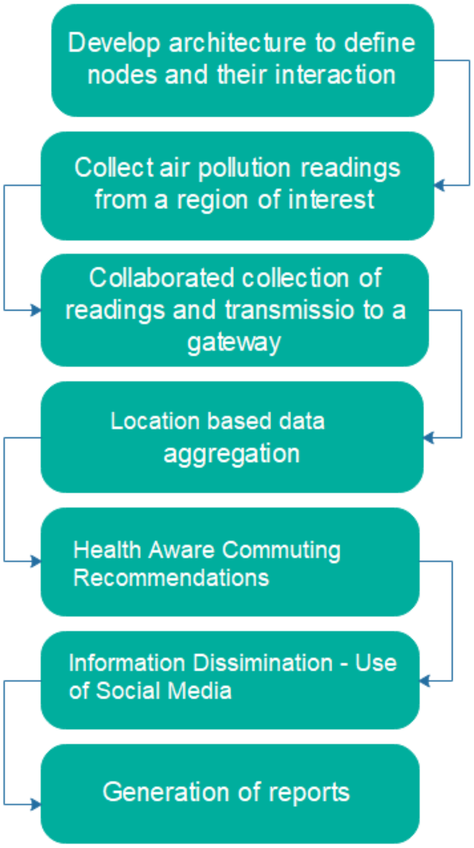

Develop architecture to define nodes and their interaction
Collect air pollution readings from a region of interest
Collaborated collection of readings and transmissio to a gateway
Location based data aggregation
Health Aware Commuting Recommendations
Information Dissimination - Use of Social Media
Generation of reports

Fig. 2. Block Diagram of System Design.

The data was clustered based on location of measurements. Locations with measurements of NO2 were recorded using GPS module integrated within this prototype Air quality measurement modules. The data was cleansed and clustered using location information. The measurements were further grouped into time slots. Each time slot consisted of 3600 Seconds i.e. One hour. Readings of NO2 within each time slot and location were averaged. The following Table I shows the categories of NO2 measurements in the air.

The research contributed to design an Android application that is capable to make health aware commuting recommendations to the public, based on the measured values of NO2 and CO in the city and the NO2 and CO categories available in table. In addition, the application can broadcast this information on different social media platforms.

TABLE I. AQ RANGES FOR NO2[11]

Range	Category
(0-50)	Good
(51-100)	Moderate
(101-150)	Unhealthy for Sensitive Groups
(151-200)	Unhealthy
(201-300)	Very Unhealthy

IV. RESULTS AND DISCUSSION

This application was available for download in Google Play store and through effective publicity. It was downloaded for more than 5000 times. Users were offered to select their user group. Users are categorized in the following groups [11].

- People with lung disease, such as asthma.
- Children and older adults.
- People who are active outdoors.

The sensor module measured the concentration of air pollutants dynamically in different areas. There were multiple copies of this module mounted on University buses which traverse throughout the city in different times of the day. This helped us to collect data from almost all populated areas of the city in different times. An intensive data cleansing was required such as to remove duplicate and incorrect values at times. The results taken from different geographic positions (locations) are shown below in graphs.

The results shown in Fig. 3 to Fig. 17 reflects substantial variation in concentration of CO and NO2.

This research considered five routes in the city and buses commuted to these routes every hour from 7:00am to 7:00pm. Based on the reading of CO and NO2. For each route the Air quality index (AQI) was calculated. Results of these routes are shown in Fig. 18. From the results, it can be observed that during peak hours 8:00am to 10:00am and 4:00pm to 6:00pm the concentration of NO2 remains in the unhealthy range. Whereas during off peak times, there are opportunities for people here they can travel on these routes. This application considering health issues maps and recommends its users to commute through these routes. It is highly recommended for people with lungs diseases and children to not commute through areas with high concentration of NO2 in the air.

Based on the value of CO and NO2 concentration at particular times, this Recommender made one of the following four recommendation:

1) Commute at selected time slot (Safe)
2) Do not commute at selected time slot
3) Take alternate route
4) Do not commute at any time slot any route.

The availability of this application to the general public along with information related to the association of concentration of CO and NO2 with potential diseases helps citizen to take travel decisions in these areas throughout the day.

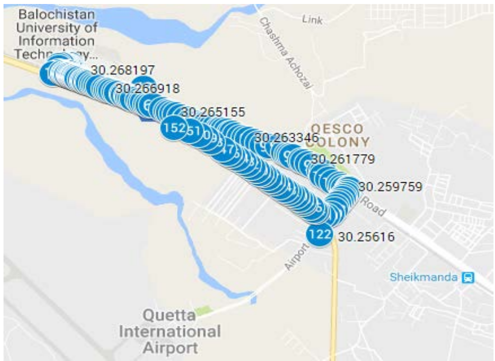


Fig. 3. Bus Route-A Location of Data Readings.

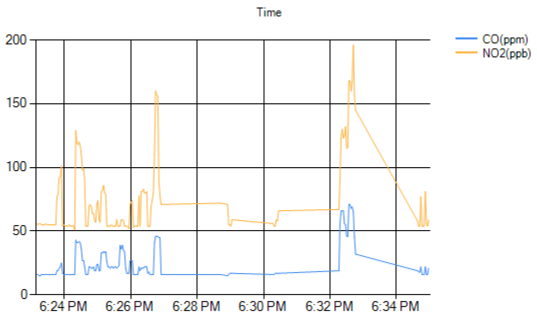


Fig. 4. Concentration Level Graph of CO (ppm) & NO2 (ppb) Verse Time, Measured by Hardware Deployed in Vehicles.

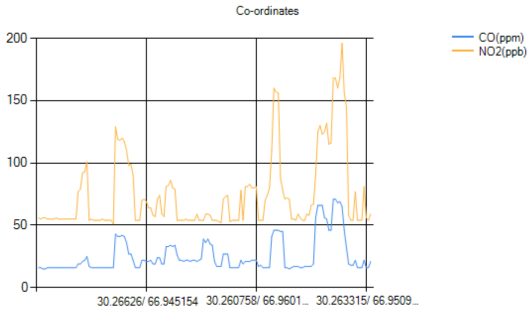


Fig. 5. CO & NO2 Values Verses GPS Co-Ordinates, Measured by Hardware Deployed in Vehicles.

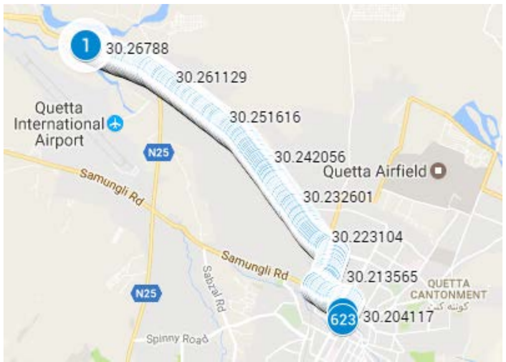


Fig. 6. Bus Route-B Location of Data Readings.

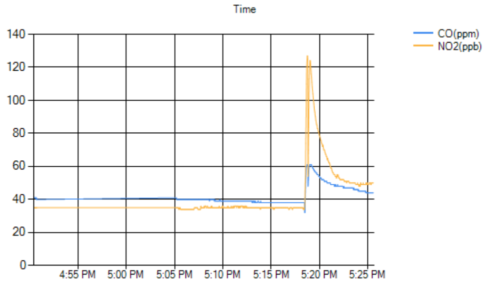


Fig. 7. CO & NO2 Values Verses GPS Co-Ordinates, Measured by Hardware Deployed in Vehicles.

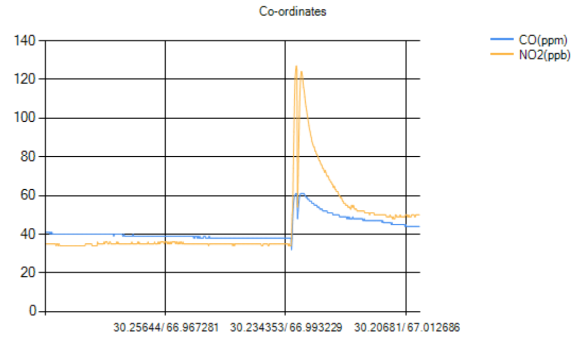


Fig. 8. CO & NO2 Values Verses GPS Co-Ordinates, Measured by Hardware Deployed in Vehicles.

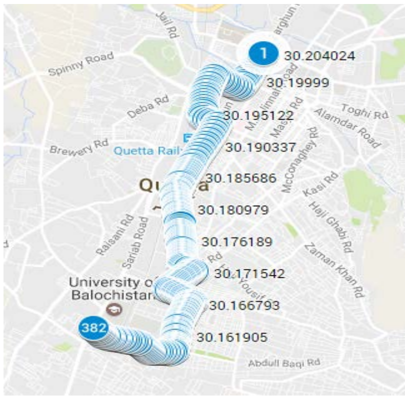


Fig. 9. Bus Route-C Location of Data Readings.

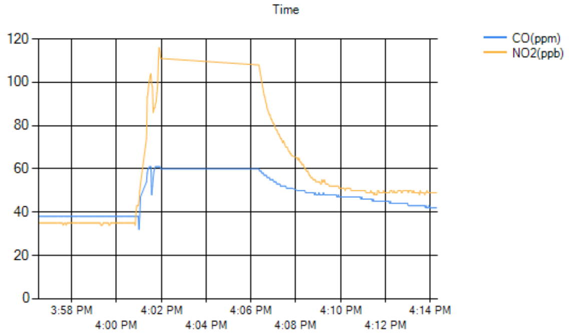


Fig. 10. CO & NO2 Values Verses GPS Co-Ordinates, Measured by Hardware Deployed in Vehicles.

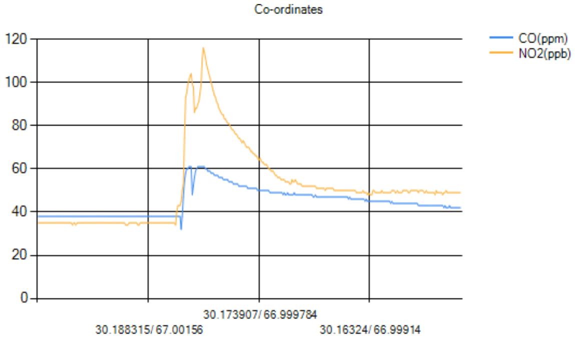


Fig. 11. CO & NO2 Values Verses GPS Co-Ordinates, Measured by Hardware Deployed in Vehicles.

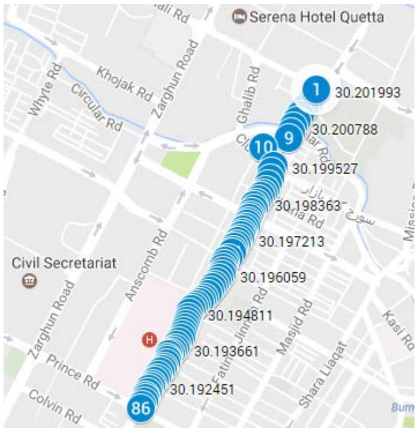


Fig. 12. Bus Route-D Location of Data Readings.

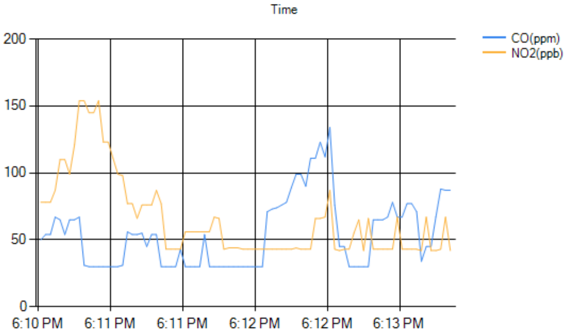


Fig. 13. CO & NO2 Values Verses GPS Coordinates, Measured by Hardware Deployed in Vehicles.

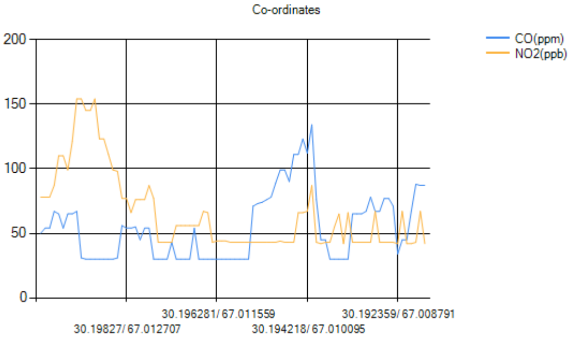


Fig. 14. CO & NO2 Values Verses GPS Coordinates, Measured by Hardware Deployed in Vehicles.

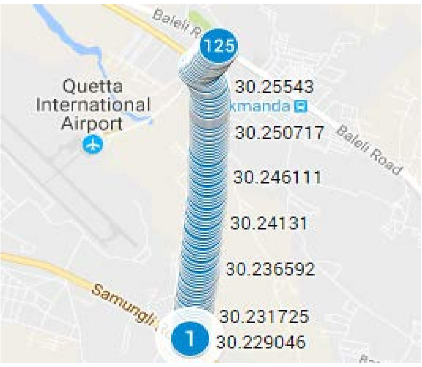


Fig. 15. Bus Route-E Location of Data Readings.

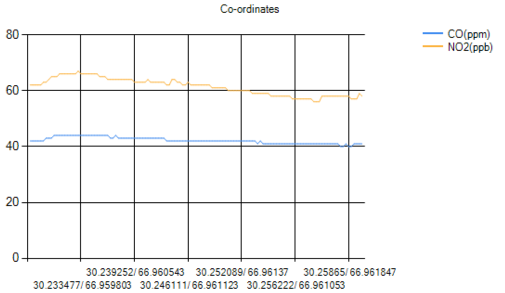


Fig. 16. CO & NO2 Values Verses GPS Coordinates, Measured by Hardware Deployed in Vehicles.

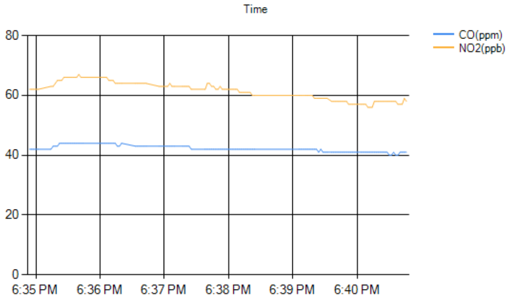


Fig. 17. CO & NO2 Values Verses GPS Coordinates, Measured by Hardware Deployed in Vehicles.

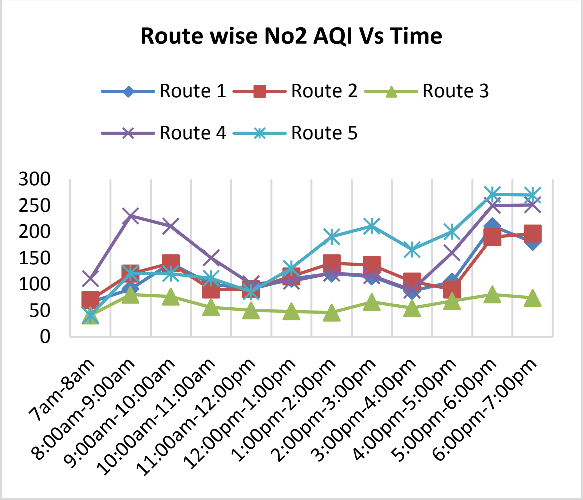


Fig. 18. Routes and NO2 AQI Against Time Slots.

V. Conclusion and Future-Work

This research contributes health aware commuting recommendation based on cost effective monitoring of air quality. This research has presented a case study of Quetta city while dividing it in five routes and measuring AQI and making commuting recommendations through android application. It is found that as effective approach to avail the opportunity to commute in any area of interest hence there is low level of NO2. The approach may be tested and in mega city with more routes, users and multiple sensors to measure other contributing factors to AQI. The Application in future may be further integrated with multiple social media platforms to create public awareness and helping them to take timely commuting decisions.

REFERENCES

[1] Ul-Amin R, ”Cooperative & cost-effective network selection: a novel approach to support location-dependent & context-aware service migration in VANETs”, PhD Thesis, University of Glasgow URL: https://theses.gla.ac.uk/.

[2] R. Al-Ali, I. Zualkernan, and F. Aloul, “A mobile GPRS-sensors array for air pollution monitoring,” IEEE Sens. J., vol. 10, no. 10, pp. 1666–1671, 2010.

[3] R. Rushikesh, C. Mohan, and R. Sivappagari, “Development of IoT based Vehicular Pollution,” pp. 779–783, 2015.

[4] P. Völgyesi, A. Nádas, X. Koutsoukos, and Á. Lédeczi, “Air Quality Monitoring with SensorMap,” 2008 Int. Conf. Inf. Process. Sens. Networks (ipsn 2008), pp. 529–530, 2008.

[5] S. C. Hu, Y. C. Wang, C. Y. Huang, and Y. C. Tseng, “A vehicular wireless sensor network for CO2 monitoring,” Proc. IEEE Sensors, pp. 1498–1501, 2009.

[6] S. R. W. H.Ali, J.k. Soe, “A Real-Time Ambient Air Quality Monitoring Wireless Sensor Network for Schools in Smart Cities,” pp. 5–10.

[7] A.-C. Firculescu and D. S. Tudose, “Low-Cost Air Quality System for Urban Area Monitoring,” 2015 20th Int. Conf. Control Syst. Comput. Sci., pp. 240–247, 2015.

[8] D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele, “Participatory Air Pollution Monitoring Using Smartphones,” 2nd Int. Work. Mob. Sens., pp. 1–5, 2012.

[9] G. V. Bhagwan and V. G. Puranik, “Real Time Air Pollution Monitoring Using Mobile Phone,” vol. 3, no. 4, pp. 447–453, 2014.

[10] S. Devarakonda, P. Sevusu, H. Liu, R. Liu, L. Iftode, and B. Nath, “Real-time air quality monitoring through mobile sensing in metropolitan areas,” Proc. 2nd ACM SIGKDD Int. Work. Urban Comput. - UrbComp ’13, p. 1, 2013.

[11] Air Quality Guide for Nitrogen Dioxide, URL: https://www3.epa.gov /airnow/NO2.pdf.

ER Model Partitioning: Towards Trustworthy Automated Systems Development

Dhammika Pieris[1]
Department of Human Resource Management
Faculty of Commerce and Management Studies
University of Kelaniya
Sri Lanka

M. C Wijegunesekera[2]
Department of Software Engineering
Faculty of Computing and Technology
University of Kelaniya
Sri Lanka

N. G. J Dias[3]
Department of Computer Systems Engineering
Faculty of Computing and Technology
University of Kelaniya
Sri Lanka

***Abstract*—In database development, a conceptual model is created, in the form of an Entity-relationship (ER) model, and transformed to a relational database schema (RDS) to create the database. However, some important information represented on the ER model may not be transformed and represented on the RDS. This situation causes a loss of information during the transformation process. With a view to preserving information, in our previous study, we standardized the transformation process as a one-to-one and onto mapping from the ER model to the RDS. For this purpose, we modified the ER model and the transformation algorithm resolving some deficiencies existed in them. Since the mapping was established using a few real-world cases as a basis and for verification purposes, a formal-proof is necessary to validate the work. Thus, the ongoing research aiming to create a proof will show how a given ER model can be partitioned into a unique set of segments and use it to represent the ER model itself. How the findings can be used to complete the proof in the future will also be explained. Significance of the research on automating database development, teaching conceptual modeling, and using formal methods will also be discussed.**

***Keywords*—Conceptual model; Entity Relationship (ER) model; relational database schema; information preservation; transformation**

I. Introduction

The Entity-Relationship (ER) model[1, 2] is widely used to create conceptual schemas (conceptual models) to represent application domains in the field of Information Systems development. However, when an ER model is transformed to a Relational Database Schema (RDS) of the relational model, some critical information modeled on the ER model may not be represented meaningfully on the RDS [3-5]. This situation causes a loss of information during the transformation process [5, 6].

Min-max constraints, role names, composite attributes, subtype/supertype hierarchies, and certain relationship types are frequently lost in the transformation process [5][13].

Previous studies undertaken by other researchers on information loss [6-11] were of varying opinion. Some researches proposed ignoring the information that is lost during the transformation process and accepting only the information, that is, actually transformed. This proposal is called information reducing transformation (e.g., [8, 9].) Researches in [7] and [10] suggested that the min-max constraints that cannot be transformed and represented on the RDS to be directly implemented in the database system via triggers and stored procedures. This is a way of bypassing the RDS. According to [11], min-max constraints can be represented as a set of functions in a separate schema, external to the RDS. The functions are then implemented in the database as a program written in extended SQL (e.g., PL/SQL or T/SQL). The method is also a way of bypassing the RDS. The research in [6] indicated that supertype/subtype hierarchies that could be lost during a transformation could be directly implemented in the database system. It is also a way of bypassing the RDS. As indicated in [10] and [11], min-max constraints can be directly implemented in user application programs. It is a way of bypassing the logical level RDS as well as the physical level database.

In summary, some previous research suggests bypassing the logical level−that is, the RDS−and implementing the lost information directly on the physical level. Some others suggest bypassing both the logical level and the physical level and implementing the lost information directly in user application programs. Some other researchers proposed ignoring the information that is lost during the transformation process, suggesting that the information that is actually preserved is adequate.

However, in contrast to bypassing the RDS and ignoring the lost information, in our study, we focus on preserving information and representing them on the logical level RDS as much as possible.

According to [12], if the information is preserved when a conceptual schema (e.g., ER model) is transformed to a logical schema (e.g., RDS) (forward transformation), the logical schema should be able to reverse back to the conceptual schema (reverse transformation) by means of reverse applying the steps of the algorithm used for the forward transformation process. We based our research on this theory proposed by [12].

We argue that if the forward transformation can create a one-to-one and onto mapping from the ER model to the RDS, the RDS could be reversed back to the ER model. The RDS could be reversed back to the ER model means, according to [12], the information is preserved in the transformation process from the ER model to the RDS.

However, during our previous studies, we found that the deficiencies that exist in the ER model and the transformation algorithm hinder such a one-to-one and onto mapping is being established in the forward transformation process. [5, 13] [14-16]. We then modified the ER model and the transformation algorithm [5, 14, 15], eliminated the deficiencies and avoid that hindrance. Accordingly we established a one-to-one and onto mapping in the forward transformation process. We wish to generalize the work and prove it formally.

It is necessary to show that the concept can be applied to any ER model representing any application domain. On the other hand, a formal proof that can justify the accuracy of the system is an essential goal in Computer Science [17].

The current work aims to show that a one-to-one and onto mapping, as defined in mathematics, exists from an ER model diagram (also called an “ER model”) to its RDS. The ER model diagram is created using the modified ER model and transformed to RDS using the modified transformation algorithm. For this purpose, we need to show that a given ER model and its RDS can be expressed as sets.

In the current work, we show that an ER model can be expressed as a set, and the set can be used as a representation of the ER model itself. For this purpose, we use a generic ER model–one that represents phenomena in symbolic notation. A generic ER model can be used as a general representative for exemplifying any ER model from any application domain[13]. We show that the generic ER model can be partitioned into unique segments that each one can represent a meaning in the real world. We call them ER-construct-units and show that such a unit cannot be divided further into smaller units while retaining its meaning. We then show that the set of ER-construct-units of the ER model can be used to represent the ER model itself.

A. *Significance of the Research*

The traditional ER model uses conventional graphical constructs to create ER model diagrams. Accordingly, a rectangle is used to represent an entity type, an oval is used to represent an attribute, and a diamond is used to represent a relationship type. The traditional ER model is regarded to be providing a true natural representation of the real world. The model is still popular and widely used for conceptual modelling of databases as well as teaching and learning the database design process (some recent examples for its use, in practice and research, are: [18-20]).

What we have modified is the traditional ER model. As a result, of the modifications introduced to the traditional ER model and the transformation algorithm, a one-to-one correspondence is established from any ER model diagram created by the modified ER model to the RDS created by the modified transformation algorithm. We argue that, if this modified approach is used, the database designing process will become a much more natural, straightforward, momentary, and trustworthy task for its learners, teachers, and practitioners.

Many automated tools are available for creating ER models for the traditional ER model and its variants. However, no such tool exists to provide a real automatic transformation from the traditional ER model to the RDS. Some tools claiming to be providing an automated transformation can only help the user visualize what he/she is doing with the computer. The user has to transform the ER model diagram to the RDS manually using pointing and clicking devices. The user can monitor and, if necessary, rectify what he/she is doing in the computer. In contrast, we argue that our modified database design approach can provide a high level of and a true nature of automation to the transformation process. Once the ER model diagram is produced, to transform it to the RDS is just a one-click away action. Thus, we believe a Computer-Aided Software Engineering tool (also called CASE tool) could be produced based on our modified approach to automate the transformation process. Tools that are limited to creating ER model diagrams only could also be extended to provide a true automated transformation. We also believe such a CASE tool that we expect will equally enhance the teaching and learning process of database design.

The current research seeks to develop a formal method and use it to validate a systems development method proposed. Thus, we hope the research will contribute significantly to the area of formal methods in software engineering.

B. *Related Research*

Kamišalić et al. [21] examine the effectiveness of learning conceptual database design. They found that the manual transformation from a conceptual model to a logical data model can increase students' understanding of the concept. Khaire and Mali [22] presented a web application that can assist in generating an ER diagram automatically. The application needs the user to fill a form it provides to get entities, attributes, and relationships in the application domain as inputs. It then gives the ER diagram as output, automatically[22]. Kuk et al. [23] also present a semi-automated method for generating an ER model from requirements stated in a natural langue. Javed and Lin [24] also undertook a similar study. The method they investigated could generate ER models automatically from requirements stated in a natural language [24]. Yang and Cao, [25] investigated how MySQL Workbench −a visual tool for data modelling −can be used for helping students improve their performance in the ER model to RDS transformation. They also investigated the effects of using MySQL Workbench, in teaching ER to relational transformation. The authors found that visualization of the transformation process could increase the students’ interest in it and their engagement with it, as well as their ability to transform the ER model’s concepts to the RDS [25]. Wu et al. [26] investigated several versions of the ER model to understand the right ER diagram convention used to teach ER modelling to undergraduate students. Accordingly, they investigated the traditional ER model, the Bachman ER model − the ER model in Bachman notation, and the UML class diagram. The authors found that the traditional ER model is much better than any other model they investigated to introduce ER modeling concepts to students [26].

We will show how our standardized ER to relational transformation process can enhance the above findings. However, the main objective of this paper is to validate formally the standardization that we had undertaken. Thus, with that view in mind, we organize the rest of this paper as follows. In Section II, we explore how a real-world small ER model can be partitioned and its ER-construct-units identified. In Section III, we deal with a generic ER model and define the ER-construct-units discussed in Section II. Section IV extends the work done in Section II with a larger ER model. ER-construct-units found in Section IV are defined in Section V. Section VI presents the conclusion, while Section VII details future research.

II. Partitioning a Real-World ER Model into Segments

An ER model is a conceptual schema represented as a diagram drawn using ER constructs such as entity types, attributes, and relationship types. It is intended to represent a user application domain in the real world.

On an ER model, the ER constructs do not exist in isolation separated from each other. Still, they exist connected logically as an arrangement that portrays a real-world meaning relevant and vital to the application domain concerned.

For instance, a regular (strong) entity type, including its attributes, is an ER construct arrangement. Fig. 1 shows a regular entity type, which is made up of three ER constructs in such a way that (i) - a primary key(PK) attribute ER construct (Emp_No”), and (ii) - a simple attribute ER construct (“Name”) are connected to (iii) - a regular entity type ER construct (“Employee”). The ER model that contains the regular entity type is drawn for representing a portion of a “Company” user application domain.

We argue that the three constructs are the minimum requirement for a regular entity type to be constructed for any application domain, not only for a “Company” application domain. Thus, what is presented in Fig. 1 is the smallest possible regular entity type arrangement. Therefore, it cannot be split or any of its three constructs removed. For instance, if its simple attribute or the PK attribute is removed, the remainder would become meaningless. Hence, each of the three constructs, the PK attribute, the simple attribute, and the regular entity type are mandatory and should exist connected as a single coherent arrangement regardless of the application domain concerned. Therefore, we consider the arrangement to be a single unit of ER constructs.

Even though Fig. 1 regular entity type, which we consider a single unit of ER constructs, cannot be split, it can be expanded by adding one or more simple attributes. For instance, the regular entity type in Fig. 2 expands the regular entity type in Fig. 1 by adding two more simple attributes: “Address” and “Gender.” Thus, the regular entity type in Fig. 1 acts as a base and allows other attributes to be added to it. In this context, we consider this single unit of ER constructs to be a base unit of ER constructs. Since it is of a regular entity type, we consider it and call it Regular-entity-base-ER-construct-unit.

Further, we call the simple attributes that are added to this Regular-entity-base-ER-construct-unit the secondary simple attributes. We call the secondary simple attributes the Regular-entity-secondary-simple-attribute-ER-construct-unit attached to a Regular-entity-base-ER-construct-unit.

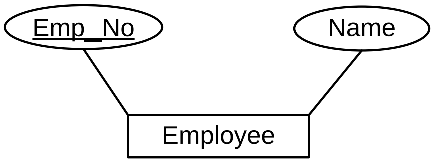


Fig. 1. A Regular Entity Type with Two Simple Attributes.

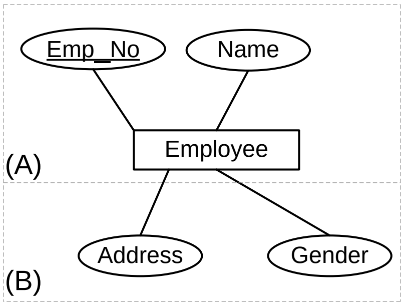


Fig. 2. (A) -The base Regular Entity Type unit, and (B) -the Secondary Simple Attribute unit that are Separated.

Both the Regular-entity-base-ER-construct-unit and the Regular-entity-secondary-simple-attribute-ER-construct-unit are shown partitioned and labelled as (A) and (B), respectively, in Fig. 2. Further, Fig. 2-(B) shows how this Regular-entity-secondary-simple-attribute-ER-construct-unit exists attached to the Regular-entity-base-ER-construct-unit (Fig. 2-(A)).

Next, in section III, we will generalize the concept using a generic ER model proposed by [13].

III. Partitioning a Small Generic ER Model and Defining its ER-Construct-units

In the generic ER model [13], the letter "e" represents a regular entity type. Consequently, e_i represents the i^{th} regular entity type, where $i \in \mathbb{N} = \{1, 2, 3 \dots\}$. Further, $k(e_i)$ represents the primary key (PK) attribute. The symbol $s_j(e_i)$ represents the j^{th} simple attribute where, $j \in \mathbb{N}$. Accordingly, the symbols $s_1(e_i)$, $s_2(e_i)$, and $s_3(e_i)$,…, $s_n(e_i)$, represent the 1^{st}, 2^{nd}, and n^{th} simple attributes of the entity type e_i. The Fig. 3, represents a generic ER-model of this nature. Notice that we reserve the notation, $s_1(e_i)$, to represent the mandatory simple attribute (section II).

In the generic ER model (Fig. 3), the partition named $b(e_i)$ shows the generic equivalent of the Regular-entity-base-ER-construct-unit, the one we showed in the partition (A) in the real-world ER model (Fig. 2)(section II). Accordingly, we formally define the first ER-construct-unit as follows.

A. Definition 1

In a generic ER model, a regular entity type, e_i, its key attribute, $k(e_i)$, and its mandatory simple attribute, $s_1(e_i)$, taken together, is defined as an ER-construct-unit and named as the “Regular-entity-base-ER-construct-unit” and denoted as $b(e_i)$. The unit is shown partitioned and named as $b(e_i)$ in the generic ER model in Fig. 3. Here, the letter “b” indicates “base.”

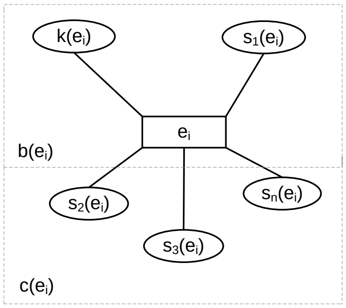


Fig. 3. A Generic ER Model that Represents a Regular Entity Type.

The unit is independent and can exist itself meaningfully. It has a semantic meaning itself. The unit acts as a base and lets other constructs to be attached to it.

In the generic ER model (Fig. 3), recall that we reserved the symbol, $s_1(e_i)$ to denote the mandatory simple attribute of the entity type e_i . Therefore, we denote a secondary simple attribute by $s_t(e_i)$, where $t \geq 2$. For instance, a set of $n - 1$, where $n > 1$ number of secondary simple attributes of a regular entity type, e_i can be denoted as $s_2(e_i)$, $s_3(e_i)$, …, $s_n(e_i)$.

In the generic ER model (Fig. 3), the partition named $c(e_i)$ shows the generic equivalent of the Regular-entity-secondary-simple-attribute-ER-construct-unit. It is the one we have shown in the partition (B) in the real-world ER model (Fig. 2) (Section II). Accordingly, we define the ER-construct-unit, as follows.

B. Definition 2

In a generic ER model, the collection of the secondary simple attribute constructs, $\{s_t(e_i) / \ t \geq 2, t \in \mathbb{N}\}$, connected to a Regular-entity-base-ER- construct-unit, $b(e_i)$ is defined as an ER-construct-unit and named as the "Regular-entity-secondary-simple-attribute-ER-construct-unit" and denoted as $c(e_i)$ (Fig. 3). The unit is shown partitioned and named as $c(e_i)$ in the generic ER model in Fig. 3. The letter "c" in $c(e_i)$ indicates the meaning "se<u>c</u>ondary." The unit, $c(e_i)$, itself does not provide any semantic meaning when it is taken alone. It provides a meaning only when it is attached to a relevant Regular-entity-base-ER-construct-unit, $b(e_i$. It always depends on its base unit, $b(e_i)$, for existence.

Fig. 3 shows how a regular entity type, e_i, in a generic ER model can be partitioned into two ER-construct-units, named, $b(e_i)$, and $c(e_i)$. It also shows how the two units:$b(e_i)$ and $c(e_i)$, can exist associated with each other and form the segment that consists of the regular entity type, e_i and the attributes connected to it, in a generic ER model. The two units forms a set: $\{ \ b(e_i), c(e_i) \ \}$. We assume the set can be used to represent the generic ER model in Fig. 3 that contains the regular entity type, e_i.

IV. Partitioning an ER Model Including a Relationship Type and Identifying its ER-Construct-units

In this section, we consider an ER model with a relationship type and then identify and partition its ER-construct-units.

Consider the real-world ER model given in Fig. 4 that represents two regular entity types, "Vehicle" and "Project," and a relationship type "AssignedTo" existing in between them. A relationship type like AssignedTo where only two entity types participate in is called a relationship type of degree two. A degree two relationship type like AssignedTo is called a binary relationship type [2]. Notice that in the current work, we only deal with binary relationship types existing in between two different regular entity types. We do not consider recursive relationship types, in the current work.

The ER model in Fig. 4 shows min-max structural constraints on the association of the two entity types with each other via the relationship type. They are shown as two bracketed pairs of values (min, max), as (m_1, x_1) and (m_2, x_2). The pair (m_1, x_1) is placed in between the entity type Vehicle and the relationship type AssignedTo, while (m_2, x_2) is placed in between the entity type Project and the relationship type. We will define and discuss the functionality of the two bracketed (min, max) pairs following how min-max structural constraints have been presented in the literature (e.g., [2]).

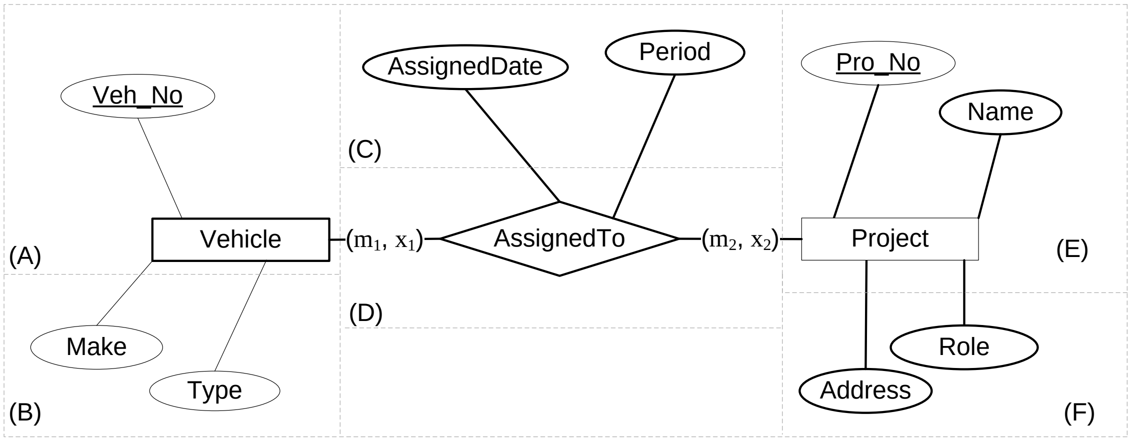


Fig. 4. An ER Model that Contains a Binary One-to-many Relationship Type and Some Attributes Attached to it.

Accordingly, the pair of variables: m_1 and x_1 lie in the range: $0 \leq m_1 \leq x_1$ and $x_1 \geq 1$, while the pair m_2 and x_2 lie in the range: $0 \leq m_2 \leq x_2$ and $x_2 \geq 1$. Variables: m_1 and m_2 represent minimum (min) values, while x_1 and x_2 represent maximum (max) values, in their respective ranges. The number m_1, in (m_1, x_1) means an entity in the entity type Vehicle should participate (via the relationship type AssignedTo) in a minimum m_1 number of entities of the entity type Project. The constraint is called the participation constraint. Notice that the number m_2 in (m_2, x_2) also bears a similar meaning.

On the other hand, the numbers x_1 in (m_1, x_1) and x_2 in (m_2, x_2) represent another constraint called cardinality ratio constraint. The constraint is expressed categorizing into three types as one-to-one, one-to-many, and many-to-many, and from one direction of the relationship type to the other.

To understand the participative constraint and the cardinality ratio constraint let us consider the following example (Example 1)─a pair of min-max structural constraints:

[(m_1, x_1) , (m_2, x_2)] ≡ [(0,3) , (1,1)]

Where, $m_1 = 0, x_1 = 3, m_2 = 1, x_2$ = 1

For instance, m_1 represents participation constraint, and $m_1 = 0$ means some entities in the entity type Vehicle may not participate in the relationship type AssignedTo and hence not associate with any entity in the entity type Project. In this case, the participation of the entity type Vehicle in the relationship type AssignedTo is called "partial" or "optional." Similarly, $m_2 = 1$ means every entity in the entity type Project can exist only if it participates in at least one AssignedTo relationship type instance with an entity in the Vehicle entity type. In this case, the participation of the entity type Project in the relationship type AssignedTo is called "total" or "mandatory."

On the other hand, $x_1 = 3$ and $x_2 = 1$ indicate a one-to-many cardinality ratio constraint, which exists in the direction from the entity type Vehicle to the entity type Project. It means an entity in the entity type Vehicle can relate with minimum 0 and maximum 3 entities in the entity type Project, but an entity in the entity type Project can relate with only one entity (maximum) in the entity type Vehicle.

Table I summarizes two more examples (Example 2 and Example 3) of min-max structural constraints. Example 2 presents a one-to-one cardinality ratio constraint, while Example 3 presents a many-to-many constraint. Notice that Example 1, mentioned above, has already presented a one-to-many constraint.The binary relationship type consists of the ER constructs: (i)- the relationship type construct "AssignedTo" attached to two regular entity types, "Vehicle" and "Project" and (ii)-a pair of min-max structural constraint constructs denoted by two bracketed pairs of values: (m_1, x_1) and (m_2, x_2). Each pair is placed on either side of the relationship type.

Assume any of the constructs: (i) or (ii), mentioned above, does not exist in the structure. Then the relationship type may not exist, and the remainder may become meaningless. Therefore, for a meaningful relationship type to exist, both constructs must exist with binding together and acting as a single unit.

TABLE I. SUMMARY OF TWO MORE STRUCTURAL CONSTRAINT EXAMPLES

Participative constraint		Cardinality ratio constraint	
Example 2			
m_1	m_2	x_1	x_2
1	0	1	1
mandatory /total	partial/optional	one-to-one	
Example 3			
m_1	m_2	x_1	x_2
1	2	3	5
mandatory /total	mandatory /total	many-to-many	

Two simple attributes: "AssignedDate" and "Period" are attached to the relationship type AssignedTo in Fig. 4. They are optional attributes. That is, they may or may not exist.

Thus, we consider the relationship type consisting of the relationship type construct and the min-max structural constraint construct to be a separate ER-construct-unit.

Since the attributes can sometimes exist attached to the relationship type, the relationship type acts as a base and allows other constructs (attributes) to be attached to it. In this context, we deem the relationship type to be a base ER-construct-unit.

The relationship type exists attached to two Regular-entity-base-ER-construct-units. If the two Regular-entity-base-ER-construct-units do not exist, the relationship type does not exist. Thus, the relationship type is a dependent unit that depends on the two Regular-entity-base-ER-construct-units. Accordingly, the relationship type ER-construct-unit depends on the Regular-entity-base-ER-construct-units for its existence. In the meantime, it acts as a base and allows other constructs (attributes) to be attached to it.

We name the relationship type to be a Binary-relationship-type-ER-construct-unit. Notice that the unit is separated and highlighted by a dashed line and labelled as (D) in Fig. 4.

The attributes attached to the relationship are optional. That is, they may or may not exist attached to the relationship type. Even if they exist, the number of them varies. Thus, the simple attributes attached to the relationship type seems to have a particular behavior inherent to them. Therefore, we consider the simple attributes attached to a Binary-relationship-type-ER-construct-unit to be a separate ER-construct-unit. We call the unit a Simple-optional-attribute-ER-construct-unit attached to a Binary-relationship-type-ER-construct-unit. Notice that this unit is separated by a dashed line and labelled as (C), on the ER schema, in Fig. 4.

The generic equivalents of the ER-construct-units: (C) and (D) in Fig. 4 will be defined in the next section.

V. PARTITIONING A MODERATE LEVEL GENERIC ER MODEL AND DEFINING ITS ER CONSTRUCT UNITS

For this purpose, we again use the generic ER model proposed by [13]. The generic ER model uses the symbol, $r_v(e_i, e_j)$, where $v \in \mathbb{N}$, for denoting a binary relationship type existing between two regular entity types e_j and e_j Attributes attached to the relationship type are denoted as $s_1(r_v(e_i, e_j))$, $s_2\left(r_v(e_i, e_j)\right)$,…, $s_t(r_v(e_i, e_j))$, where $t \in \mathbb{N}$. The min-max values are denoted as variables: m_1, x_1, m_2,.and x_2 Fig. 5 shows a binary relationship type existing in a generic ER model.

In the generic ER model (Fig. 5), the partition named $b(\, r_v(e_i, e_j))$ shows the generic equivalent of the Binary-relationship-type-ER-construct-unit, which we have shown in the partition (D) in the real-world ER model (Fig. 4). Accordingly, we formally define the ER-construct-unit as follows.

A. Definition 3

In a generic ER model, the arrangement that consists of the two ER constructs: (i) − a relationship type construct, $r_v(e_i, e_j)$, which is attached to two regular entity base ER construct units, $b(e_i)$ and $b(e_j)$, and (ii) − a min-max structural constraint construct denoted by two bracketed pairs of values: (m_1, x_1) and (m_2, x_2) where each bracketed pair is placed on either side of the relationship type, is defined to be an ER-construct-unit. The unit is named as the Binary-relationship-type-ER-construct-unit and denoted as $b(\, r_v(e_i, e_j))$. The unit is shown partitioned and named as $b(\, r_v(e_i, e_j))$ in the ER model in Fig. 5. The letter "b" indicates the meaning "base".

Notice that depending on the actual numerical values of the min-max variables, the relationship type may get either of the forms: one-to-one, one-to-many, or many-to-many. However, the constitution and the shape of the ER-construct-unit are not to be changed for any form of the relationship type: one-to-one, one-to-many, or many–to-many.

In the generic ER model (Fig. 5), the partition named - $p(\, r_v(e_i, e_j))$ shows the generic equivalent of the Simple-optional-attribute-ER-construct-unit, the one we have shown in the partition (C) in the real-world ER model (Fig. 4). Accordingly, we formally define the ER-construct-unit as follows.

B. Definition 4

In a generic ER model, the collection of the simple attributes attached to a Binary-relationship-type-ER-construct-unit, $b(\, r_v(e_i, e_j))$, is defined to be an ER construct unit. The unit is named as the Simple-optional-attribute-ER-construct-unit attached to Binary-relationship-type-ER-construct-unit, $b(\, r_v(e_i, e_j))$. The unit is partitioned and denoted as $p(\, r_v(e_i, e_j))$ in the ER model in Fig. 5. The letter "p" represents the meaning "optional". The unit is an optional unit, that is, it may or may not exist attached to a unit, $b(\, r_v(e_i, e_j))$. If it exists, its number of attributes may vary.

Accordingly, Fig. 5 shows how a Binary- relationship type, $r_v(e_i, e_j))$, in a generic ER model can be partitioned into two ER-construct-units, named, $b(\, r_v(e_i, e_j))$, and $p(\, r_v(e_i, e_j))$. It also shows how the two units: $b(\, r_v(e_i, e_j))$ and $p(\, r_v(e_i, e_j))$ can exist associated with each other and form the relationship type, $r_v(e_i, e_j))$ in a generic ER model.

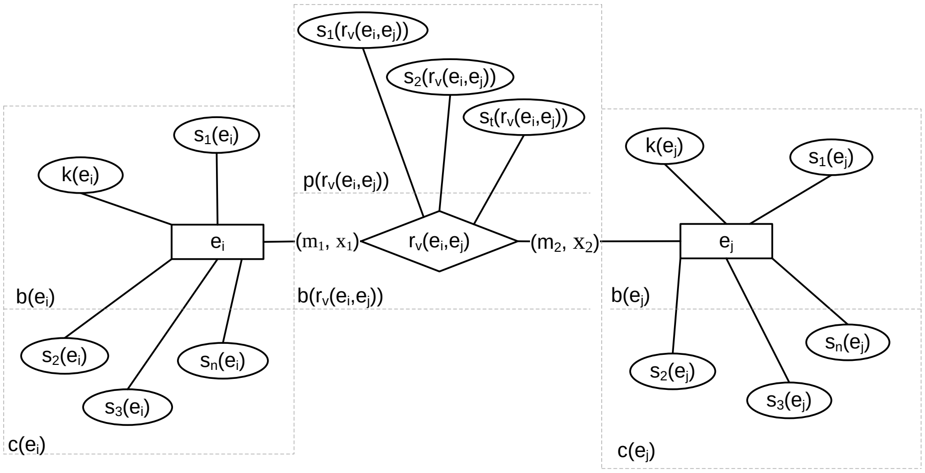


Fig. 5. A Generic ER Model Containing a Binary One-to-Many Relationship Type Attached to Two Regular Entity Types.

VI. Conclusion

We have shown (Section III) that the regular entity type, e_i, in the ER model (Fig. 3) can be partitioned into two distinct ER construct units, $b(e_i)$ and $c(e_i)$. The same partitions and the ER construct units: $b(e_i)$ and $c(e_i)$ exist in the generic ER model in Fig. 5. Similarly, the regular entity type, e_j, in the generic ER model (Fig. 5) can also be partitioned into two ER-construct-units, $b(e_j)$ and $c(e_j)$. We also showed that the binary-relationship type, $r_v(e_i, e_j))$, in the generic ER model (Fig. 5) can be partitioned into two ER-construct-units, $b(\ r_v(e_i, e_j))$, and $p(\ r_v(e_i, e_j))$.

Accordingly, the entire generic ER model in Fig. 5 can be partitioned into six ER-construct-units, namely, $b(e_i)$, $c(e_i)$, $b(r_v(e_i, e_j))$, $p(r_v(e_i, e_j))$, $b(e_j)$, and $c(e_j)$. The six partitions are distinct: that is, any one of them does not overlap or penetrate into another. They all together cover the entire generic ER model (Fig. 5).

The six distinct ER-construct-units form a set: $\{\, b(e_i)$, $c(e_i)$, $b(r_v(e_i, e_j))$, $p\left(r_v(e_i, e_j)\right)$, $b(e_j)$, $c(e_j)\}$. We assume that the set can be used to represent the generic ER model (Fig. 5).

On the other hand, a generic ER model can represent any real-world ER model [13]. Thus, we conclude that any real-world ER model that contains a binary relationship type that exists between two regular entity types can be viewed as a set of six elements and the set can be used as a representation of the ER model.

VII. Future Research Imerging from the Current Research

The current paper presents a part of an ongoing reach. Its results will be used in the future for further research expected. Accordingly, in future research, we will transform the moderate level generic ER model (Fig. 5) to a relational database schema (RDS). We will use the modified transformation algorithm for this purpose. We will then partition the RDS into segments, which we call Relation-schema-units. Next, we show that a mapping that is one-to-one and onto exists from the set representing the generic ER model to the set representing its RDS. We will then show that the information represented on the ER model is preserved on the RDS.

VIII. Implications of the Research Series

We argued that a one-to-one and onto correspondence from the ER model to the RDS not only preserve information from the ER model to the RDS. It also should be a basis for automating the transformation process from the ER model to the relational model. In section 1, we stated that a CASE tool can be created for automating the process.

We believe the CASE tool we expect can extend the work of Khaire and Mali [22]. The tool can be integrated with the web application that they have proposed. The CASE tool can then be used to automatically transform an ER model produced by the web application to the relational model. The CASE tool we expect should be able to be integrated with any other CASE tool that creates ER models (e.g. ERDplus - https://erdplus.com/) to transform them to the RDS automatically. Further, a CASE tool we expect also can extend the works of [23], and [24] (Section 1), in the same manner, mentioned above.

Going beyond the visualization of a computer-aided transformation process proposed by Yang and Cao [25], the CASE tool we expect could undertake the entire transformations process and perform it purely automatically without letting a user be intervened at intermediate stages for making adjustments. Even if the traditional ER model is claimed to be more suitable for teaching ER modeling concepts [26], in our view, the database designing process cannot be limited to just ER modeling only. Once an ER model diagram is created, it needs to be transformed to the RDS. The created RDS should be accurate and a one that preserves the information of its predecessor ER model. Without obtaining the skill that how an ER model can be transformed to the RDS, accurately and with preserving information, the database design and learning process is deemed to be incompleted. We argue that our modified approach comprising the ER model and the transformation algorithm that we have modified can fill this gap. It provides a hassle-free learning process. The reason the ER modeling and transformation rules are now apparent, straightforward, and ambiguous free. They provide a one-to-one transformation from the ER model to the RDS, which will also automate the transformation process. An automated tool can help students to validate their manual transformations and iteratively improve them until a correct RDS is reached as the output. The same advantage is equally applicable to the practitioners as they no longer need worrying about how models can be transformed from one to the other from the ER model to the RDS. A CASE tool will do the job for them.

Except for our ongoing researches for formal validation of our approach, empirical researches can be undertaken with learners, teachers, and practitioners aiming to assess our claims about the impact of the approach on improving the efficiency and productivity of them. If a CASE tool is produced, it can also be used as a tool for empirical validation of the approach.

References

[1] P. P. S. Chen, "The entity-relationship model: toward a unified view of data," ACM Trans. Database Syst, vol. 1, pp. 9-36, 1976.

[2] R. Elmasri and S. B. Navathe, Database Systems: Models, Languages, Design and Application Programming, 6 ed. Chennai: Pearson, 2013.

[3] E. F. Codd, "A relational model of data for large shared data banks," Commun. ACM, vol. 13, pp. 377-387, 1970.

[4] K. Kumar and S. K. Azad, "Relational Database Design: A Review," International Journal of Computer Applications, vol. 176, pp. 14-18, 2017.

[5] D. Pieris, "Reversible Database Design From the Entity-Relationship(ER) model" unpublished manuscript, 2015.

[6] R. C. Goldstein and V. C. Storey, "Data abstractions: Why and how?," Data & Knowledge Engineering, vol. 29, pp. 293-311, 1999.

[7] A.-J. Harith, D. Cuadra, and P. Martínez, "PANDORA CASE TOOL: Generating Triggers For Cardinality Constraints In RDBMS," 2003.

[8] C. Fahrner and G. Vossen, "A survey of database design transformations based on the Entity-Relationship model," Data & Knowledge Engineering, vol. 15, pp. 213-250, 1995.

[9] C. Batini, S. Ceri, and S. B. Navathe, Conceptual database design: an Entity-relationship approach: Benjamin-Cummings Publishing Co., Inc. Redwood City, CA, USA ©1992, 1992.

[10] A.-J. Harith, D. Cuadra, and P. Martínez, "Applying a Fuzzy approach to relaxing cardinality constraints, Database and Expert Systems Applications," in Database and Expert Systems Applications. vol. 3180, ed, 2004, pp. 654-662.

[11] D. Cuadra, P. Martínez, E. Castro, and A.-J. Harith, "Guidelines for representing complex cardinality constraints in binary and ternary relationships," Software and Systems Modeling, pp. 1-19, 2012.

[12] J.-L. Hainaut, "The transformational approach to database engineering," in Generative and Transformational Techniques in Software Engineering, ed: Springer, 2006, pp. 95-143.

[13] D. Pieris, M. C. Wijegunasekera, and N. G. J. Dias, "An Improved Generic ER Schema for Conceptual Modeling of Information Systems," presented at the Asia International Conference on Multidisciplinary Research 2019 (AIMR’19), Colombo, Sri Lanka, 2019.

[14] D. Pieris, "Modifying the entity relationship modeling notation: towards high quality relational databases from better notated ER models," arXiv preprint arXiv:1306.5690, 2013.

[15] D. Pieris, "A novel ER model to relational model transformation algorithm for semantically clear high quality database design," arXiv preprint arXiv:1306.6734, 2013.

[16] D. Pieris, M. C. Wijegunasekera, and N. G. J. Dias, "ER to Relational Model Mapping: Information Preserved Generalized Approach," presented at the 20th International Postgraduate Research Conference, University of Kelaniya, Sri Lanka, 2019.

[17] A. A. Almeida, A. C. Rocha-Oliveira, T. M. F. Ramos, F. L. de Moura, and M. Ayala-Rincón, "The Computational Relevance of Formal Logic Through Formal Proofs," in Formal Methods Teaching Workshop, 2019, pp. 81-96.

[18] Muhammad Ahsan Raza, S. R. M. Rahmah, A. Noraziah, and R. A. Hamid, "A Methodology for Engineering Domain Ontology using Entity Relationship Model," International Journal of Advanced Computer Science and Applications(IJACSA), vol. 10, pp. 326-332, 2019.

[19] Puja, P. Poscic, and D. Jaksic, "Overview and Comparison of Several relational Database Modelling Metodologies and Notations," in 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019, pp. 1641-1646.

[20] N. Amran, H. Mohamed, and F. D. S. Bahry, "Developing Human Resource Training Management (HRTM) Conceptual Model Using Entity Relationship Diagram (ERD)," International Journal of Academic Research in Business and Social Sciences, vol. 8, pp. 1444–1459, 2018.

[21] A. Kamišalić, M. Heričko, T. Welzer, and M. Turkanović, "Experimental study on the effectiveness of a teaching approach using barker or bachman notation for conceptual database design," Computer Science and Information Systems, vol. 15, pp. 421-448, 2018.

[22] A. V. Khaire and P. B. Mali, "Towards Automated Generation of ER-Diagram using a Web Based Approach," IOSR Journal of Computer Engineering vol. Volume 18, pp. 37-43, 2016.

[23] K. Kuk, M. Angeleski, and B. Popovic, "A Semi-automated generation of Entity-Relationship Diagram based on Morphosyntactic Tagging from the Requirements Written in a Serbian Natural Language," in 19th International Symposium on Computational Intelligence and Informatics, 2019, pp. 85-92.

[24] M. A. Javed and Y. A. Lin, "Iterative Process for Generating ER Diagram from Unrestricted Requirements," in 13th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2018), 2018, pp. 192-204.

[25] L. Yang and L. Cao, "The Effect of MySQL Workbench in Teaching Entity-Relationship Diagram (ERD) to Relational Schema Mapping," International Journal of Modern Education and Computer Science, vol. 8, 2016.

[26] P. Y. Wu, D. A. Igoche, and P. J. Drauss, "Media versus Message: Choosing The ER Diagram To Teach ER Modeling," in Proceedings of the EDSIG Conference ISSN, 2017, p. 3857.

Generic Framework Architecture for Verifying Embedded Components

Lamia ELJADIRI[1]
LIMSAD Laboratory in Mathematics and Computer Science
Department, Faculty of Sciences FSAC
University Hassan II, Casablanca, Morocco

Ismail ASSAYAD[2]
LIMSAD Laboratory in Mathematics and Computer Science
Department, ENSEM
University Hassan II, Casablanca, Morocco

***Abstract*—This dissertation presents a framework for the formal verification of standard embedded components such us bus protocol, microprocessor, memory blocks, various IP blocks, and a software component. It includes a model checking of embedded systems components. The algorithms are modeled on SystemC and transformed on Promela language (PROcess or PROtocol MEta LAnguage) with the integration of LTL (Linear Temporal Logic) properties extracting from state machines in order to reduce verification complexity. Thus, SysVerPml is not only dedicated to verifying generated properties but also for the automation integration of other properties in models if needed. In the following, we will provide the answer to the problems of component representation on the design system, what properties are appropriate for each component, and how to verify properties.**

Keywords—Algorithms; automation; embedded components; embedded systems; formal verification; framework; LTL properties; Promela; SystemC; SysVerPml; system design

I. Introduction

Verification can be applied to discover errors early in the SOC (System On Chip) design against properties expressed as part of the requirements. Worth to mention that the cost to find errors and to make correction in the product line increases ten times like what industry study demonstrates [1]; it is revealed that verification accounts for 55% in totality project time between 2012 and 2016.

The formal verification technology is divided into three methods: equivalence checking, model checking, and theorem proving [2], [3].

Equivalence checking is a technique based on mathematical approach to verify the equivalence of a reference or golden model to the implementation of the model [4].

Model checking is an algorithmic technique for determining whether a system satisfies a formal specification expressed as a temporal logic formula, where properties are the direct representation of a design's behavior [5].

Finally, the theorem proving method has the ability to decompose a problem especially the case of microprocessor verification. More details on theorem proving can be found in [6].

The three formal methods are generally used as formal verification techniques. However, model checking is particularly used in protocol verification. Model checking method [7] treats all the possible behavior of the design model.

The method called Undounded Model Checking (UMC) is based on the translation of the model checking problem into the satisfaction problem of a propositional formula, unlike the Bounded Model Checking (BMC) the encoding of the formulas is different. While the two techniques shares the encoding of the states and the transition relation of the model as explained on the article [8].

This article gives an overview of our Model Checker Platform named SysVerPml; the tool allows creating an abstract model of the design instead of translating SystemC programs to formal models, and then checking them using verification tool SPIN (Simple Promela INterpreter).

Model checking focuses on the state-space explosion problem. The main idea in our approach is that the number of states of a design is exponential to the number of variables and the width of each variable. To attain this first aim as explained in our previous article [9] the modeling methodology of a system must exhibit the execution semantics instead of encompassing it inside an execution-scheduler. Moreover, in order to allow new and old systems integration, any process interaction which might be useful for inter-system integration must not be cut in the final system model. The challenge of this approach is to guarantee that the abstract model is exact to the granularity of programs behaviors states. For that we use the code-level way of verification, as explained in the article [10], which has the advantage of permitting compositional verification of programs by keeping their incomplete interactions.

In the following, we first state the verification environment supported by our approach describing the different plug-in component used by the framework. Second, study case is taken as an example for the proposed method with the complete transformation procedure for the SRAM component. We conclude the resume with tests of the performance verification of our framework followed by conclusion.

II. Verification Environment

By the collaboration and exploitation of core integration technology, we can focus on core competencies to invent development technology as our platform SysVerPml will allow us. The SysVerPml tools have been developed over the Eclipse development environment. The open source integrated development environment (IDE) Eclipse developed by IBM, Object Technology International (OTI), and eight other companies [11], [12[, [13], [14]. This IDE mainly allows providing an extensible platform for building software. As shown in Fig. 1 the major advantage is that it gives extensible facilities which makes possible to implement tools of our framework by plug–ins such us the use of SystemC plug-in, IPXACT plug-in, and JSpin Java GUI for SPIN (graphical user interface for the SPIN Model Checker).

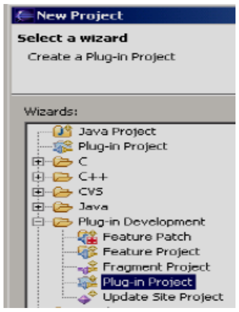


Fig. 1. Adding Plug-in Project from the Plug-in Development and Accepting Content Default Settings.

Nowadays, SystemC is an embedded system modeling language that has a lot of features and can be used to develop prototypes of embedded system. It is rich by its data types library and compilation environments of the C ++ language. It adds primitives to be able to write parallel processes, signals, clocks, as well as some concepts of a component language. SysVerPml has been designed to support SystemC plug-in. SystemC plug-in has been utilized to create a SystemC project based on C/C++ Development Toolkit (CDT) plug-ins in Eclipse Platform, SystemC, Cygwin packages required for building GCC compiler, and Managed Build System (MBS) pre-defines many useful macros and allows tool integrators and users to define additional macros. The CDT plug-ins supports a C/C++ Editor, Debugger, Launcher, Parser, Search Engine, Content Assist Provider and a Makefile generator [15]. To do the installation we followed the steps described at the guide for getting started with SystemC development, it contains a chapter for setup of Eclipse together with Cygwin and SystemC [16].

IP-XACT is another standard enabling the assembly of IP components (Intellectual Property blocks); it describes especially the interconnection interfaces, some communication components and associated protocols, using an Architecture Description Language (ADL). The ADL makes possible to define the interfaces of certain types of bus and protocols. The IP-XACT format respects the syntax construction rules specified in XML's Abstract Syntax Tree (AST). IP-XACT has been designed to address all these issues by providing a standardized data exchange format which has both the flexibility to represent SystemC models and the rigor to allow information to be automatically extracted and used in flow automation and advanced verification by Spin using Promela language.

In order to realize the transformation between SystemC and IP-XACT, we use Eclipse IP-XACT plug-in [17] as a means to import the IP component descriptions from the first model ScModel which provides database along with methods and structural information such as variables, functions, events, ports, processes, constructors and module instances, and from the second model PtrModel which include assertions with re-usable properties and the system declaration. So we realize the stream described in our previous article [10] and we pass the structural model conform to the SystemC behavioral model as a call parameter to retrieve a complete model as main file output. In this file, we create an instance of the embedded component with their attributes and the parameter configurations. Component properties are established by port-signal bindings.

IP-XACT is successful at ensuring syntactic formats compatibility and the interpretation's uniqueness of their descriptions to make component interoperability if needed, but it is not simulatable and it has neglected the behavioral aspects and components properties verification. Further, the purpose of this SystemC main file is to enable a simulation for the IP-XACT model. In a previous work, we described [9] that our translation to SystemC can also be seen as a translation into a set of automata. Each process and each function is translated into one produced automaton by composing produced SystemC models without any change. The SysVerPml framework enables to check safety properties for each SystemC program of the product line once at design time, without the need for additional time to redo the verification process every time programs are involved in the creation of new system prototypes as explained in our work [18]. After the simulation and gathering of results, a Promela file is generated. In this file, specifications can be given in Linear Temporal Logic (LTL) formulas;

The plugin consists of two main components, a compiler which compiles Promela code, and an interpreter. We used the graphical front-end JSPIN. The JSPIN tool executes SPIN commands in the background in response to user actions. It provides a clear overview of the many options in SPIN that are available for performing animations and verifications. JSPIN was built using the Java SWING library and consists of three adjustable panes, displaying text. The left one displays the Promela source files, the lower one messages from SPIN and JSPIN and the right one is used to display the output of printf statements and of data from animations [19].

As we shall see in the article JSPIN tool will attempt to do automatically verification limiting human intervention and returning one of three results; whether it be a state where properties are satisfied, or properties are not satisfied so a counterexample will be given, or Indeterminate if the state space is such that the tool cannot compute a result in a reasonable amount of time [20].

In order to demonstrate the importance of the SysVerPml framework the case studies of some embedded components have been published in preceding articles; the verification results of FIFO component have been published in [21] and the verification results of Bus AMBA AHB have been published in [10]. In this dissertation we provide an application example related to memory SRAM (static RAM), this component have two views following the model described in Fig. 2.

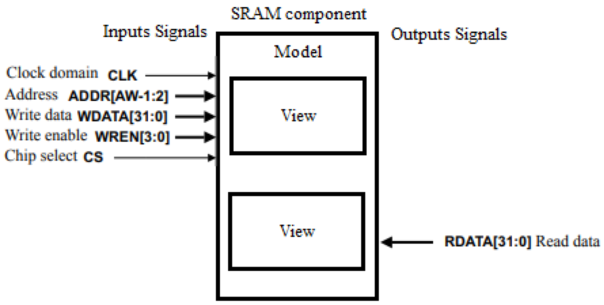


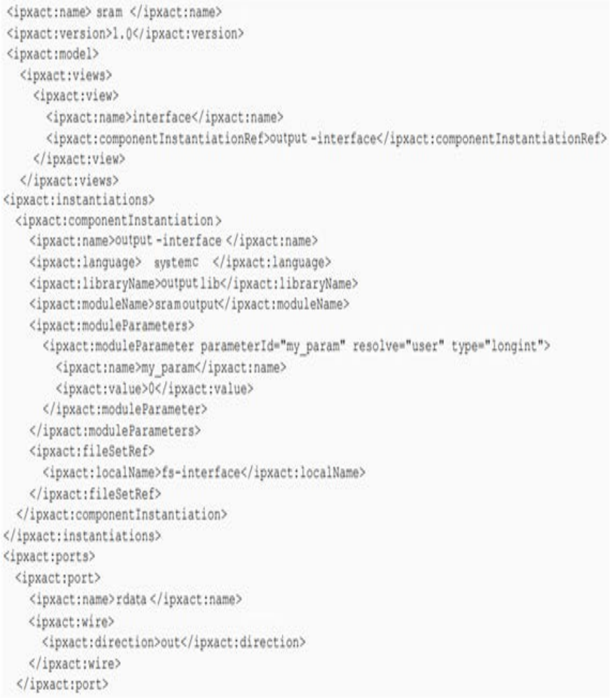
Fig. 2. The Characteristics of the SRAM Model.

We report the IP-XACT description introduced in the previous part of this article. We use the namespace ipxact, below in Fig. 3 we show the output view of the SRAM model, in which port is denoted with RDATA.

The component definition <ipxact:component> contains information to Promela file and the SystemC model about the component. This information is situated within a <ipxact:parameter> element, identified with the <ipxact:value> and <ipxact:name> tags.

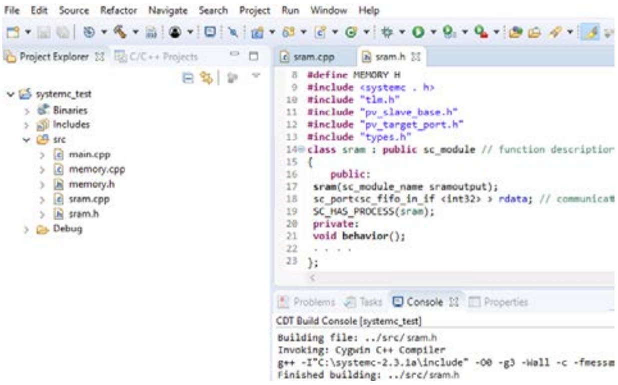
Each component interface that uses SysVerPml mapping is defined in the generated file as: Inputs, outputs, the combination of inputs and outputs and the parameters.

We can combine inputs and outputs in a single component interface definition, but we havent't possibility to combine parameters and inputs/Outputs because these elements are defined in the pair name-value of <ipxact:parameter> which indicates to the SysVerPml generator that there is a SystemC mapping.



```
 <ipxact:name> sram </ipxact:name>
 <ipxact:version>1.0</ipxact:version>
 <ipxact:model>
   <ipxact:views>
     <ipxact:view>
       <ipxact:name>interface</ipxact:name>
       <ipxact:componentInstantiationRef>output -interface</ipxact:componentInstantiationRef>
     </ipxact:view>
   </ipxact:views>
<ipxact:instantiations>
  <ipxact:componentInstantiation >
    <ipxact:name>output -interface </ipxact:name>
    <ipxact:language>  systemc  </ipxact:language>
    <ipxact:libraryName>output lib</ipxact:libraryName>
    <ipxact:moduleName>sram output</ipxact:moduleName>
    <ipxact:moduleParameters>
      <ipxact:moduleParameter parameterId="my_param" resolve="user" type="longint">
        <ipxact:name>my_param</ipxact:name>
        <ipxact:value>0</ipxact:value>
      </ipxact:moduleParameter>
    </ipxact:moduleParameters>
    <ipxact:fileSetRef>
      <ipxact:localName>fs-interface</ipxact:localName>
    </ipxact:fileSetRef>
  </ipxact:componentInstantiation>
</ipxact:instantiations>
<ipxact:ports>
  <ipxact:port>
    <ipxact:name>rdata </ipxact:name>
    <ipxact:wire>
      <ipxact:direction>out</ipxact:direction>
    </ipxact:wire>
  </ipxact:port>
```

Fig. 3. The IPXACT Document Tree.



```
#define MEMORY H
#include <systemc . h>
#include "tlm.h"
#include "pv_slave_base.h"
#include "pv_target_port.h"
#include "types.h"
class sram : public sc_module // function descriptior
{
    public:
 sram(sc_module_name sramoutput);
 sc_port<sc_fifo_in_if <int32> > rdata; // communicat
 SC_HAS_PROCESS(sram);
 private:
 void behavior();
 . . . .
};
```

CDT Build Console [systemc_test]

```
Building file: ../src/sram.h
Invoking: Cygwin C++ Compiler
g++ -I"C:\systemc-2.3.1a\include" -O0 -g3 -Wall -c -fmessa
Finished building: ../src/sram.h
```

Fig. 4. The SystemC Interface Capture.

We have the possibility to import the generated IP-XACT file for use and update if needed by the use of the SysVerPml generator; the header file describes the design of Memory and it is entirely integrated into the SystemC model illustrating how information contained in IPXACT file can be used for a behavioral implementation as we observe in Fig. 4.

Furthermore, we can use easily the interface of JSPIN tool; downloaded from the Github link [22]; a tool that track bugs into the encoding programs and it can verify whether a specification is satisfied or make a counterexample of symbolic formulas. By the way, it makes possible to edit as well as to update the LTL formulas written inside of Promela model in respect of semantic transformation from SystemC model. JSPIN tool is an elementary part of our SysVerPml platform and it makes possible to run simulation and formal verification directly. We note well that JSPIN's main focus is the SpinSpider component. SpinSpider allows us to demonstrate the properties in case of concurrent processes.

The generated file in the PtrModel module is represented by the structured classes. These classes gave us the advantage to efficiently represent the semantic results and allow us to represent both the ports and the properties of the component.

III. Case Study

This section discusses the use of our approach to verify some properties of the SRAM design used in interaction with a CPU model which contains working microengines - a set of threads in each microengine – and all of them want access to SRAM component.

We have developed the translator, which takes the SystemC design as an input and generates the Promela encoding with the integration of properties as explained in the previous section. The translator uses IPXACT to extract from the SystemC design description that is useful for performing the transformation to Promela language.

Remember that the verification of the resulting Promela models from the SystemC models provided by JSPIN tool to completely verifying SRAM component.

To make length of this paper brief we express with LTL the most functional properties, such as non-starvation, safety and deadlock.

The non-starvation property for the events that are related to SRAM controller of a CPU means that if an SRAM access request comes from a thread 0 of a microengine 0 for example is enqueued, it is eventually committed in the next 400 SRAM occurrences. This property can be formulized with the LTL formula as shown in (1) in this way:

AG (microengine0_thread0_sram_enqueued ⇒XF [1:400]
(microengine 0_thread 0_sram_done)) (1)

The safety property of the memory access is stored in a scheduling FIFO to handle the occurred order of the events sram_enqueued (the SRAM access request is enqueued), sram_dequeued (the SRAM access request is dequeued) and sram_done (the SRAM access request is committed), which makes necessary that always after an SRAM request by a thread 1 of a microengine 1 for example, it cannot be done before it is dequeued. As shown in (2) this property can be expressed with the LTL formula like this:

AG (microengine1_thread1_sram_enqueued ¬
microengine1_thread1_sram_done U
microengine1_thread1_sram_ dequeued) (2)

The deadlock property to prevent problems with shared resource, for each SRAM access on CPU, the data readout and the memory address referenced must be similar, and always all the SRAM references represented by addr are made in execution with the same order. As shown in (3) the LTL formula can be expressed with the following:

AG (addr(sram_enqueued[i]) = addr(sram_enqueued_CPU[i])
^ data(sram_done[i]) = data(sram_done_CPU[i])) (3)

We assume that the SRAM access request is put into a scheduling FIFO by a thread 1 of a microengine 1 for example and then eventually committed; always the memory address should be the same as shown in (4).

AG (addr(microengine1_thread1_sram_enqueued[i]) =
addr(microengine 1_thread1_sram_done[i])) (4)

Table I lists the average values of performance metrics using by SPIN verification process. The average values were computed over the set of pre-defined specification properties to check without errors the functional properties of SRAM component.

TABLE I. VERIFICATION DATA

LTL formulas	**SPIN Metrics**			
	States generated	***Transitions number***	***Memory used***	***Verification time***
1	6700	$3.0*10^5$	1KB	100s
2	5739	$7.0*10^6$	50Bytes	24s
3	10267	$3*10^5$	40KB	6s
4	5710	$7.0*10^6$	12Bytes	60s

The units used in this table are; B= Bytes, s = second.

IV. CONCLUSION

In this paper, we have reported our effort to implement SysVerPml platform and the impressive component's modeling and checking gain obtained by transforming SystemC models to Promela encodings. This remarkable gain is achieved by modules which decomposes the implementation of our tool and make it modular. This modularity facilitates modifications inside of IPXACT description and LTL properties. We have provided an application example related to a sessions that implements the SRAM component.

REFERENCES

[1] H. D. Foster, "Trends in Functional Verification: A (2016) Industry Study", whitepaper, Mentor Graphics.

[2] Edmund M. Clarke and Jeannette M. Wing. "Formal Methods: State of the Art and Future". In ACM Computing Survey, volume 28–4, pages 626–643, (December 1996).

[3] Carl-Johan Seger. "An Introduction to Formal Verification". Technical Report 92–1, Department of Computer Science, University of British Columbia, Canada, (June 1992).

[4] D. D. Gajski, S. Abdi, A. Gerstlauer, and G. Schirner. "Embedded System Design: Modeling, Synthesis and Verification". Springer, (2009).

[5] E. M. Clarke and E. A. Emerson. "Design and Synthesis of Synchronization Skeletons Using Branching-Time Temporal Logic". In Logic of Programs, Workshop, pages 52–71, London, UK, (1981). Springer-Verlag.

[6] M. J. C. Gordon and T. F. Melham, "Introduction to HOL: A Theorem Proving Environment for Higher Order Logic", Cambridge University Press, (1993).

[7] Orna Lichtenstein and Amir Pnueli. "Checking that finite state concurrent programs satisfy their linear specification". In Proceedings of the 12th ACM SIGACT-SIGPLAN POPL'85, pages 97–107, New York, NY, USA, (1985). ACM.

[8] Nina Amla, Robert Kurshan, Kenneth L. McMillan, and Ricardo Medel, "Experimental Analysis of Different Techniques for Bounded Model Checking", Springer-Verlag Berlin Heidelberg (2003).

[9] A. Ismail, E. J. Lamia, Z. Abdelouahed, and N. Tarik, "The behavior, interaction and priority framework applied to systemc-based embedded systems", in 13th IEEE/ACS International Conference of Computer Systems and Applications, AICCSA 2016, Agadir, Morocco, (November 29- December 2, 2016).

[10] Ismail Assayad, Lamia Eljadiri, "A platform for systematic verification of embedded components in IP-XACT, SystemC and Promela", In ICSDE 2018, Rabat, Morocco, (October 18-19, 2018).

[11] "Eclipse Platform Technical Overview". Technical Report, Object Technology International (OTI) Inc… (2001). http://www.eclipse.org /whitepapers/eclipseoverview.pdf

[12] Holzner Steve, "Eclipse", O'Reilly, (April 2004).

[13] Holzner Steve, "Eclipse Cookbook", O'Reilly , (June 2004).

[14] Shavor Sherry, D'Anjou Jim, Fairbrother Scott, Kehn Dan, Kellerman John, McCarthy Pat, "The Java Developer's Guide to Eclipse",Addison-Wesley, (2003).

[15] "Managed Build Extensibility Reference Document (for CDT2.1)", http://www.eclipse.org/cdt/

[16] "Guide for getting started with SystemC development", by senior consultant Kim Bjerge, Danish technological institute (2007).

[17] http://www.eclipse.org/dsdp/dd/ipxact/gettingstarted/QuickStart.html

[18] Lamia Eljadiri, Ismail Assayad, and Abdelouahed Zakari, "Generic Verification of Safety Properties For SystemC Programs Using Incomplete Interactions", In ICSDE 2018, Rabat, Morocco, (October 18-19, 2018).

[19] BEN-ARI, Mordechai, "Principles of the Spin Model Checker". Springer, (2008). – ISBN 978–1–84628–769–5

[20] Robert C. Armstrong, Ratish J. Punnoose, Matthew H. Wong, Jackson R. Mayo, “Survey of Existing Tools for Formal Verification”, Sandia National Laboratories, Albuquerque, New Mexico 87185 and Livermore, California 94550, (December 2014).

[21] Ismail Assayad, Lamia Eljadiri, Abdelouahed Zakari, “Systematic Verification of Embedded Components with Reusable Properties”. In WINCOM 2017, Rabat, Morocco, (November 01-04, 2017).

[22] https://github.com/motib.

Architectural Proposal for a Syllabus Management System using the ISO/IEC/IEEE 42010

Anthony Meza-Luque[1]
Bantotal Perú S.A.C
Arequipa, Perú

Alvaro Fernández Del Carpio[2], Karina Rosas Paredes[3]
Jose Sulla-Torres[4]
Systems Engineering Program
Universidad Católica de Santa María
Arequipa, Perú

***Abstract*—The efficiency in the academic and administrative procedures of higher education clearly marks competitive advantage in aspects of quality, which consists in the continuous improvement and improvement for the achievement of educational objectives. In our institution, the syllabus treatment is currently carried out manually, delaying many educational processes. Therefore, it is proposed to innovate through a software architecture approach based on the standard "ISO/IEC/ IEEE 42010: Systems and software engineering - Architecture Description" to describe the architecture Syllabus Management System software. It is developed in three stages: Analysis, Design and Verification. This will allow professors to develop their research, training, teaching and presentation of timely reports, with the measurement of achieved skills and abilities of students, managers and academic authorities, and to make decisions based on the results obtained by the tool allowing an improvement in the quality of the contents and development flow of the syllabus.**

Keywords—Management; architecture; software; syllabus; skills; ISO/IEC IEEE 42010

I. Introduction

The syllabus is a plan for teaching and learning, as it contains the important meaning between the professor and students. However, most current program management systems offer simple functionality including creating, modifying, and retrieving the unstructured program [1].

The management of syllabus in a quality educational environment within a teaching-learning process has become an urgent need for universities [2]. The "Universidad Católica de Santa María" (UCSM) has obtained the ISO 9001 standard for its quality management systems (QMS) in order to meet the needs of customers and stakeholders. Thus, it is very appropriate to implement good practices specifically in the continuous improvement for the achievement of educational objectives.

The properly chosen software architecture helps to overcome potential problems and allows you to take advantage of this model [3]. For the software architect, it is essential to understand what a software architecture ready application is and what requirements it must meet.

In this context, for the academic management of the educational content of the subjects taught at the university, they have been innovating through a software architecture approach, improving the accessibility and usability of software products through standards for minimize errors in the development of syllabus, an important instrument in the development of subjects.

Currently there are several proposals for software architecture, but as far as it has been reviewed, very few oriented towards syllabus management. In [4], they conducted a comparative study of the structure of the curriculum at Latin American universities, discovering that Syllabus are still seen as a registration document and not as a learning tool.

For this reason, the Syllabus Management System is proposed and developed in three stages. The analysis stage, with the detail of the functional and non-functional requirements, visits and interviews with the directly involved actors (professors and managers) to gather needs and current problems that lie mainly in response times since the process is carried out manually. The design stage, that starting from a prototype, automates the manual process, generating information on subjects from the Center for Academic Development to the different professional schools. This process minimizes common mistakes that the professor made, focusing only in sections relevant to its work. Finally, the verification stage, to validate the results of the previous stages. The architecture description can be modeled using the ISO/IEC/IEEE 42010 standard that allows for alternative options and decisions to be selected, where the rationale and trade-offs for each decision are documented and understood as necessary to inform subsequent decisions to stakeholders [5]. These concerns combined with the environment and system scenarios provide an architectural design context that clarifies the motivation to make decisions.

These functionalities and characteristics will allow professors to develop their research, training, teaching and presentation of timely reports. The measurements of skills and abilities achieved by students, managers and academic authorities are included in the proposal as well as to make decisions based on the results obtained. This tool helps to get an improvement in the quality of the contents and development flow of the syllabus through an intuitive and friendly system.

This article is organized as follows: the background is explained in Section II, the methodology in Section III, and the results of experimentation in Section IV, and finally the conclusions and future work in Section V.

II. Background

In universities, the syllabus document defines the contents of a course, and other important information that ensures the quality of teaching-learning. This document, in addition to capturing the contents of learning, includes the mechanisms for the achievement of learning and the development of skills and abilities. It is an instrument in which the contents are defined by units or themes, credits are specified, in some cases pre-requisite courses are established and the learning outcomes to be achieved are specified.

In the literature, various proposals have emerged for the construction of technological supports and mechanisms that help the good development of study programs. Such is the case of [6], which presents an abstract hierarchical model for syllabus management based on Model Directed Architecture, specifying a set of attributes such as levels, hours, credits, etc. It generates various types of reports, such as for accreditation and general purpose. The model is linked to the Open Syllabus system, a system for creating, editing and publishing syllabus.

In [7], an XML-based syllabus repository system integrates syllabus information from a set of universities was proposed. This system includes a template for data entry, the process of integration and search of syllabic contents. This system makes it easier for students to search for courses.

Similarly, [8] developed a service model for syllabus in order to standardize the method of classification of syllabus items and allow collaboration between environments. They modeled the syllabus scheme to facilitate the search between different domains, developed the automatic generation of the syllabus based on a markup language and implemented multilayer search agents. In addition, the model includes data creation, storage and retrieval functions.

Likewise, in order to assist in the search for topics of the courses in the study programs, [9] introduced a web-based tool using probabilistic models. This allows to identify the degree of similarity between the content of a syllabus in relation to a given topic. The tool extracts similar courses from a set of highly ranked universities. The benefit of the tool is twofold, as it helps students understand courses and professors to improve their study program.

In itself, World Wide Web technologies have transformed the design, development, implementation and deployment of decision support systems. The DSS web-based academic literature focuses primarily on applications and implementations, and only a few articles examine architectural issues or provide design guidelines based on empirical evidence [10]. For the development of technological proposals, as is the case of the present investigation, the software architecture approach is important, since it improves the accessibility and usability of software products for the benefit of high quality advanced education systems throughout the world [11]. It also allows improving learning experiences through collaborative services that are context-aware: software architecture and prototype system [12].

In [13], they analyzed the courses of archival study programs in North America. In identifying the convergences and divergences of the topics, they sought to understand the relationship between the two courses and obtain information on how these courses continue to serve as an integral component of archival studies education. The research examined three different aspects of the syllabus: textbooks, required articles and weekly topics. The syllabus was analyzed as separate data sets (RM syllabus and ERM syllabus), which was followed by a comparative analysis of the two types of syllabus. This may allow the Design of Knowledge Management Syllabus [14].

Júnior, Misra and Soares [15] indicate that software architecture as a development product is useful for technical activities, such as describing the views and concerns of the future software products, as well as for management activities, including assigning tasks to each team and as an input for project management activities. A major problem in describing software architecture is knowing what elements should be included in the architecture and at what level of detail, towards a Reference Architecture [16].

In [17] address the problem of how the software architectures of a System of Systems (SoS) should be described. For this purpose, they present an approach based on the standard "ISO / IEC / IEEE 42010: Systems and software engineering - Architecture description" to describe the software architectures of a SoS, as well as to develop an approach to modeling SoS using an architecture description language (ADL) [18].

As described, most jobs vary in the application of ISO/IECIEEE 42010. In addition, the approach used depends on the nature of course management at a University. Taking into account that most of these works are implemented for the design of general architecture, that is why in this work we try to use the suggestions of the standard in the university academic field. The objective is to use the guide of the ISO/IEC/IEEE 42010 standard to obtain a good software architecture in the management of syllabus in the university.

III. Elaboration: Architectural Proposal

Given the need for improvement within the Syllabus Management process in the UCSM and with the aspiration to have a standard and minimize errors in its preparation; the analysis, design and results points are detailed as follows:

A. Analysis

An agenda of meetings with interested parties was established to have a dialogue and understand the context in which syllabus management is currently taking place. These meetings allowed us to understand the process cycle from general to detail (see Fig. 1). The current guidelines established by the Center for Academic Development of which all professors use are also documented.

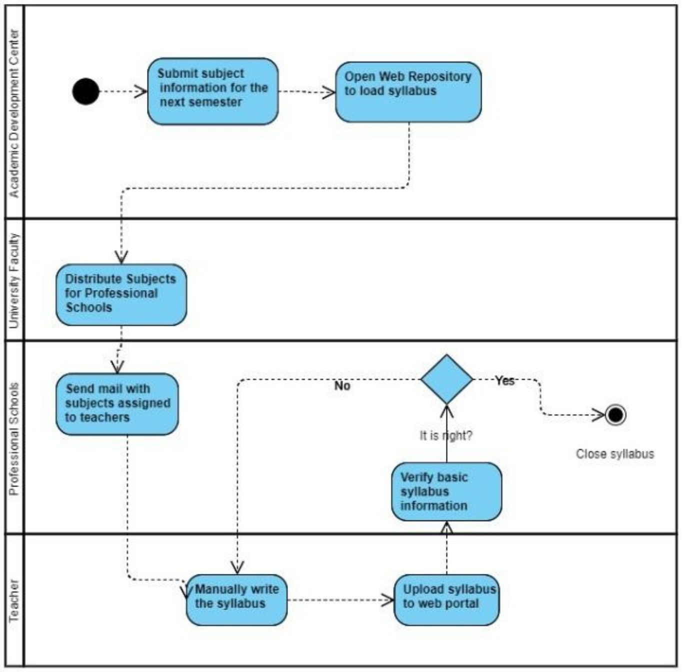

Academic Development Center
Submit subject information for the next semester
Open Web Repository to load syllabus
University Faculty
Distribute Subjects for Professional Schools
Professional Schools
Send mail with subjects assigned to teachers
No
Yes
It is right?
Close syllabus
Verify basic syllabus information
Teacher
Manually write the syllabus
Upload syllabus to web portal

Fig. 1. Current Syllabus Management Process.

With the information collected we proceed to detail the functional and non-functional requirements that will contribute to the improvement within the process of Syllabus Management.

Functional Requirements:

- Allow the creation, modification and elimination of users.
- Perform the creation, modification and elimination of roles.
- Navigate correctly in the menu corresponding to the assigned role.
- Consult the syllabus entered through a query interface.
- Be able to update the status of the syllabus to move on to its next stage.
- Execute the information load, transferring it from the database of the "Universidad Católica de Santa María" to the database for Syllabus Management.
- Develop the content of the syllabus focusing on the sections that require professor participation.
- Generate PDF file as final product after developing the syllabus correctly.

Nonfunctional Requirements:

- The information entered into the system will be protected from unauthorized access and disclosure.
- Ensure the integrity of the information that has been entered into the system. Redundant or unnecessary information will not be allowed.
- Ensure adequate access to users according to their assigned role. This will happen through users that professors already have assigned with Windows environment authentication.
- Facilitate the understanding of the system with the help of user manuals, which will allow professors to have a flow guide implemented in the system.
- Allow optimal supports to be made by documenting the design of the system that will help the maintenance user understand the system agilely.
- Implement a system incident log by means of error and audit logs, so that anomalies that have occurred may be monitored.

B. *Design*

A prototype of the Syllabus Management system is currently under development that allows automating and streamlining this process. For this, the process definition has been made and a new model to be executed is proposed (Fig. 2).

This new model automates two important points:

1) In the first place, there is automation of the issuance of information on the subjects by the Center for Academic Development towards the different professional Schools. After executing the SSIS packages, the predefined load of already established information is performed and the Syllabus Management database is fed for the process to be carried out.

2) Secondly, the professor only focuses on developing those sections within the syllabus that cannot be obtained with a predefined load. This minimizes common mistakes.

C. *External System Architecture*

In this section, we present an overview of the standard ISO/IEC/IEEE 42010 entitled "Architecture Description" [19] on which our approach is based.

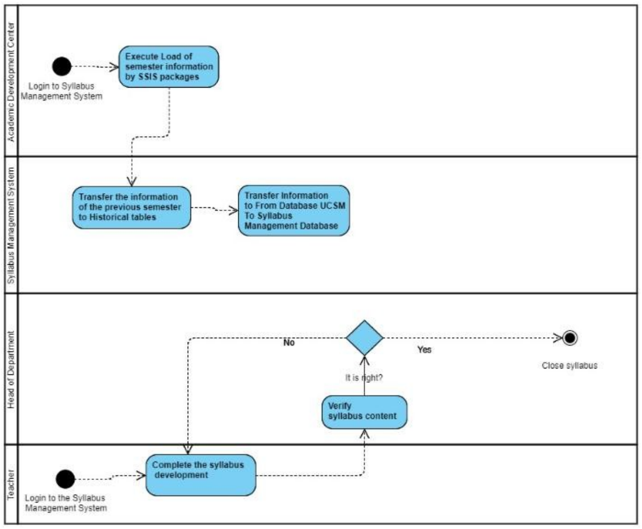

Academic Development Center
Login to Syllabus Management System
Execute Load of semester information by SSIS packages
Syllabus Management System
Transfer the information of the previous semester to Historical tables
Transfer Information to From Database UCSM To Syllabus Management Database
Head of Department
No
Yes
It is right?
Close syllabus
Verify syllabus content
Teacher
Login to the Syllabus Management System
Complete the syllabus development

Fig. 2. Proposal for the New Syllabus Management System.

In Fig. 3, a system is situated in an environment. It exhibits an architecture expressed by an architecture description. A system has a number of stakeholders. Each stakeholder has interest in the system presented by a number of concerns.

To obtain an architecture description, the first step is to identify the stakeholders having concerns considered fundamental to the architecture of the system and to identify these concerns. The second step is to identify the architecture viewpoint: providing a name for the viewpoint, providing a listing of architecture-relevant concerns to be framed by this architecture viewpoint, providing a listing of the typical stakeholders of a system and identifying each model kind used in the viewpoint. To identify a model kind, we must provide its name, describe the conventions for models and identify the notation used in models [17].

1) Concerns and stakeholders: We present the stakeholders of the Smart Building System:

- Professors (Professors) of the Professional Program are the users and operators of the system.
- The department heads of the faculty are the ones who review the system.
- A software development application represents courses and their syllabus, developers, builders, and system maintainers.

Then, we present, the concerns considered fundamental to the architecture of the system:

- The purposes of the syllabus management system are to guarantee the correct entry, validation and monitoring of syllabus for courses.
- To guarantee the suitability and feasibility of the architecture to achieve the purposes of the system, our objective is to implement this system based on SOA technologies.
- To intercept the risks and potential impacts of the system for its stakeholders throughout its life cycle and to ensure the maintenance capacity and evolution capacity of the system, we will use a monitoring mechanism that allows intercepting SOAP messages [20].

2) Viewpoint and model kind: For the Syllabus Management System illustrated in our case study, we propose a structural viewpoint. In fact, a structural viewpoint deals with the purposes of each system participating in our proposal, the suitability of the architecture for achieving the system's purposes and the feasibility of constructing and deploying the system.

The external architecture of the system will be based on web (see Fig. 4) and the interaction can be done through the internet or intranet of the university. Professors connect through Windows authentication with the users they were provided with. The information is validated and access is provided.

A relationship is generated between the UCSM database and the database used to manage the syllabus. The database feed to manage the syllabus is done through the execution of SSIS packages. The SSIS packages were developed in the Microsoft SQL Server Integration Services tool that provides the necessary platform to perform required data extraction, transformation and loading (ETL) from one Database to another.

The server that supports the web application is the same where the various solutions that the UCSM have put into production for their development are hosted.

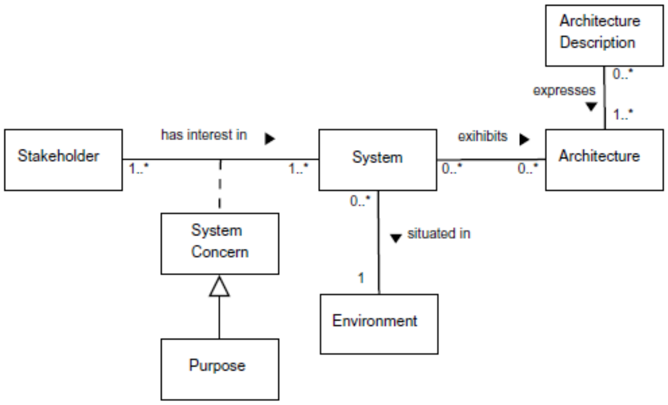


Fig. 3. Context of Architecture Description [19].

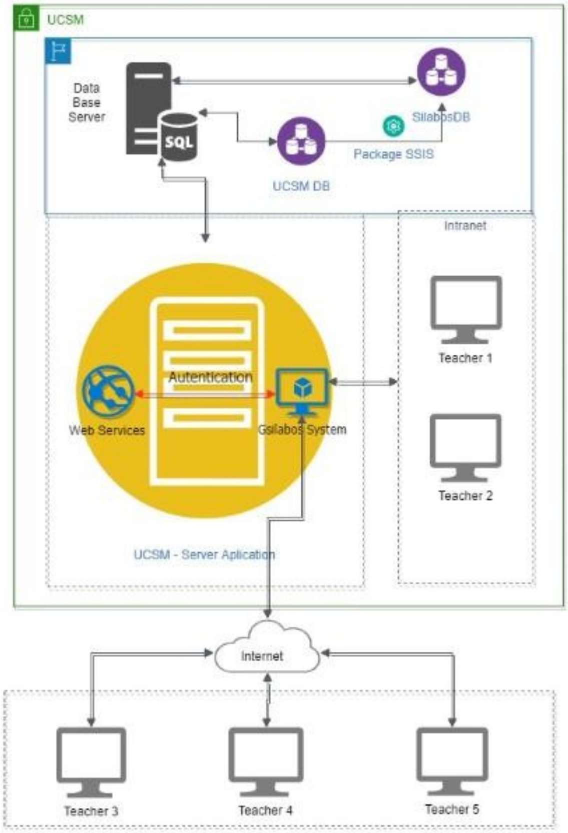


Fig. 4. External Architecture of the Syllabus Management System.

D. Internal System Architecture

The internal architecture of the system is based on MVC (see Fig. 5). Various technologies that enriched its functionality and aesthetics were integrated into this three-layer architecture.

For the view part, the Bootstrap 4 framework was integrated, which will help with typography, forms, buttons, boxes, navigation menus, etc. Also, the JQuery library was added to manipulate the DOM elements effectively. AJAX to be able to handle asynchronously the obtaining or sending of information without reloading the page. Data tables to show the information in the queries and that these are looked at in a distributed and dynamic way.

ReportViewer, a component that allows you to design reports, was used in this case to generate the final product that is the syllable in PDF format. Razor was used to embed C # code in views and add functionality. Log4net was applied to be able to store the record of errors or incidents for audit. Linq and Lambda were used to manage the information in the database.

Likewise, Codefirst was applied to build the database model, and SQL Server was included as the database engine where the system data will be hosted.

E. Main Screen Design

This section presents the designs of the main screens of the syllabus management system. Fig. 6 shows the interface of the syllabus structure. From here, we proceed to enter or modify the syllabus information.

Fig. 7 shows the specification of the contents and competences grouped by units as well as the matrix corresponding to the research activities according to the learning phases of the academic semester.

Fig. 8 presents a summary of the file upload status. In this section the SSIS packages are executed and their processing is verified; if it was successful or had an error.

Fig. 9 shows the generation of the report in pdf of the syllabus document, according to the format established by the UCSM.

F. Testing

The tasks for the implementation of the support tool for Syllabus Management were tested, and the following results were obtained (see Table II). The Priority Task is in Table I.

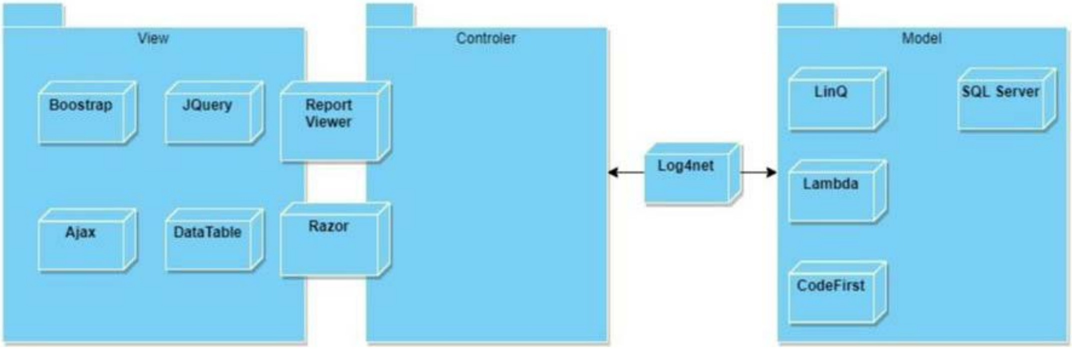


Fig. 5. Internal Architecture of the Syllabus Management System.

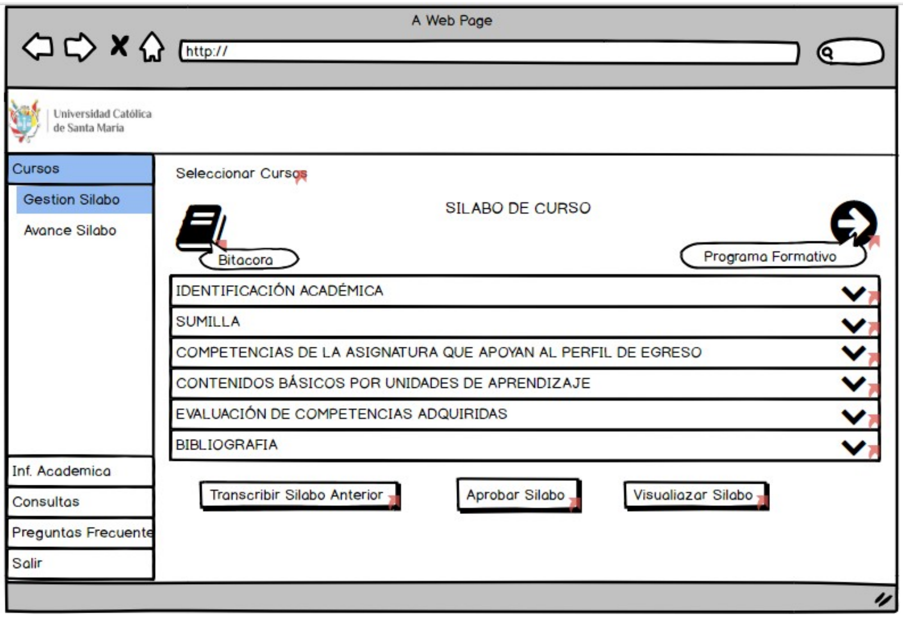
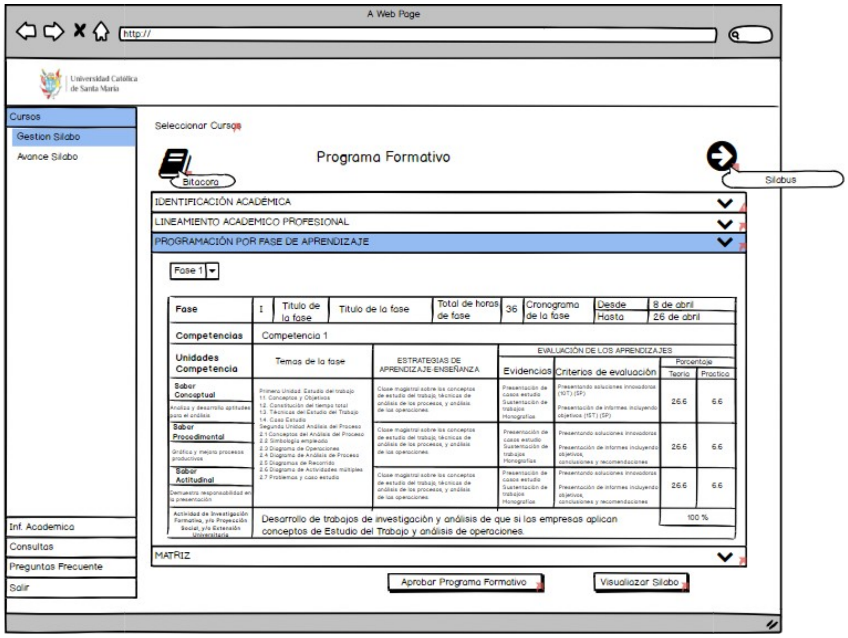


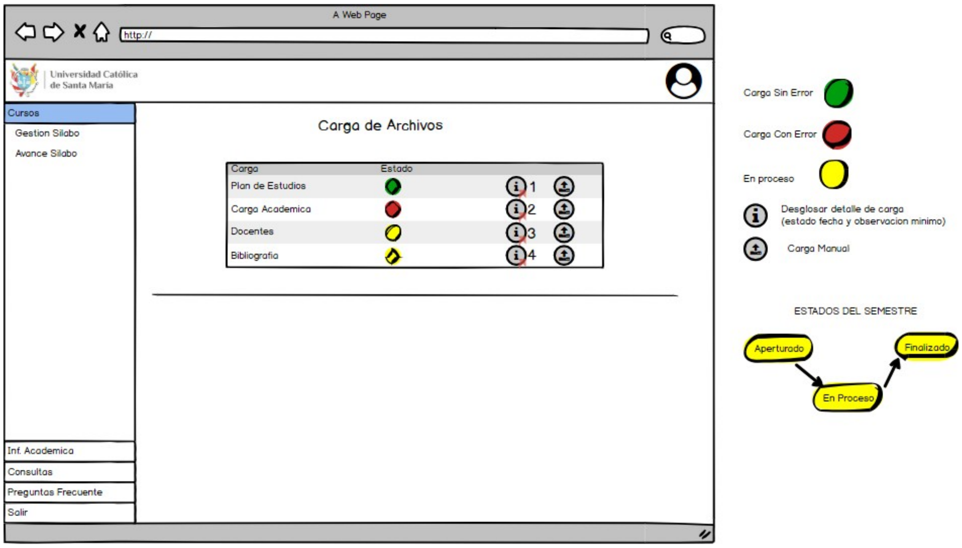
Fig. 6. Drop-Down Menu for the Syllabus Section.



Fase	I	Titulo de la fase	Titulo de la fase	Total de horas de fase	36	Cronograma de la fase	Desde Hasta	8 de abril 26 de abril
Competencias	Competencia 1							

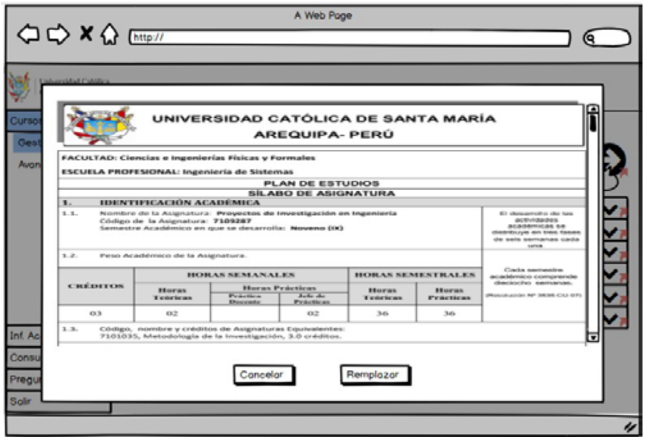
Unidades Competencia	Temas de la fase	ESTRATEGIAS DE APRENDIZAJE ENSEÑANZA	EVALUACIÓN DE LOS APRENDIZAJES			
			Evidencias	Criterios de evaluación	Porcentaje	
					Teoria	Practica
Saber Conceptual Analiza y desarrollo aptitudes para el análisis	Primera Unidad: Estudio del trabajo 1.1. Conceptos y Objetivos 1.2. Constitución del tiempo total 1.3. Técnicas del Estudio del Trabajo 1.4. Caso Estudio Segunda Unidad Análisis del Proceso 2.1 Conceptos del Análisis del Proceso 2.2 Simbologia empleado 2.3 Diagrama de Operaciones 2.4 Diagrama de Análisis de Proceso 2.5 Diagramas de Recorrido 2.6 Diagrama de Actividades múltiples 2.7 Problemas y caso estudio	Clase magistral sobre los conceptos de estudio del trabajo, técnicas de análisis de los procesos, y análisis de las operaciones.	Presentación de casos estudio Sustentación de trabajos Monografias	Presentando soluciones innovadoras (10T) (SP) Presentación de informes incluyendo objetivos (15T) (SP)	26.6	6.6
Saber Procedimental Gráfica y mejora procesos productivos		Clase magistral sobre los conceptos de estudio del trabajo, técnicas de análisis de los procesos, y análisis de las operaciones.	Presentación de casos estudio Sustentación de trabajos Monografias	Presentando soluciones innovadoras Presentación de informes incluyendo objetivos, conclusiones y recomendaciones	26.6	6.6
Saber Actitudinal Demuestra responsabilidad en la presentación		Clase magistral sobre los conceptos de estudio del trabajo, técnicas de análisis de los procesos, y análisis de las operaciones.	Presentación de casos estudio Sustentación de trabajos Monografias	Presentando soluciones innovadoras Presentación de informes incluyendo objetivos, conclusiones y recomendaciones	26.6	6.6
Actividad de Investigación Formativa, y/o Proyección Social, y/o Extensión Universitaria	Desarrollo de trabajos de investigación y análisis de que si las empresas aplican conceptos de Estudio del Trabajo y análisis de operaciones.				100 %	

Fig. 7. Drop-Down Menu of the Training Program Section.



Carga	Estado		
Plan de Estudios	🟢	ⓘ 1	⊕
Carga Academica	🔴	ⓘ 2	⊕
Docentes	🟡	ⓘ 3	⊕
Bibliografia	🟡	ⓘ 4	⊕

Fig. 8. Load Menu.



UNIVERSIDAD CATÓLICA DE SANTA MARÍA
AREQUIPA- PERÚ

FACULTAD: Ciencias e Ingenierías Físicas y Formales

ESCUELA PROFESIONAL: Ingeniería de Sistemas

PLAN DE ESTUDIOS

SÍLABO DE ASIGNATURA

1. IDENTIFICACIÓN ACADÉMICA

1.1. Nombre de la Asignatura: Proyectos de Investigación en Ingeniería
Código de la Asignatura: 7109287
Semestre Académico en que se desarrolla: Noveno (IX)

El desarrollo de las actividades académicas se distribuye en tres fases de seis semanas cada una

1.2. Peso Académico de la Asignatura.

CRÉDITOS	HORAS SEMANALES			HORAS SEMESTRALES	
	Horas Teóricas	Horas Prácticas		Horas Teóricas	Horas Prácticas
		Práctica Docente	Jefe de Prácticas		
03	02		02	36	36

Cada semestre académico comprende dieciocho semanas.
(Resolución Nº 3838-CU-07)

1.3. Código, nombre y créditos de Asignaturas Equivalentes:
7101035, Metodología de la Investigación, 3.0 créditos.

Fig. 9. Final Product, PDF File of the Syllabus in the Format Provided by the Center for Academic Development.

TABLE I. PRIORITY TASK – VALUE

Priority	Value
Low	1
Medium	2
High	3

TABLE II. TASKS TESTED

Id	Task	Pre Requisite (Id Task)	Priority	Expected Results	State	Department Head Review	Functional Testing
1	Application Login		1	Successful teacher access	Development - Integration	No	In Integration Test
2	Application Home		2	Recognition of assigned role. Load views associated with role	Development – Integration	No	In Integration Test
3	File upload Curriculum		3	Successful execution of SSIS packages. Correct information migration to BD	Development – Integration	No	In Integration Test
4	Load Academic Load Files		3	Successful execution of SSIS packages. Correct information migration to BD	Development – Integration	No	In Integration Test
5	Professor File Upload Screen		3	Successful execution of SSIS packages. Correct information migration to BD	Development – Integration	No	In Integration Test
6	Screen Load Bibliography Files		3	Successful execution of SSIS packages. Correct information migration to BD	Development – Integration	No	In Integration Test
7	Select Course Screen	3, 5	1	Course recognition by teacher. Association of information by course	Validated	Yes	Done
8	Display Syllable Screen	3, 4, 5, 6	3	Course information load. Creation of syllabus.	Validated	Yes	Done
9	Training Program Screens	3, 4, 5, 6	3	Course information load. Training Program Creation	Validated	Yes	Done
10	Enable Syllables	3	2	Display of entered syllabus. Change of states within the management process	Validated	Yes	Done
11	Display Syllable Screen	3	2	Display of entered syllabus. Search for past syllabus. Option to display the PDF document of the syllabus.	Validated	Yes	Done
12	Display Screen PDF Syllables	8, 9	3	Syllabus in PDF format	Development – Integration	No	Done

IV. CONCLUSION

This work presents a tool as a syllabus support which allows an improvement in the quality of the contents by providing the tool with suggestions for professors in the various sections that compose it. For this purpose, we had the idea of using the rules and the basis provided by the standard ISO/IEC/IEEE 42010 "Systems and Software Engineering - Architecture Description".

The evaluations carried out by the pilot show that the flow of syllabus development was intuitive and friendly. The various sections that make up the syllabus, which can be provided by the Center for Academic Development and that do not require professor intervention, are automatically resolved through the load. When you start the syllabus management flow, you immediately see the relationship between the syllabus and its preloaded information.

Likewise, the probability of error decreases because the professor has to implement a large percentage of options within the syllabus through selection; for example, the competencies of the graduation profile associated with the subject, which is expected information. That information requires to edit the criteria and participation; for example, the competences of the subject.

The relationship between the different sections of the syllabus, such as the Academic Identification and Formative Program was successful since it is verified that the data entered in the first one are displayed with coherence in the second one.

V. FUTURE WORKS

Activities related to business intelligence will be included in the tool for analytical processing, text mining and predictive analysis of syllabus content. Topics of the courses and teaching-learning strategies will be guided by an intelligent assistant for avoiding inconsistencies when it is filled by professors. The tool will suggest curricula recommendations from technical organizations when defining subjects of the course.

Finally, textual and graphical reports will be generated to show syllabic compliance during this execution.

ACKNOWLEDGMENTS

This research has been funded by the Universidad Católica de Santa María, through the project "Syllabus management through business intelligence techniques for support in decision making at the FCIFF-UCSM".

References

[1] H. S. Chung and J. M. Kim, “Semantic model of syllabus and learning ontology for intelligent learning system,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2014.

[2] A. Lile and E. Muka, “University/Student Management System: a case study of Sports University of Tirana.,” Proc. Multidiscip. Acad. Conf., 2018.

[3] A. Sharma, M. Kumar, and S. Agarwal, “A Complete Survey on Software Architectural Styles and Patterns,” in Procedia Computer Science, 2015.

[4] M. Tapia-Leon, M. L. Merchán-Gavilanez, A. Carrera-Rivera, and S. Luján-Mora, “Comparative Study Of Syllabus Structure In Latin American Universities,” in ICERI2017 Proceedings, 2017.

[5] K. E. Harper and J. Zheng, “Exploring Software Architecture Context,” in Proceedings - 12th Working IEEE/IFIP Conference on Software Architecture, WICSA 2015, 2015.

[6] O. Gerbé, J. Raynauld, and R. Camarero, “Learning outcomes in a model-based approach to curriculum design,” in 5th International Conference on Internet and Web Applications and Services, ICIW 2010, 2010, pp. 550–553.

[7] S. Chatvichienchai, “An XML-based syllabus repository system for inter-university credit exchange systems,” in ICCTD 2009 - 2009 International Conference on Computer Technology and Development, 2009.

[8] T. Kawaba, T. Tsuchiya, and K. Koyanagi, “Research on inter-domain collaborative syllabus services with cross-retrieval - IEEE Conference Publication,” 2012. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6268513. [Accessed: 30-May-2020].

[9] T. Sekiya, Y. Matsuda, and K. Yamaguchi, “A web-based curriculum engineering tool for investigating syllabi in topic space of standard computer science curricula,” in Proceedings - Frontiers in Education Conference, FIE, 2017.

[10] H. K. Bhargava, D. J. Power, and D. Sun, “Progress in Web-based decision support technologies,” Decis. Support Syst., 2007.

[11] G. R. Morales and J. P. Benedí, “Towards a reference software architecture for improving the accessibility and usability of open course ware,” in ACM International Conference Proceeding Series, 2017.

[12] N. Dimakis, L. Polymenakos, and J. Soldatos, “Enhancing learning experiences through context-aware collaborative services: Software architecture and prototype system,” in Proceedings - Fourth IEEE International Workshop on Wireless, Mobile and Ubiquitous Technology in Education, WMUTE 2006, 2006.

[13] D. C. Force and J. Zhang, “Knowledge discovery from within: An examination of records management and electronic records management syllabi,” Rec. Manag. J., 2016.

[14] R. Yucel, “Designing Knowledge Management Syllabus: How Business Administration Students Learn at Summer School?,” in Proceedings of the 9TH International Conference on Intellectual Capital, Knowledge Management & Organisational Learning, 2012.

[15] A. A. C. Júnior, S. Misra, and M. S. Soares, “ArchCaMO - A Maturity Model for Software Architecture Description Based on ISO/IEC/IEEE 42010:2011,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019.

[16] M. Vidoni and A. Vecchietti, “Towards a reference architecture for advanced planning systems,” in ICEIS 2016 - Proceedings of the 18th International Conference on Enterprise Information Systems, 2016.

[17] M. Chaabane, I. B. Rodriguez, and M. Jmaiel, “System of systems software architecture description using the ISO/IEC/IEEE 42010 standard,” in Proceedings of the ACM Symposium on Applied Computing, 2017.

[18] M. Chaabane, I. Bouassida Rodriguez, R. Colomo-Palacios, W. Gaaloul, and M. Jmaiel, “A modeling approach for Systems-of-Systems by adapting ISO/IEC/IEEE 42010 Standard evaluated by Goal-Question-Metric,” Sci. Comput. Program., 2019.

[19] International Organization of Standardization, “ISO/IEC/IEEE 42010:2011 - Systems and software engineering -- Architecture description,” ISOIECIEEE 420102011E Revis. ISOIEC 420102007 IEEE Std 14712000, 2011.

[20] M. Chaabane, F. Krichen, I. B. Rodriguez, and M. Jmaiel, “Monitoring of service-oriented applications for the reconstruction of interactions models,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2015.

Preparing Graduates with Digital Literacy Skills Toward Fulfilling Employability Need in 4IR Era: A Review

Khuraisah MN[1], Fariza Khalid[2], Hazrati Husnin[3]
Faculty of Education
Universiti Kebangsaan Malaysia
Bandar Baru Bangi, Malaysia

***Abstract*—This systematic review aims to review and synthesize employer expectations towards digital skills among graduates, steps, and measurements taken by higher education institutions to prepare students and harness motivation among students to make themselves competitive and marketable toward fulfilling employability needs in 4IR era. It was designed based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). Articles published between January 2016 and 2020 were sought from three electronic databases: Science Direct, Scopus, and Web of Science. Additional items gain from the Universiti Kebangsaan Malaysia repository are also considered to be reviewed. All papers were reviewed, and the quality assessment was performed. Twenty articles were finally selected. Data were extracted, organized, and analyzed using a narrative synthesis. The review identified three overarching themes: (1) Employer perspectives on their expectation from young graduates. (2) Institutions' views on how they should prepare their students for the 4IR era. (3) Students' perspectives on how they could motivate themselves. The systematic review provides insightful information on the required digital literacy skills among young graduates, expectations of the industry player, and how digital literacies can be developed in the institutions.**

Keywords*—Digital literacy; computer literacy; information literacy; employability

I. Introduction

The Fourth Industrial Revolution (4IR) defined as a complete digitization process that connects orders from the customer and manufacturing up to the after-sales service [1]. The term attracted attention to many leading industry players after the World Economic Forum in 2016 and became influential agenda to many developed countries. The pace for 4IR is fast in comparison to the previous industrial revolution, thus cause disruptive innovation through all social systems [2].

It gave a broad impact on most current job characteristics, not least resulting in the loss of traditional jobs and creating a new job opportunity. To stay relevant in the job market, one needs to have skills and the will to learn and re-learn [3]. The new job field will focus more on non-routine activities and include higher cognitive levels and soft skills. Therefore, this is an important trait which graduates must fulfill to meet the need of employers and to make themselves relevant and marketable.

The process of preparing student marketability in higher institution education is significantly critical. Students today are extensively exposed to technology and used it in most of their daily undertakings [4]. This group is known as digital natives, a term introduced by Prensky [5]. However, does being born after the digital era make these generations automatically digital literate?

Kirschner and Bruyckere [6] believed that the term digital native is too general and can be questionable. Even though students of this generation are used to the ever-connected digital world, they may not always be able to utilize the current technology as expected entirely. There is a digital literacy gap between them, and this literacy gap needs to be identified and minimize during the study years, making the students well prepared for the employability.

II. Digital Divide and Digital Literacy

Previously, the digital divide defined as the gap between those who have access to computers and the Internet with those who do not. This gap arises due to several barriers, such as consumer socioeconomic factors, high technology access costs, and complex system interfaces [7]. Prior studies on the digital divide focused on demographic, socioeconomic, gender, and ethnic, which ultimately led to the "have or not have" computer and internet access as the foundation for their argument [5, 8, 9, 10].

However, even though more people have access to the Internet through smartphones, this does not mean that the digital divide has closed [11]. Previous findings show that members of minority groups, young people, low-income, and less educated are more dependent on smartphones. Tsetsi and Rains [12] claimed that dependence on smartphones might bridge the digital divide, but it caused another problem, which was the information gap. In this case, the new digital divide is no longer due to the availability or accessibility of the Internet. However, it does look at the differences in internet usage, namely, as a second-level digital divide [13].

"In the 21st century, the term illiteracy no longer refers to those who do not know how to read and write, but it is to those who have not learned how to learn". The excerpt is from

Alvin Toffler's book Future Shock [12], an American author and futurist. Toffler, in his book, goes a long way, expecting that students will face too many options. Therefore, they should be taught how to perform the process of classification and reclassification of information, evaluate information, and be able to move abstract objects into concrete, and vice versa.

Focus is now not only on the issue of what one needs to know about technology but how the technology effectively and critically being utilized [14]. Additionally, able to evaluate the reliability of online content [15], and relate the information collected with the area of specialization [16]. Employers nowadays need digitally-savvy employees who can conduct their work effectively and seamlessly through ever-changing technologies and emerging media [17].

Therefore, this review paper led to presume that digital literacy is a must-have skill for young graduates to make themselves employability ready. Thus, the systematic review provides insightful information on the required digital literacy skills among graduates, the expectation of the industry player, and how the development of digital literacies by the higher education institution identified at the end of this systematic review. Its help foster the understanding and add essential knowledge in preparing graduates toward fulfilling employability need in 4IR era.

III. THE REVIEW

A. Aims

This literature review has two aims. The first aim is to explore published research studies for digital literacy expectations and employability preparedness among graduates. The second aim is to review and synthesis the digital literacy skills needed by the students at higher education institutions to meet the requirements of employability in the 4IR era. The critical questions that the researchers want to answer are:

- What are the expectations among employers that the students need to fulfill?
- How does the institution prepare its students with digital literacy before graduate?
- How can the students be more motivated to improve their digital literacy?

B. Design

This review adapts and adopts a systematic approach and conducted in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standard [18]. PRISMA used as a guideline to identify articles base on the pre-specified eligibility requirements as a framework for answering determine research questions. It allows the researcher to minimize biases by using explicit and systematic methods. The researcher then systematically read through each article and noting common themes on employer expectations, higher education institution preparation, and student's motivation toward development to join the job market.

C. Search Strategy

A Boolean search strategy was applied. Three electronic databases used, namely Scopus, Science Direct, and Web of Science. The keyword used is ("digital literacy" OR "technology literacy" OR "computer literacy" OR "information literacy") AND ("employability"), with some different syntax according to the database used.

D. Inclusion and Exclusion Criteria

The first level of filtering method done by using the build-in refine option provided by each database used. The inclusion criteria were (1) articles published from January 2016 to 2020, (2) English and Malay language publication, and (3) document type: research article. The exclusion criterion was (1) 2015 and before, (2) using a language other than English and Malay and (3) document type conference, book, book chapter, and proceeding.

The second level of filtering method conducted manually by going through each title abstract before carefully determine whether to review the articles or reject it based on (1) material is a review paper (2) Population is not related to tertiary education (3) Not-in-line with research objectives (4) Specific for library literacy (5) Research Note.

E. Search Outcomes

In line with the Preferred Reporting Item for Systematic Reviews and Meta-Analysis (PRISMA) researcher adopt and adapts the steps and showing the process using Fig. 1, Article searching and filtering process using the systematic approach.

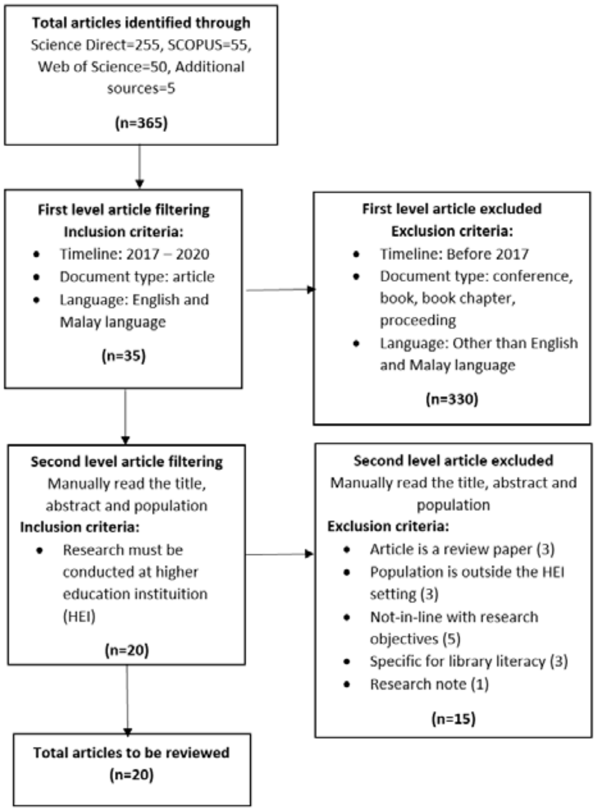

Total articles identified through
Science Direct=255, SCOPUS=55, Web of Science=50, Additional sources=5
(n=365)
First level article filtering
Inclusion criteria:
• Timeline: 2017 – 2020
• Document type: article
• Language: English and Malay language
(n=35)
First level article excluded
Exclusion criteria:
• Timeline: Before 2017
• Document type: conference, book, book chapter, proceeding
• Language: Other than English and Malay language
(n=330)
Second level article filtering
Manually read the title, abstract and population
Inclusion criteria:
• Research must be conducted at higher education instituition (HEI)
(n=20)
Second level article excluded
Manually read the title, abstract and population
Exclusion criteria:
• Article is a review paper (3)
• Population is outside the HEI setting (3)
• Not-in-line with research objectives (5)
• Specific for library literacy (3)
• Research note (1)
(n=15)
Total articles to be reviewed
(n=20)

Fig. 1. Article Searching and Filtering Process using the Systematic Approach.

This study is conducted under GG-2019-068 grants.

By using the specific search keyword, 365 articles identified. In the first level selection, only 35 items met the criteria. 330 articles excluded due to publication before 2016; documents are either conference, book, book chapter, or proceeding and published in a language other than English and Malay language.

In the second filtering phase, abstracts reading revealed 15 articles were irrelevant and not in-line with the current research objectives. Its include review paper, the population that is not related to higher education institution (HEI), specific for library literacy and research note.

These steps provide us with a systematic way to identify relevant literature for a review process. Finally, 20 articles were selected to be carefully read and synthesize their key findings to answer the research questions. The summary of the research articles is shown later in Table II.

IV. Result and Discussions

Table I shows the summary of the selected studies in terms of design, method, and location.

The research findings are structured according to the research questions and thoroughly discussed further details.

TABLE I. Summary of Types of Design, Method, and Sample of Selected Articles

Design	**Method**	**Sample / Settings**
Qualitative (8) Quantitative (7) Mix-Method (4) Action Research (1)	Diary writing (1) Phenomenology (1) Content Analysis (3) Exploratory (2) Case Study (1) Causal (1) Confirmatory Analysis (1) Reflective (2) Interview (5) Cross-sectional (1) Time series (1) Multinomial logistic regression analysis (1)	Students (14) Industry player (4) Academician (3) Others (3)

TABLE II. Summary (in Date Order) Research Articles Included in the Review

Study	*Aim*	*Design/Method*	*Sample/Settings*	*Major Finding*
The employability skills among students of Public Higher Education Institution in Malaysia [19]	This paper attempt to shows the relationship between employability skills and demographic data among graduates in public universities of Malaysia. Additionally, the researcher is also trying to proves other impacting factors for employability among the graduates.	A quantitative – cross-sectional survey	554 undergraduate students throughout several universities in Malaysia.	1. The study raised the importance of the institution to equip its students with information-technology environment friendly. 2. The researcher emphasizes the need to establish a good connection between students and the industrial player by having an internship program. 3. Teaching and learning approaches, facilities, and technology provided by the institution must in-line with the industry demands.
How does the pedagogical design of a technology-enhanced collaborative academic course promote digital literacies, self-regulation, and perceived learning of students? [17]	This paper focused on learning processes that enabled the development of various digital literacies conceptualized by the Digital Literacy Framework: photo-visual, information, reproduction, branching, social-emotional, and real-time thinking skills.	Qualitative – Diary writing	78 (82%) out of 95 students enrolled in four consecutive semesters of a graduate course in education.	1. The study suggests self-regulation and learning new technologies as an integral part of digital literacies. 2. The researcher recommends to include self-regulation learning skills in the Digital Literacy Framework.

Information skills for business acumen and employability: A competitive advantage for graduates in Western Sydney [20]	This paper aims to identify the information skills needed from the graduate before they enter the job market. The researcher also interested to look for current literacy skills thought in the university that can be applied when the student comes to the industry.	Qualitative – phenomenology studies.	24 participants, which consist of 12 graduates and 12 employers who have experience of supervising graduates.	1. Researcher confirms that information skill is a critical factor in competitive workplaces; therefore, the ability to evaluate information is the most valuable skills. 2. Participants emphasized the importance of information skills for individual endurance and autonomy, especially in enabling lifelong learning and adaptability. 3. Gathering information for problem-solving is close to a real situation in the working environment, thus engaging students in more profound and active learning opportunities. 4. Focusing on commonly used tools in the industry may be helpful for new graduates. 5. Other essential skills needed by the industry is a collaborative skill.
Student Perception On Graduate Employability In Era Of Fourth Industrial Revolution [21]	The objectives of this study are to investigate students' understanding of 4IR and their perception towards graduates' employability issues in this era.	Quantitative – Exploratory study	97 respondent selected using simple random sampling among students from the Faculty of Economy and Management, Universiti Kebangsaan Malaysia	1. The main finding in this study shows that's students have a fair understanding of 4IR and aware of the changes in the industry. 2. Results from the survey also show that students must have the willingness to build new skills to cope with the technology changes.
Strategic university practices in student information literacy development [22]	The researcher aims are to explore and understand how university practice has an impact on the development of students' information literacy.	Qualitative - interviews	31 respondent from various public research universities in Malaysia consists of students, lecturers, and librarians.	1. The most apparent findings in this paper are the acknowledgment that lecturers, librarians, students, and management all play an essential part in developing information literacy skills. 2. The inability to conduct independent information searching and manipulation will lead to a lack of problem-solving skills. 3. The outcome-based learning approach will motivate students to gather data from different resources and then construct new knowledge to meet their needs. 4. Applying various strategies in the teaching and learning approach will develop student information literacy skills. 5. Librarian suggests a collaborative information literacy program development between the library and academic department will have a significant impact on the development of information literacy among students.
Redesigning curriculum in line with Industry 4.0 [23]	This paper focused on redesigning curriculum and teaching practice, in-line with the 4IR trend.	Qualitative method – content analysis	72 publications about 4IR from 2013-2018 using a strict filtering process.	1. Universities need to be more determined to equip the upcoming generation with the skills related to capture, analyze, and communicate data using IT infrastructures. 2. Propose Model Curriculum Matrix include Big Data, IoT, Cloud Computing, AI, and AR.
Employability Skills Model for Engineering Technology Students [24]	This researcher seeks to create a forecast model for predicting the chances of attaining high-level employability skills among students of engineering technology. By identifying missing factors, the researcher aims to reduce the unemployment rate with the introduction of an invention program.	Quantitative - multinomial logistic regression analysis	1. 204 engineering technology students from various technical institutions chosen using a random sampling technique.	1. The result of this study indicates the ability to use information technology in innovative ways, extra-curriculum activities, and the industrial-based program used to predict the level of employability skills. 2. Study shows interesting factor identified affecting employability skills among graduates is parents occupational.

Changing skills for architecture students employability: Analysis of job market versus architecture education in Egypt [25]	Researcher focus on evaluating the transferable skills through two different courses, namely: (i) Architecture Education course (ii) Architecture, Engineering, and Construction with the industry job market in Egypt and to propose an approach to fill the gap between both.	Mix method: i. Questionnaires ii. Interviews	2. Questionnaires distributed using purposive sampling, distributed among fresh graduates. 3. Interview with professional architects, CEOs, founders of architectural firms, and academicians that have 15-30 years of working experience.	1. Strong evidence shows that by applying technology in teaching and learning activities will encourage and help the students to build their 21-century skills. 2. The result of this study also indicates that lecturers should give more flexibility and empowered the students in coursework development. 3. Interestingly, the interview shows that providing the students with an opportunity to connect with the company, increasing the timeframe of internship, promoting academic exchange, and encouraging participation in conferences in likely increase students' employability rate.
The digital culture of students of pedagogy specializing in the humanities in Santiago de Chile [26]	This study aimed to gather data that might give prospective input about teachers in a humanities course that come from the millennial generation. The researcher is interested in investigating whether cyber-culture helps future teachers using digital technology even though the humanities course lacks technology subject.	Quantitative – Exploratory research	118 students of Pedagogy in Secondary Education selected through a population study and surveys with a non-probabilistic sample.	1. The result of this study shows a technological transition has consequences for initial teacher training. 2. Teacher training should consider adapting changes and new cultural conditions from digital technology as a complement to the educational process.
An exploratory study of digital workforce competency in Thailand [27]	The researchers aim to define the individual skills desired by the digital workforce, whether the capabilities of the digital workforce can be group into categories, and establish the expectations of demand for competencies of the digital workforce.	Mix method i. Qualitative- content analysis and interview ii. Quantitative Exploratory research	1. Select 30 respondents from 289 IT organizations for an interview session. 2. 389 IT experts were selected using a purposive sampling method for the quantitative part.	1. The interview sessions suggest that types of competencies needed by the digital workforce are attitude, essential IT for work, critical-thinking, general communication, IT foundations, IT support, lifelong-learning, problem-solving, and teamwork. 2. Based on the questionnaire survey, it helps us to understand that competencies related to soft skills and foundation knowledge in IT are most required, followed by IT technical ability and IT management capabilities. 3. The most prominent finding emerges from the interview session are, almost all IT experts in Thailand expect graduates to show skills in lifelong learning, personal attitude, and dependable.
How the flipped classroom affects knowledge, skills, and engagement in higher education: Effects on students' satisfaction Luis [28]	The main objective is to present a successful flipped classroom proposal in higher education to understand better its influence in terms of knowledge, skill, and engagement.	Mix method i. Qualitative – in-depth interview. ii. Quantitative – instruments development	1. Conduct an in-depth interview with senior and knowledgeable academic scholars. 2. 160 students enrolled in the Macroeconomics module and the dual-degree program in Management & Law and Business Administration.	1. The finding of this study suggests a 4D-FLIPPED Model that consists of out-of-class activities, feedback, in-class activities, and the use of technology. 2. The researcher strongly emphasizes the need for two-way feedback as an effective way to link out-of-class activities and in-class-activities. 3. Additionally, the study results show that technology usage help established lecture outside of the classroom, thus creating more active learning inside the school.

Building student employability through interdisciplinary collaboration: an Australian Case Study [29]	The focus of this study is to have an in-depth understanding of the importance of information literacy to develop employability skills among students through the capstone project.	Qualitative - Case study	Respondents are Biological Sciences capstone course students at Macquarie University, Sydney, Australia.	1. Researcher finds that students give more attention to discipline-specific literacy and only give little attention to general information literacy. 2. The results reveal a specific emphasis on four dimensions of learning, namely self-awareness, opportunity awareness, decision making, and learning about the transition.
Classification Techniques for Predicting Graduate Employability [30]	The main objectives of this research are to identified factors affecting graduate employability.	Quantitative – time series	43863 UKM graduates data instance from the Malaysia Ministry of Education tracer study year 2011-2017.	1. The result from data tracer shows that age, course, faculty, co-curriculum activities, marital status, industrial training, and English proficiency are the most influential factors that determine graduate employability rate. 2. However, studies show that the most crucial factors are age, industrial training, and faculty.
Information Literacy in Practice: Content and Delivery of Library Instruction Tutorials [31]	The focus of this study is to analyze the content of library tutorials and match it with the level of difficulties based on Bloom Taxonomy.	Qualitative – content analysis	Online tutorials made publicly available through academic library web sites. The source for tutorials for this content analysis is the winners of ACRL's Excellence in Academic Libraries Award.	1. The result suggested that information literacy competencies can be matched to Bloom's Taxonomy under the first three categories, namely remembering, understanding, and applying. 2. The researchers suggest considering developing higher-order thinking skills when designing the online tutorial material.
Resignification of educational e-innovation to enhance opportunities for graduate employability in the context of new university degrees [32]	The researchers intend to present innovative teaching ways in higher education institutions for teaching methodologies and how it's related to employability ready among the graduates.	Action research	Undergraduate and master's degree students undertake two subjects: i. Event Organization, ii. Protocol and Institutional Relations Faculty of Economics and Business, University of Alicante, Spain.	1. Results demonstrate training program allowed graduates to improve their employability and career development opportunities by encouraging active participation and self-directed learning. 2. Participants believed technology literacy is essential, and it is crucial to have a wide range of skills and knowledge of digital tools, especially ones related to communication.
Employability skills of maritime business graduates: industry perspectives [33]	The focus of this study is to investigate current and future industry employability skills for maritime business graduates.	Mix-method	Conduct focus groups, individual interviews, and an online survey with senior managers in maritime organizations from Australia, USA, and Canada.	1. The most apparent finding that emerges from the interview session is that maritime business graduates must have excellent communication skills, problem-solving, adaptability, self-management, teamwork, digital literacy, and technology literate for them to be employable ready. 2. Noticeably that the maritime industry is moving toward digitization, therefore increase the demand for digital literacy and technology knowledge and skills. 3. The survey shows that the use and management of technology are the current skills that maritime business graduate focusing.
A new educational pattern in response to new technologies and sustainable development. Enlightening ICT skills for youth employability in the European Union [34]	This paper aims to identify the importance of training in ICTs to get a job. It intended to analyze the relation between ICTs' knowledge and employment for young people in the European Union.	Quantitative – Causal research	Respondent age 16 and 24 years old at the European Union.	1. The importance of emphasizing informal education on ICT is identified as an essential element to be employment ready. 2. The researcher highlights the importance of the promotion of self-learning environments as well as the long-life learning in ICTs to increase youth employability.

Motivational factors predicting ICT literacy: First evidence on the structure of an ICT motivation inventory [35]	This article focusses on presenting the construction and first validation of an ICT motivation inventory based on social cognitive theory. The researcher intends to predict computer-related knowledge and skills, deduced from ICT-related usage motives, self-effacing, and self-regulation.	Quantitative - Confirmatory factor analyses	323 German students between 16 and 27 years of age.	1. Significant findings of this paper have developed the concept of ICT motivation, measurement of motivation, and first empirical results on the measure's dimensional structure and construct validity. 2. Another critical finding is motivational and metacognitive ICT characteristics are essential for the acquisition and development of ICT knowledge and skills.
Encouraging student learning of control by embedding freedom into the curriculum: students perspectives and products. [36]	This paper aims to look at a novel mechanism for encouraging students to take more ownership of their learning in terms of control topics and improve learning engagement.	Qualitative – Reflective research	Undergraduate Bioengineering student.	1. The researcher proves that when the lecturer assigns extensive usage of modern technology within assignments, students show enthusiasm to complete it. 2. Another vital point is by encouraging students to develop teaching materials is the final part of the learning cycle will provide a useful learning experience.
"Old school" meets "new school": Using books and tablets to improve information literacy and promote integrative learning among business students [37]	This article aims to discuss the use of an ordinary book versus tablet program and highlight how these programs can enhance students' communication and collaboration skills, thus enhance employability.	Qualitative – reflective research	First-year business students, Nipissing University School of Business, Canada	1. The study shows that a common book program offers the potential for a variety of activities that improve the students' literacy and communication skills. 2. The use of a tablet allows students to held academic research during lectures and efficiently communicate the information with their classmates. 3. The researcher believes that peer feedback is an essential element of in-class research.

V. Employers: Skills Needed in the Fourth Industrial Era

Almost all areas of work in today's digital information societies require digital literacy skills. Competencies related to Information-Communication-Technology (ICT) are an integral part of employability [33, 20]. Considering the increasingly changing technological climate, self-regulated and lifelong learning are critical factors in keeping up with recent innovations in the industry [35].

Employers are not expecting fresh graduates to know everything when they begin their careers. Most of the employers are willing to train their newly recruit staff. Nevertheless, what they are looking for is the willingness to learn. Therefore it is safe to infer that self-learning is the most required skills by employers [20, 21, 25, 27, 28, 33, 34].

Today, we are living in an era where unlimited information is available and accessible. Locating and accessing data can be considered as lower-order skills, as these tasks are process-based and generally do not require students to analyze or synthesize information. Therefore, according to Saunders [31], the ability to evaluate sources and content would be considered as higher-order information skills. Graduates expected to be able to select, synthesize, and to leverage information in decision making. These skills will ultimately provide a competitive advantage to them [20, 22]. Another vital skill highly demanded among employers in this digitization era is analytical skills [20, 25, 33]. Employees were supposed to be able to think objectively, holistically, gather data, and evaluate it to support the decision-making process.

Communication skills are also pertinent [30]. Effective communication requires empathy, active listening, written, and verbal skills. Also, the employer expects the employee to have functional computer literacy, experience in the use of core computer application, and knowledge in data processing and information dissemination. Students graduating in this era must and should be able to communicate effectively in all electronic forms [33].

New media literacy and virtual collaboration are two additional critical skills needed in the future workforce. These skills will provide employees with the opportunity of integrating productivity-software or technology into their work and encouraging them to collaborate easily [34]. The collaboration skills is another expected skills required by the current industrial revolution era [17, 23, 29, 34, 37].

Minimize the gaps in soft skills is essential to many employers. Jewell [20] strongly recommended that higher education institutions improve communication, interpersonal and critical skills for students in classrooms. Also, the student should be allowed to choose suitable media when presenting the information they have attained, as this will nurture the students' soft skills gradually. Applying outcome-based learning could also encourage the student to construct knowledge, based on the information gathering activities [22].

According to Khodeir and Nessim [25], there are two quality categories among graduates, which are technical and non-technical skills. Universities seldom get a negative review of the technological capabilities of their students. Nevertheless, lack of employability skills among students is the critical reason for the frustration among employers. Similarly, Mang [37], in their research, found out that business graduates often become specialized in one particular area, but unable to relate it with world knowledge which is beyond their area of expertise. If graduates can combine highly develop meta-skills, it will provide them with more flexibility and lead to a successful career.

VI. Institutions: Preparing Students with Digital Literacy Skills

Generally, the Higher Education Institution (HEI) is the last formal institution used to prepare students before they embark into the industry. Therefore, HEI must design its curriculum and nurture the needed skills during the students' study period.

However, Ayale-Pérez and Joo-Nagata [26] say that the millennial generation grown up with technologies. Therefore, even if their initial training in universities did not emphasize in the applied computer or computer-related subjects, this generation could still be able to work with technological products and services naturally. On the other hand, Mang [37] believed that today's ease of information access often provides students with a false feeling of competency. Even if they could easily find information, students cannot still critically evaluating the data. Therefore, nurture these skills must be during the studies period that is critically needed.

Whereas discipline-specific mastery is vital, exposure to integrated learning provides students with flexibility is also required in the industry [37]. Incorporating information and communications teaching in universities may be troublesome, but it will be beneficial in preparing students for the work market [34].

The student should be assisted in the development of mastery office computer applications to improve digital literacy among them. Assignment should be versatile, allowing students to select information services and approaches freely. Flexibility, experiential, and active learning are likely to involve students in more in-depth education, thereby improving the skills required in the industry [20].

Next, by adding more practical implications, real-life projects, and embedded digital literacy skills in the teaching and learning approach, it will encourage and assist students in building their digital literacy skills [22, 23]. A method such as a flipped classroom will provide better exposure and influence in terms of knowledge, skills, and engagement to the students [28].

Furthermore, regardless of how familiar the digital natives with technology, one must admit that technological changes are so fast. Thus, doing continuous training is significant [34]. HEI needs to keep up with these trends and tried to embrace it in their teaching approach. Ellahi [23] suggested that a curriculum framework should encapsulate the elements of big data, Internet of things, cloud computing, artificially intelligent, and augmented reality, thus producing university graduates that meet the expectation of 4IR industry.

Another way to establish familiarity between study environment and job environment is by providing students with an internship program [19, 24, 30]. It is also the responsibility of the institution to offer excellent facilities and technology, in-line with the trend and demand of the industry player [19].

Based on the articles reviewed, it is noticeable that the responsibility of developing employability skills among students should not be limited in Information Technology classes. Instead, it can incorporate it in all subjects, in term of teaching approach and assignments design. Selecting appropriate technology to be used is the responsibility of both lecturers and students, supported by the institution. Only then, all the needed skills will be embedded in the student's mind organically.

VII. Students: Motivation Needed Toward Developing Digital Literacy Skills

The expectation among employers and preparation by the institution is only half the effort in preparing the students for employability. To survive the fast pace of the 4IR era, students need to have strong motivation towards developing skills required themselves.

Besides, when lecturers are more flexible and embracing their students in designing coursework, these will give students the feeling of empowerment. Students' empowerment would have a significant impact on them [17, 25, 32, 36]. It gave the ultimate motivation to learn by providing them with a reason to conduct a particular exploration. Also, setting autonomous goals and applying participatory educational approaches will foster critical thinking and the development of professional skills as needed in the industry [32].

Students need to play an active role in their learning. By doing so, they can acquire a series of abilities associated with content-knowledge education that will make them more desirable when they graduate [28].

Clearly, in the selected articles reviewed, self-motivation can be nurtured by giving students more power to explore and use a wide variety of technologies available. Students' empowerment will encourage engagement in their learning, therefore, making them more motivated to learn and explore. These characteristics of self-learning and willingness to learn are among the most desirable skills in the industry.

VIII. An Emerging Finding of Digital Literacy

Most activities using ICT by young adults nowadays are predominantly for entertainment and social interaction purposes, which including game playing and passive forms of media consumption such as video streaming. These ICT-related activities do not require extensive technological or information-related knowledge and skills. Widespread usage in entertainment and social media interaction does not add to digital literacy skills [35].

Senkbeil & Ihme [35] further stated that not all ICT activities, entertainment, and social interaction activities, in particular, enhance ICT knowledge and skills. Assumed based on the list of articles reviewed, experience with the information-related task, is a vital prerequisite to acquire functional ICT knowledge and then consider as digitally literate.

In measuring the digital literacy level, Blau et al. [17] utilize Digital Literacy Framework (DLF). Digital Literacy includes namely, social-emotional thinking by separately addressing communication issues, different levels of teamwork – psychological ownership towards a collaborative outcome. The research finding emphasized the importance of independent study in higher education and students' ability to tailor learning experiences to meet their individual needs. Therefore there is a need to include self-regulation learning skills in the DLF.

IX. Conclusion

Based on the synthesis of literature selected, we can conclude that digital literacy not only focuses on technological literacy but more on how we use technology in decision making. The processes of developing digital literacy skills should not be limited to the library or computer classes. Instead, if lecturers could find exciting ways to encourage their students, as a native digital, students will develop the skill needed, intuitively. What the students need is a reason to do so.

As we are now moving toward the era of the fourth industrial revolution, employers expect graduates to have competencies related to ICT, able to do self-learning and excellent information skills that come together with analytical skills. When there is so much information available, know how to look for information is no longer appreciated. Instead, graduates expected to have the ability to select, synthesize and leveraging data in decision making. Other than that, the competency to communicate effectively using all suitable electronic form and the medium is also a vital skill.

The focus when facing the 4IR era is the acknowledgment that education institutions play an essential role in preparing and develop students' skills and knowledge to meets the demand of industrial players. Lecturers are encouraged to use and embed technology in the teaching approach. Students should be given more empowerment in designing assignments so that it will provide them with the motivation to look for various forms of information and exciting ways to deliver their findings. This skill will naturally prepare the students for employability. The fast pace of the 4IR era requires every stakeholder, including students, academic institutions, and industry players ready to face ever-changing technology and ways to do things with a strong will to learn and re-learn.

Acknowledgment

This study is conducted under GG-2019-068 grant.

References

[1] M. Wilkesmann and U. Wilkesmann, "Industry 4 . 0 – Organizing Routines or Innovations ?," J. Inf. Knowl. Manag. Syst., vol. 48, no. 2, p. pp.238-254, 2018.

[2] S. Xu, H. H. Yang, J. MacLeod, and S. Zhu, "Social media competence and digital citizenship among college students," Convergence, vol. 25, no. 4, pp. 735–752, 2019.

[3] S. Ra, U. Shrestha, S. Khatiwada, S. W. Yoon, and K. Kwon, "The rise of technology and impact on skills," Int. J. Train. Res., vol. 17, no. sup1, pp. 26–40, 2019.

[4] S. Yong and P. Gates, "Born Digital : Are They Really Digital Natives ?," Int. J. e-Education, e-Business, e-Management e-Learning, vol. 4, no. 2, pp. 2–5, 2014.

[5] M. Prensky, "Digital Natives, Digital Immigrants Part 1," Horiz., vol. 9, no. 5, pp. 1–6, 2001.

[6] P. A. Kirschner and P. De Bruyckere, "The myths of the digital native and the multitasker," Teach. Teach. Educ., vol. 67, pp. 135–142, 2017.

[7] D. E. Frederick, "The Fourth industrial revolution and the digital divide," Libr. Hi Tech News, vol. 36, no. 7, pp. 12–17, 2019.

[8] N. Mohd Daud, C. Siong Choy, A. Aris, I. S. Mohamed, R. Kamarudin, and R. Zainuddin, "The effects of students' backgrounds and attitudes on computer skills in Malaysia," Int. J. Manag. Educ., vol. 1, no. October, pp. 371–388, 2007.

[9] S. Naidoo and J. Raju, "Impact of the digital divide on information literacy training in a higher education context," South African J. Libr. Inf. Sci., vol. 78, no. 1, pp. 34–44, 2012.

[10] S. Salam, M. Yang, A. Shaheen, M. Movahedipour, and J. Zeng, "ICT and students performance in Pakistan," Hum. Syst. Manag., vol. 36, no. 4, pp. 277–284, 2017.

[11] A. J. A. M. van Deursen and J. A. G. M. van Dijk, "The digital divide shifts to differences in usage," New Media Soc., vol. 16, no. 3, pp. 507–526, 2014.

[12] E. Tsetsi and S. A. Rains, "Smartphone Internet access and use: Extending the digital divide and usage gap," Mob. Media Commun., vol. 5, no. 3, pp. 239–255, 2017.

[13] A. J. A. M. van Deursen and J. A. G. M. van Dijk, "The first-level digital divide shifts from inequalities in physical access to inequalities in material access," New Media Soc., vol. 21, no. 2, pp. 354–375, 2019.

[14] D. Buckingham, "Defining digital literacy: What do young people need to know about digital media?," in Nordic Journal of Digital Literacy, vol. 2010, no. 4, VS Verlag für Sozialwissenschaften, 2010, pp. 59–71.

[15] G. Polizzi, "Digital literacy and the national curriculum for England: Learning from how the experts engage with and evaluate online content," Comput. Educ., p. 103859, 2020.

[16] C. Mang, N. Brown, and L. Piper, "'Old school' meets 'new school': Using books and tablets to improve information literacy and promote integrative learning among business students," Int. J. Manag. Educ., vol. 15, no. 3, pp. 449–455, 2017.

[17] I. Blau, T. Shamir-Inbal, and O. Avdiel, "How does the pedagogical design of a technology-enhanced collaborative academic course promote digital literacies, self-regulation, and perceived learning of students?," Internet High. Educ., vol. 45, no. December 2019, p. 100722, 2020.

[18] L. Shamseer et al., "Preferred reporting items for systematic review and meta-analysis protocols (Prisma-p) 2015: Elaboration and explanation," BMJ, vol. 349, no. January, pp. 1–25, 2015.

[19] M. Z. Abd Majid, M. Hussin, M. H. Norman, and S. Kasavan, "The employability skills among students of Public Higher Education Institution in Malaysia," Malaysian J. Soc. Sp., vol. 16, no. 1, pp. 36–45, 2020.

[20] P. Jewell, J. Reading, M. Clarke, and L. Kippist, "Information skills for business acumen and employability: A competitive advantage for graduates in Western Sydney," J. Educ. Bus., vol. 95, no. 2, pp. 88–105, 2020.

[21] L. Wei Sieng and A. Noradilah, “Persepsi Pelajar Terhadap Kebolehpasaran Graduan dalam Era Revolusi Perindustrian 4 . 0 (Student Perception On Graduate Employability In Era Of Fourth Industrial Revolution) LAI WEI SIENG * & NORADILAH AZIZ,” J. Pers. Pelajar, vol. 22, no. 2, pp. 121–127, 2019.

[22] A. K. Aidah, A. B. Kamariah, and M. S. Parilah, “Strategic university practices in student information literacy development,” Psychol. Appl. to Work An Introd. to Ind. Organ. Psychol. Tenth Ed. Paul, vol. 4, no. 19, pp. 253–259, 2019.

[23] R. M. Ellahi, M. U. Ali Khan, and A. Shah, “Redesigning curriculum in line with industry 4.0,” Procedia Comput. Sci., vol. 151, no. 2018, pp. 699–708, 2019.

[24] A. K. Zatul Iradah and M. Siti Mistima, “Employability skills model for engineering technology students,” J. Tech. Educ. Train., vol. 11, no. 2, pp. 79–87, 2019.

[25] L. M. Khodeir and A. A. Nessim, “Changing skills for architecture students employability: Analysis of job market versus architecture education in Egypt,” Ain Shams Eng. J., , 2019.

[26] T. Ayale-Pérez and J. Joo-Nagata, “The digital culture of students of pedagogy specializing in the humanities in Santiago de Chile,” Comput. Educ., vol. 133, no. June 2018, pp. 1–12, 2019.

[27] V. Siddoo, J. Sawattawee, W. Janchai, and O. Thinnukool, “An exploratory study of digital workforce competency in Thailand,” Heliyon, vol. 5, no. 5, p. e01723, 2019.

[28] L. R. Murillo-Zamorano, J. Á. López Sánchez, and A. L. Godoy-Caballero, “How the flipped classroom affects knowledge, skills, and engagement in higher education: Effects on students’ satisfaction,” Comput. Educ., vol. 141, no. October 2018, 2019.

[29] S. Lin-Stephens et al., “Building student employability through interdisciplinary collaboration : an Australian Case Study Building student employability through interdisciplinary collaboration : An Australian Case Study,” Coll. Undergrad. Libr., vol. 26, no. 3, pp. 234–251, 2019.

[30] Z. Othman, S. W. Shan, I. Yusoff, and C. P. Kee, “Classification techniques for predicting graduate employability,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 8, no. 4–2, pp. 1712–1720, 2018.

[31] L. Saunders, “Information Literacy in Practice: Content and Delivery of Library Instruction Tutorials,” J. Acad. Librariansh., vol. 44, no. 2, pp. 269–278, 2018.

[32] R. M. T. Valdés, A. S. Soriano, and C. L. Álvarez, “Resignification of educational e-innovation to enhance opportunities for graduate employability in the context of new university degrees,” J. New Approaches Educ. Res., vol. 7, no. 1, pp. 70–78, 2018.

[33] P. S. L. Chen, S. Cahoon, H. Pateman, P. Bhaskar, G. Wang, and J. Parsons, “Employability skills of maritime business graduates: industry perspectives,” WMU J. Marit. Aff., vol. 17, no. 2, pp. 267–292, 2018.

[34] J. Picatoste, L. Pérez-Ortiz, and S. M. Ruesga-Benito, “A new educational pattern in response to new technologies and sustainable development. Enlightening ICT skills for youth employability in the European Union,” Telemat. Informatics, vol. 35, no. 4, pp. 1031–1038, 2018.

[35] M. Senkbeil and J. M. Ihme, “Motivational factors predicting ICT literacy: First evidence on the structure of an ICT motivation inventory,” Comput. Educ., vol. 108, pp. 145–158, 2017.

[36] J. A. Rossiter, L. Barnett, E. Cartwright, J. Patterson, N. Shorten, and J. Taylor, “Encouraging student learning of control by embedding freedom into the curriculum: student perspectives and products,” IFAC-PapersOnLine, vol. 50, no. 1, pp. 12149–12154, 2017.

[37] C. Mang, N. Brown, and L. Piper, “‘Old school’ meets ‘new school’: Using books and tablets to improve information literacy and promote integrative learning among business students,” Int. J. Manag. Educ., vol. 15, no. 3, pp. 449–455, 2017.

A Comparative Study of Eight Crossover Operators for the Maximum Scatter Travelling Salesman Problem

Zakir Hussain Ahmed
Department of Mathematics and Statistics, College of Science
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Kingdom of Saudi Arabia

***Abstract*—The maximum scatter traveling salesman problem (MSTSP), a variation of the famous travelling salesman problem (TSP), is considered here for our study. The aim of problem is to maximize the minimum edge in a salesman's tour that visits each city exactly once in a network. It is proved be NP-hard problem and considered to be very difficult problem. To solve this kind of problems efficiently, one must use heuristic/metaheuristic algorithms, and genetic algorithm (GA) is one of them. Out of three operators in GAs, crossover is the most important operator. So, we consider eight crossover operators in GAs for solving the MSTSP. These operators have originally been designed for the TSP which can also be applied on the MSTSP after some modifications. The crossover operators are first illustrated manually through an example and then executed on some well-known TSPLIB instances of different types and sizes. The obtained comparative study clearly demonstrates the usefulness of the sequential constructive crossover operator for the MSTSP. Finally, a relative ranking of the crossover operators is reported.**

Keywords—Traveling salesman problem; maximum scatter; genetic algorithms; crossover operators; sequential constructive crossover

I. Introduction

The travelling salesman problem is a famous problem (TSP) that aims to find shortest tour of a salesman who starts his journey from depot node and visit all remaining n nodes (cities) such that each node is to be visited only once and then returns to the depot. It is a NP- Hard problem [1] that is very easy to define but difficult to solve. Several researches have been done to deal with the problem and consequently numerous good algorithms have been reported in the literature. However, few circumstances require different restrictions on the acceptability of a tour as solution. One such restriction is to maximize the minimum cost edge in a tour of the salesman, which is named as maximum scatter TSP (MSTSP). In MSTSP, given a weighted graph, the aim is to find a Hamiltonian circuit so that the minimum cost edge is maximized. That is, the aim is to make each point away from (scattered) its previous and next points in the circuit. It is also called the max-min 1-neighbour TSP. In the max-min m-neighbor TSP, the aim is to maximize the minimum cost between any city and all its m-neighbours in the Hamiltonian circuit. These problems are close to the bottleneck TSP (BTSP) [2].

The MSTSP, defined first in [3], has application in operations involving heating workpiece, where it is equally important to keep each point away from its immediate ancestor and successor along with its m-neighbors for allowing cooling period in each operation. It has application in some other manufacturing processes that attach metal sheets together. After required alignment, the topmost sheet has some pre-specified points where riveting operations are applied to attach the sheets together. To avoid nonuniform deformation of the sheets, it is required to arrange the riveting process such that the distance between any rivet and its next rivet is very large; that means, the riveting operations must be scattered. It has application in some kind of medical imaging also. During imaging physical functions by Dynamic Spatial Reconstructor, radiation sources are positioned on the upper half of a circular ring and sensors are positioned directly opposite on the lower. The 'firing sequence' decides the sequence of radiation sources along with their associated sensors, generally periodically. The sensors gather energy intensity which goes through the patient positioned in the middle of the ring. It is required that if the i^{th} source is activated, then its neighbour sources (for example, $(i–1)^{th}$, $(i+1)^{th}$, $(i + 2)^{th}$, etc.) must not be activated, and hence some amount of scattering occurs [1]. The problem can be applied to a case where someone is falsely accused of a crime and given is death penalty. Now, he tries to escape from the police by visiting different safe places across his country to avoid the capture. Throughout his journey, he looks for a tour such that the smallest distance between consecutive places is very big [4].

The MSTSP can be formally defined as follows: Let a network with a set of n nodes, considering node 1 as depot node and a travel cost (time or distance, etc.) matrix $C=[c_{ij}]$ of order n connected with ordered pair (i, j) of nodes is given. Let $(1=\alpha_0, \alpha_{1,} \alpha_{2,} ,....,\alpha_{n-1} , \alpha_n=1) \equiv \{1\rightarrow\alpha_1\rightarrow\alpha_2\rightarrow.... \rightarrow\alpha_{n-1}\rightarrow1\}$ be a tour. The tour cost is defined as $\min\{c_{\alpha_i,\alpha_{i+1}}: i = 0, 1, 2, \ldots., n-1\}$. The aim is to find a tour that has maximum tour cost. The problem can be transformed to a BTSP by the transformation $d_{ij} = L\text{-}c_{ij}$ where $D = [d_{ij}]_{nxn}$ is equivalent BTSP's cost (or distance) matrix and L is very large number [5].

Since the problem is NP-hard, obtaining optimal solution using exact method is very hard, if not possible. The moderate sized TSP instances have been effectively solved by using operations research methods, like branch-and-bound [6], lexisearch [7], branch-and-cut [8] and local search [9]. As the problem size increases, obtaining exact solution is very hard. For solving large sized instances, one must go for heuristic algorithms, which, of course, don't promise to obtain optimal solution of a problem instance; however, they give near exact solution very quickly. Hence heuristic algorithms are used to solve some difficult problems. The most current algorithms that are used to solve various difficult optimization problems are termed as metaheuristics. There are metaheuristic algorithms based on simulated annealing [10], tabu search [11], insertion heuristic [12], ant colony algorithm [13], genetic algorithms [14], variable neighbourhood method [15], etc. However, genetic algorithms (GAs) are extensively applied methods amongst modern metaheuristics, and hence, we are applying GAs to solve the MSTSP.

Genetic Algorithms (GAs) first developed by John Holland in 1975, based on imitating the Darwinian survival-of-the-fittest theory among different species created by arbitrary changes in the chromosomes' structure in the natural biology [14]. They are powerful and robust metaheuristic algorithms for solving large-sized problem instances. They have been fruitfully applied to numerous combinatorial optimization problems to find their solutions. Each feasible solution of a problem may be assumed as a chromosome whose fitness is measured by its objective function value [16].

In general, simple GAs begin using randomly created a set of chromosomes called initial population, also termed as pool of genes, and then apply, mainly three, genetic operators to produce new, and possibly, better populations in subsequent generations. The first operator is selection which probabilistically copies and discards some of the chromosomes of the present generation to the next generation. Crossover is the second operator that selects randomly a pairs of chromosomes and mates to produce new chromosomes. The third operator is mutation, which randomly alters some position values (genes) of a chromosome. Crossover is very powerful operator in the GA search. Mutation diverges the GA search space. Generally, probability of applying mutation operator is fixed very low comparative to probability of crossover operator [14].

The crossover operators which have been developed for the usual TSP are also applied on the variant TSP after some modification. Since the MSTSP is a variant TSP, we consider eight crossover operators in simple GAs for solving the MSTSP. The crossover operators are first illustrated manually through an example and then executed on some well-known TSPLIB instances of different types and sizes. The obtained comparative study clearly demonstrates the usefulness of the sequential constructive crossover operator [16] for the MSTSP. Finally, a relative ranking of the crossover operators is reported.

This paper is organized as follows: A survey of the literature for the MSTSP is reported in Section II. Section III develops simple genetic algorithms using eight crossover operators for the problem, whereas, Section IV reports computational experiments for eight crossover operators. Finally, Section V presents conclusion and future works.

II. Related Work

There are few literatures about MSTSP, and the relevant papers are as follows. Arkin et al. [1] developed the first method for solving the problem. The problem was shown be NP-hard and unless P = NP, any no constant-factor approximation method can be designed. They developed factor-2 (which is best factor) approximation method with the triangle inequality for the max-min 1-neighbor TSP, for the cycle and path versions. Further, the method expanded to obtain a factor-2 approximation solution for the max-min 2-neighbor TSP, for cycle as well as some cases of path version. They also developed methods for the max-min 2-neighbor TSP with the triangle inequality, for both the path and cycle versions. The methods also expanded to obtain an approximation solution for path version of the max-min m-neighbor TSP.

Chiang [17] developed approximation methods for the max-min 2-neighbor TSP that follows the triangle inequality. He developed approximation methods for the path and cycle versions by improving methods in [1]. As mentioned, both algorithms are much simpler. John [4] also studied many works of MSTSP and its relevant models. Kabadi and Punnen [18] obtained an approximation method for the MSTSP that satisfies the triangle inequality, which is claimed to be the best bound for the case. Hoffmann et al. [19] extended the algorithm in [1] that produces optimal solutions for the nodes on a line to a regular mxn-grid. As reported, in some particular cases, the algorithm takes linear computational time to find an optimal tour.

The MSTSP is close to the BTSP, where the aim is to minimize the maximum cost edge in a Hamiltonian circuit [20]. Exact algorithms based on lexisearch approach have been developed ([21], [22]). Also, hybrid algorithms have been proposed for solving the problem ([23], [24]). Another closely related problem of the MSTSP is the maximum TSP (MaxTSP), in which the aim is to maximize total length of a tour in a Hamiltonian circuit [25]. A hybrid GA is proposed for solving the problem [26].

Dong et al. [27] proposed the multi-salesmen version of the MSTSP, multiple MSTSP (MMSTSP). They developed three improved GAs using greedy initialization, hill-climbing and simulated annealing algorithms to improve GAs for solving the MMSTSP. As claimed the improved algorithms are efficient algorithms and can reveal several characteristics in finding the solution of the problem.

A multi-start iterated local search approach is proposed in [28] for the MSTSP. Two local search algorithms based on insertion and modified 2-opt moves have been developed as part of our approach. To investigate the effectiveness of the method, it is tested on the TSPLIB instances, and found very good results.

III. Simple Genetic Algorithms for the MSTSP

Beginning with an initial population, a simple GA recurrently applies three genetic operators, selection, crossover and mutation, until the stopping criterion is satisfied. Though GA is among the best metaheuristic algorithms, but its performance verily depends on initial chromosome population, three operators and some parameters [14] that are discussed in this section.

A. Chromosome Representation and Initial Population

There are numerous ways to represent solutions as chromosomes for the TSP and its variants. Path representation is considered for the MSTSP that lists labels of nodes so that no any node is repeated in a chromosome. Suppose, {1, 2, 3, 4, 5, 6, 7, 8} represents the node labels in an 8-node instance, then the tour {1→7→2→3→8 → 4→6→ 5 →1} can be denoted by (1, 7, 2, 3, 8, 4, 6, 5). The objective function is defined as the sum of the costs of edges in the tour. Since the problem is a maximization problem, fitness and objective functions are same. Usually a simple GA begins with a pool of chromosomes called initial population. Here randomly created initial population is considered.

B. Selection Operator

In selection process, strings/chromosomes are replicated to the mating pool of next generation based on probabilities associated with their fitness function values. By transferring a higher portion of fitter chromosomes to the next generation, selection imitates the Darwinian survival-of-the-fittest in natural biology. Here, no any new chromosome is formed. Generally, the proportionate selection is applied in which any chromosome is chosen based on a probability that is calculated as proportional to its fitness function value. For example, roulette wheel selection, tournament selection, stochastic remainder, etc. are some of them. We consider stochastic remainder selection method [29] for our GAs.

C. Crossover Operators

Crossover operators selects two parent chromosomes and a point throughout the length of the chromosomes and exchanges their information after the crossover point. It performs a very significant role in GAs. Several good crossover methods are suggested for the TSP in the literature which are supposed to be good for the MSTSP. For example, partially mapped crossover [30], ordered crossover [31], alternating edges crossover [32], cycle crossover [33], edge recombination crossover [34], generalized N crossover [35], greedy crossover [32], sequential constructive crossover [16] are some of them. We are going to investigate these eight crossover methods.

1) Partially mapped crossover operator. The partially mapped crossover (PMX) uses two crossover points and produces two offspring chromosomes [30]. It defines exchange mappings in the segment between the crossover points. It is the first crossover operator designed for the TSP in GAs. We illustrate the PMX through the 8-node example instance along with its cost matrix given in Table I and the parent chromosome pair P_1: (1, 5, 4, 7, 8, 2, 3, 6) and P_2: (1, 8, 3, 4, 5, 6, 2, 7) with costs 3 and 1 respectively. We start journey (computation) from the first gene (headquarters), node 1.

Let the arbitrarily assumed cut points are after 3rd and 6th genes that are marked with "|", as follows:

P_1: (1, 5, 4 | 7, 8, 2 | 3, 6) and

P_2: (1, 8, 3 | 4, 5, 6 | 2, 7)

TABLE I. THE COST MATRIX

Node	1	2	3	4	5	6	7	8
1	0	10	15	95	66	55	29	2
2	61	0	55	22	50	72	1	58
3	45	50	0	69	7	89	22	78
4	91	67	75	0	35	27	34	89
5	60	36	90	31	0	50	61	77
6	3	82	20	70	39	0	77	28
7	16	57	26	86	53	19	0	69
8	13	14	54	8	84	37	87	0

The mapping segments are between these cut points. So, the exchange mappings are 7↔4, 8↔5 and 2↔6. These mapping segments are copied to the offspring chromosomes as follows:

O_1: (1, *, * | 7, 8, 2 | *, *),

O_2: (1, *, * | 4, 5, 6 | *, *)

We now add some more genes from the alternative parent chromosomes that do not form invalid chromosome as follows:

O_1: (1, *, 3 | 7, 8, 2 | *, *),

O_2: (1, *, * | 4, 5, 6 | 3, *)

The node 8 should be in the place of first * in O_1 which comes from P_2, but, since it is available in O_1, so after checking the mapping 8↔ 5, node 5 is placed there. The second * in O_1 should be 2 which comes from P_2, but, since it is available in O_1, so after checking the mapping 2↔6, node 6 is place there. Finally, 4 is added at third *. So, the first complete offspring becomes.

O_1: (1, 5, 3 | 7, 8, 2 | 6, 4) with cost 14.

Similarly, one can create the second complete offspring as:

O_2: (1, 8, 7 | 4, 5, 6 | 3, 2) with cost 2.

2) Ordered crossover operator. To create offspring chromosomes, the ordered crossover (OX) selects a subsegment of a route from one parent chromosome and then preserves the relative order of genes from the other one [31]. We choose the same parent chromosomes and cut points marked with "|" as:

P_1: (1, 5, 4 | 7, 8, 2 | 3, 6) and

P_2: (1, 8, 3 | 4, 5, 6 | 2, 7)

We always fix first gene as 'node 1'. At first, the offspring are created by simply copying the segments between these cuts into the offspring as:

O_1: (1, *, * | 7, 8, 2 | *, *),

O_2: (1, *, * | 4, 5, 6 | *, *)

Now, starting from 2nd cut of one parent chromosome, the genes (un-available) from the other chromosome are copied in the same sequence. The order of genes in P_2 from the 2nd cut is

{2 →7→8→3→4→5 →6}. After ignoring the already available genes 7, 8 and 2 in O_1, the order becomes {3→4→5 →6}, which is added in O_1 starting from the 2nd cut point:

O_1: (1, 5, 6 | 7, 8, 2 | 3, 4) with cost 14.

Similarly, second offspring is created as:

O_2: (1, 8, 2 | 4, 5, 6 | 3, 7) with cost 2.

3) Alternating edges crossover operator. The alternating edges crossover (AEX) operator considers a chromosome as a cycle of arcs [32] that creates only one offspring by choosing alternative arcs from the parents. In case of invalid offspring, random arc is chosen to create valid offspring. We choose the same example chromosomes P_1: (1, 5, 4, 7, 8, 2, 3, 6) and P_2: (1, 8, 3, 4, 5, 6, 2, 7).

At the beginning the arc (1, 5) is chosen from P_1 and the arc (5, 6) from P_2 are added to the offspring. Next, the arc (6, 1) is chosen P_1, as 6 is the last node, but this arc creates a cycle. So, an arc leaving node 6 to an unvisited node is chosen randomly. Suppose the arc (6, 2) is chosen. Next, the arc (2, 7) from P_2, (7, 8) from P_1 and then (8, 3) from P_2 are added to the current offspring. Finally, the following offspring may be created:

O: (1, 5, 6, 2, 7, 8, 3, 4) with cost 1.

All arcs present in the offspring (O) are inherited from either of the parents.

4) Cycle crossover operator. The cycle crossover (CX) creates offspring in which every node and its corresponding location are originated from either of the parent chromosomes [33]. We choose the same example chromosomes P_1: (1, 5, 4, 7, 8, 2, 3, 6) and P_2: (1, 8, 3, 4, 5, 6, 2, 7).

The first gene is 1 and for the 2nd position, we choose randomly either 5 or 8. Suppose we choose node 5, then the offspring chromosome becomes:

O_1: (1, 5, *, *, *, *, *, *)

All genes in the offspring is chosen from either of the parents in the same location, so the next gene to should be 8, which is located in P_2 just below the present node 5. In P_1, this node 8 is located at 5th position; so, the offspring chromosome becomes:

O_1: (1, 5, *, *, 8, *, *, *)

Since, next node to be selected is 5 that is already available in O_1; thus, a cycle is completed and so, the remaining blank locations will be filled up by the genes of those locations that are present in P_2. This way the offspring is built as follows:

O_1: (1, 5, 3, 4, 8, 6, 2, 7) with cost 1.

Similarly, the 2nd offspring is created:

O_2: (1, 8, 4, 7, 5, 2, 3, 6) with cost 2.

5) Edge recombination crossover operator. The edge recombination crossover (ERX) is proposed in [34]. Most operators consider the position and the order of the node. This operator considers the links between these nodes. To apply this operator, we first construct the edge list of the parents P_1: (1, 5, 4, 7, 8, 2, 3, 6) and P_2: (1, 8, 3, 4, 5, 6, 2, 7).

Table II shows the edge list of all the nodes for the given example. Since the 1st gene of the offspring is 1, the nodes 6, 5, 7 and 8 are the candidates for the next gene. The nodes 6, 7 and 8 have three edges: initially four node minus the present node 1. Similarly, the node 5 has two edges. Among them, node 5 has minimum edges, thus it is chosen, and the offspring becomes (1, 5).

Node 5 has edges to nodes 4 and 6, so node 4 is chosen randomly as both have equal two edges, and the offspring becomes (1, 5, 4).

Node 4 has edges to nodes 7 and 3. Nodes 7 and 3 have 2 and 3 edges. So, node 7 is chosen next and the offspring becomes (1, 5, 4, 7).

Node 7 has edges to nodes 8 and 2. Nodes 8 and 2 have 2 and 3 edges. So, node 8 is chosen next and the offspring becomes (1, 5, 4, 7, 8).

Node 8 has edges to nodes 2 and 3. Both nodes have 2 edges. So, node 2 is chosen randomly and the offspring becomes (1, 5, 4, 7, 8, 2).

Node 2 has edges to nodes 3 and 6. Both nodes have 1 edge. So, node 3 is chosen randomly and the offspring becomes (1, 5, 4, 7, 8, 2, 3). This way the final offspring is created as: (1, 5, 4, 7, 8, 2, 3, 6) with cost 3. Here all edges are chosen from either of the parents.

6) Generalized n-point crossover operator. Radcliffe and Surry [35] developed generalized N crossover (GNX). Suppose N=2, and P_1: (1, 5, 4, 7, 8, 2, 3, 6) and P_2: (1, 8, 3, 4, 5, 6, 2, 7). Now, if crossover points are 4 and 6, then the bold face nodes would usually be selected by G2X. Suppose the segments are tested in the order (3, 2, 1). Then the 3rd segment of random parent, suppose of P_2, will be added to give the proto child (*, *, *, *, *, *, 2, 7). Next, the nodes in 2nd segment from P_1 will be tested in random order. The node 8 is accepted to give the proto child (*, *, *, *, 8, *, 2, 7). Then nodes in 1st segment from P_2 is tested, and nodes 1, 3 and 4 are accepted to give the final proto child after the 1st phase: (1, *, 3, 4, 8, *, 2, 7).

Now, the untested segments of both parents are the tested in arbitrary order. Only the 2nd segment for P_2 is relevant here and node 6 is accepted. So, the proto child after the 2nd phase is (1, *, 3, 4, 8, 6, 2, 7).

Since this offspring is yet incomplete, we fill it up randomly. So, the final offspring may be (1, 5, 3, 4, 8, 6, 2, 7) with cost 1. Only four edges are chosen from either of the parents.

TABLE II. The Edge List of the Nodes for the Parents P_1 and P_2

Node	**Edge list**	**Node**	**Edge list**
1	6, 5, 7, 8	**5**	1, 4, 6
2	8, 3, 6, 7	**6**	3, 1, 5, 2
3	2, 6, 8, 4	**7**	4, 8, 2, 1

4	5, 7, 3	**8**	7, 2, 1, 3

7) Greedy crossover operator. The greedy crossover (GX) selects the first node randomly [32]. Since the MSTSP is a maximization problem, hence some steps of the GX must be modified. So, our modified GX for the problem is as follows. In each step, total four neighbor nodes of the present node are considered from the parents, and the (unvisited) node having the largest cost is selected, because it is best at present. If either this best node or all neighbour nodes are available in the offspring, then any other unvisited node is chosen randomly. GX produces one offspring only from the parents. We consider the same chromosomes P_1: (1, 5, 4, 7, 8, 2, 3, 6) and P_2: (1, 8, 3, 4, 5, 6, 2, 7).

We have the initial offspring (1). The nodes 5 and 8 are neighbour nodes of node 1 with their costs 66 and 2 respectively. Having higher cos, the node 5 is better, so, it is added to the offspring: (1, 5).

The nodes 4, 1, 6 and 4 are neighbour nodes of node 5 with their costs 31, 60, 50 and 31 respectively. Though the node 1 is the best, since it is available in the offspring, node 2 is chosen randomly and added to the offspring: (1, 5, 2).

The nodes 3, 8, 7 and 6 are neighbour nodes of node 2 with their costs 55, 58, 1 and 72 respectively. Node 6 is added to the offspring, as it the best node: (1, 5, 2, 6).

The nodes 3, 2 and 5 are neighbour nodes of node 6 with their costs 20, 82 and 39 respectively. Though node 2 is the best, since it is available in the offspring, node 3 is chosen randomly and added to the offspring: (1, 5, 2, 6, 3). Finally, the complete offspring may be: (1, 5, 2, 6, 3, 4, 7, 8) with cost 13.

8) Sequential constructive crossover operator. The sequential constructive crossover (SCX) operator creates only one offspring by using better arcs available in the parents' structure ([16], [36]). Additionally, sometimes it uses better arcs those are not available in either of the parents' structure. It sequentially searches both parent chromosomes and selects first legitimate (unvisited) node that appears after the present node. If no any legitimate node is available in either of the parents, it sequentially searches from the beginning of the chromosome and then compares their associated cost to decide the next node of the offspring chromosome. This operator is found to be very effective for the TSP and some other problems ([37]-[40]). The SCX is slightly modified for the MSTSP as below:

Step 1: Start from 'node 1' (i.e., current node p =1).

Step 2: Sequentially search both parent chromosomes and consider the first 'legitimate node' (the node that is not yet visited) appeared after 'node p' in each parent. If no 'legitimate node' after 'node p' is present in any of the parents, search sequentially from the starting of the parent and consider the first 'legitimate node', and go to Step 3.

Step 3: Suppose the 'node α' and the 'node β' are found in 1st and 2nd parent respectively, then for selecting the next node go to Step 4.

Step 4: If $c_{p\alpha} > c_{p\beta}$, then select 'node α', otherwise, 'node β' as the next node and concatenate it to the partially constructed offspring chromosome. If the offspring is a complete chromosome, then stop, otherwise, rename the present node as 'node p' and go to Step 2.

We consider the same example P_1: (1, 5, 4, 7, 8, 2, 3, 6) and P_2: (1, 8, 3, 4, 5, 6, 2, 7). Node 1 is the 1st gene. After node 1, nodes 5 in P_1 and 8 in P_2 are legitimate nodes with costs c_{15}=66 and c_{18}=2. Since $c_{15}>c_{18}$, node 5 is accepted and the offspring becomes (1, 5). =

After node 5, nodes 4 in P_1 and 8 in P_2 are legitimate nodes with costs c_{54}=31 and c_{56}=50. Since $c_{56}>c_{54}$, node 6 is accepted and the offspring becomes (1, 5, 6).

After node 6, nodes 4 in P_1 and 2 in P_2 are legitimate nodes with costs c_{64}=70 and c_{62}=82. Since $c_{62}>c_{64}$, node 2 is accepted and the offspring becomes (1, 5, 6, 2).

After node 2, nodes 3 in P_1 and 7 in P_2 are legitimate nodes with costs c_{23}=55 and c_{27}=1. Since $c_{23}>c_{27}$, node 3 is accepted and the offspring becomes (1, 5, 6, 2, 3).

After node 3, there is no legitimate node in P_1 and node 4 is legitimate in P_2. So, for P_1, search continues from its starting and finds same legitimate node 4 with c_{34}=69. So, node 4 is accepted and the offspring becomes (1, 5, 6, 2, 3, 4). Finally, offspring (1, 5, 6, 2, 3, 4, 7, 8) with cost 13 is obtained.

D. Mutation Operator

Mutation operator increases variety in the population by applying random changes in the population. For example, swap mutation, inversion mutation, insertion mutation, adaptive mutation [14], etc. are some of them. We have implemented swap mutation for our simple GAs.

E. Control Parameters

Control parameters rule the genetic process at some extent. They are - population size that decides number of chromosomes available during the process, crossover probability that fixes the probability of performing crossover between parents, mutation probability that fixes the probability of performing gene-wise mutation and stopping criterion that fixes when to stop the genetic process [16]. A simple GA may be summarized as follows:

```
SimpleGA( )
{ Initialize population randomly;
Evaluate the population;
Generation = 0;
While stopping criterion is not satisfied
{ Generation = Generation + 1;
Select better chromosomes by selection operator;
Perform crossover using crossover probability ($P_c$);
Perform mutation using mutation probability ($P_m$);
Evaluate the population;
 }
}
```

IV. C COMPUTATIONAL EXPERIENCES AND DISCUSSIONS

To perform compare study among eight different crossover operators, simple GAs using these crossover operators have been encoded in Visual C++ on a Laptop with i7-1065G7 CPU@1.30 GHz and 8 GB RAM under MS Windows 10, and then run for twenty TSPLIB instances [41]. Out of the twenty, the nine instances ftv33, ftv38, ftv44, ft53, ftv64, ft70, ftv70, kro124p and ftv170 are asymmetric, and the remaining eleven instances dantzig42, eil51, st70, lin105, ch130, kroA150, si175, d198, pr226, a280 and lin318 are symmetric. We run GAs for different setting of parameters, and selected parameters are listed in Table III.

Fig. 1 presents results for ftv170 (by considering only 100 generations) by all GAs. Each curve is for one crossover, and it shows improvement of current solution in the successive generations. The figure shows some variations of SCX and shows that SCX is the best. ERX also has some variations and is place in second position. But GX and AEX have no variations and get trapped in local maximum very quickly and shown to be the worst.

The comparative study among the eight simple GAs are summarized in two tables: Tables IV and VIII. These tables are prepared similarly: each row is for an instance and each column is for one GA using a particular crossover operator. The result is defined best solution cost, average solution cost, standard deviation (S.D.) of solution costs, and average convergence time (in second). The best result for a particular instance among all GAs is marked by bold face.

TABLE III. PARAMETERS FOR THE GAS

Parameters	**Values**
Population size	50
Crossover probability	100%
Mutation probability	10%
Termination criterion	1,000 generations
No. of runs for each instance	50 times

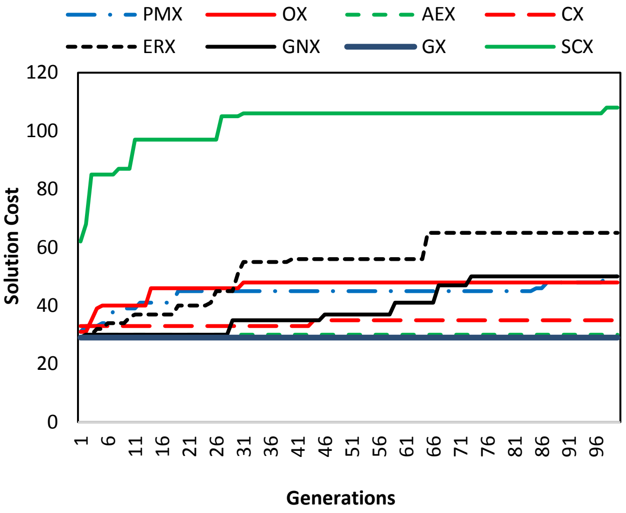


Fig. 1. Result by GAs using different Crossover Operators for ftv170.

From the Table IV, it is seen that the crossovers OX, AEX, CX, ERX and GX could not obtain either best solution or best average cost for any asymmetric instance. The crossover PMX obtains best average costs with lowest S.D. for the instances ftv33, ftv38 and ftv44, whereas SCX obtains best lowest average costs with lowest S.D. for the remaining six instances. So, SCX is shown to be the best. These results are shown in Fig. 2 that also shows the usefulness of crossover SCX. The crossovers ERX and GNX are competing, and GX is the worst.

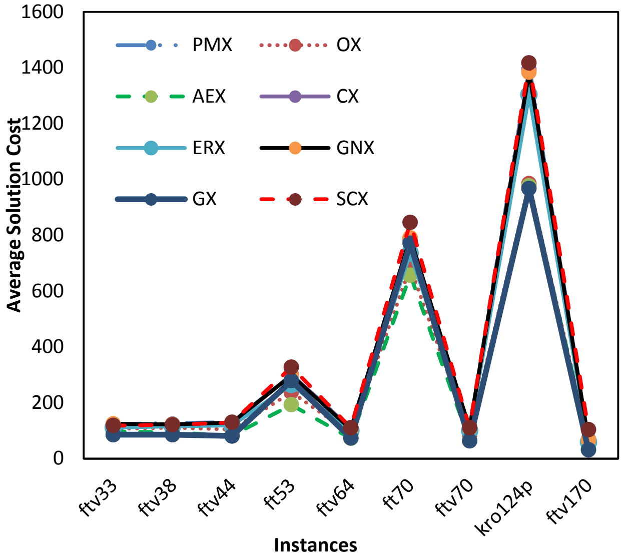


Fig. 2. Average Solution Cost by different GAs for Asymmetric Instances.

To confirm whether SCX-based GA average is statistically and significantly different from the averages found by other crossover-based GAs, Student's t-test is performed. It is to be mentioned that 50 runs have been performed for each instance. Following t-test formula is used here [42]:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{SD_1^2}{n_1 - 1} + \frac{SD_2^2}{n_2 - 1}}}$$

where,

$\bar{X}_1 - average\ of\ first\ sample,$

$SD_1 - standard\ deviation\ of\ first\ sample,$

$\bar{X}_2 - average\ of\ second\ sample,$

$SD_2 - standard\ deviation\ of\ second\ sample,$

$n_1 - first\ sample\ size,$

$n_2 - second\ sample\ size,$

The values of $\bar{X}_2$ and SD_2 are found by the SCX-based GA, and $\bar{X}_1$ and SD_1 values are found by other GAs. The t-statistic are reported in Table V. The t-values may be positive or negative. Since the problem is maximization problem, the negative value shows that SCX found better solution than the competitive crossover. In the positive case, the competitive crossover found better solution. The confidence interval at the

95% confidence level ($t_{0.05}$ = 1.96) is used. When t-value is greater than 1.96, the difference between them is significant. In this condition, if t-vale is negative then SCX-based GA solution is better, otherwise the competitive crossover-based GA solution is better. If t-value is less than 1.96, then there is no significant difference between the obtained values. The table also shows the information about the crossovers that found significantly better solutions.

On four instances there is no statistically significant difference between SCX and PMX. On three instances SCX is found better than PMX, whereas, PMX is found better than SCX on two instance. There is no significant difference between SCX and CX on two instances. On five instances SCX performed better than CX, whereas, on two instances CX is better than SCX. Next, there is no significant difference between SCX and GNX on three instances. SCX is better than GNX on five instances, whereas, GNX is better than SCX on only one instance. On all nine instances, SCX is found better than OX, AEX and ERX. From this study we can say that SCX is the best for asymmetric instances.

TABLE IV. COMPARATIVE STUDY OF 8 CROSSOVER-BASED GAS FOR ASYMMETRIC TSPLIB INSTANCES

Instance	**n**	**Result**	**PMX**	**OX**	**AEX**	**CX**	**ERX**	**GNX**	**GX**	**SCX**
ftv33	34	Best Sol	**134**	122	107	133	122	133	99	125
		Avg. Sol	**123.25**	106.25	96.8	121	111.4	124.4	85.5	118.3
		S.D.	6.72	8.1	11.25	6.16	5.57	5.64	11.15	5.04
		Avg. Time	0.02	0.04	0.07	0.06	0.85	0.04	0.08	0.07
ftv38	39	Best Sol	**137**	121	114	135	125	136	103	133
		Avg. Sol	**126.70**	110.70	94.35	123.35	115.35	122.30	85.45	121.10
		S.D.	6.07	5.28	8.79	4.64	7.12	8.49	11.17	5.80
		Avg. Time	0.03	0.05	0.07	0.08	1.14	0.04	0.11	0.11
ftv44	45	Best Sol	140	125	94	142	130	**143**	99	137
		Avg. Sol	**130.6**	104.7	82.45	125.40	118.80	129.00	81.20	129.85
		S.D.	6.56	8.68	8.87	8.00	5.48	7.85	10.07	5.34
		Avg. Time	0.03	0.05	0.08	0.10	1.44	0.05	0.14	0.14
		Best Sol	329	275	223	321	286	327	323	**360**
ft53	53	Avg. Sol	299.35	237.50	192.95	300.30	265.30	297.90	279.00	**327.80**
		S.D.	15.07	12.66	12.51	15.46	11.59	16.16	31.84	10.98
		Avg. Time	0.04	0.05	0.10	0.13	1.99	0.06	0.22	0.22
ftv64	65	Best Sol	**123**	106	85	122	110	118	88	120
		Avg. Sol	110.7	90.25	72.60	109.30	99.70	105.90	73.90	**110.90**
		S.D.	7.88	7.37	5.99	7.31	5.10	7.74	7.44	5.76
		Avg. Time	0.05	0.10	0.15	0.19	2.87	0.06	0.28	0.31
ft70	70	Best Sol	816	707	673	823	768	822	816	**884**
		Avg. Sol	778.95	685.95	656.60	785.3	735.35	791.05	770.85	**845.85**
		S.D.	25.70	7.67	11.01	17.57	24.53	23.2	26.78	17.90
		Avg. Time	0.06	0.07	0.17	0.22	3.45	0.07	0.30	0.32
ftv70	71	Best Sol	121	111	85	118	108	**125**	81	**125**
		Avg. Sol	109.95	87	63.75	106.95	99.95	108.1	63.1	**110.70**
		S.D.	6.57	6.57	8.22	5.42	5.17	5.31	6.39	6.23
		Avg. Time	0.06	0.09	0.15	0.19	3.49	0.07	0.25	0.29
kro124p	100	Best Sol	1562	1083	1097	1486	1498	**1666**	1069	1553
		Avg. Sol	1406.50	984.35	976.30	1392.30	1302.80	1383.70	966.90	**1416.50**
		S.D.	82.81	43.62	50.12	51.6	83.35	95.47	52.42	81.07
		Avg. Time	0.09	0.14	0.29	0.42	6.98	0.11	0.54	0.68
ftv170	171	Best Sol	71	70	78	71	63	73	37	**112**
		Avg. Sol	63.90	62.40	59.95	67.95	59.35	65.60	31.20	**104.15**
		S.D.	2.74	3.43	16.02	1.83	1.80	2.62	2.68	4.09
		Avg. Time	0.16	0.31	0.28	0.66	13.81	0.11	0.06	1.36

TABLE V. THE T-VALUES AGAINST SCX AND THE INFORMATION ABOUT CROSSOVERS THAT FOUND SIGNIFICANTLY BETTER SOLUTIONS

Instance	**PMX**	**OX**	**AEX**	**CX**	**ERX**	**GNX**	**GX**
ftv33	4.13	-8.84	-12.21	2.37	-6.43	5.65	-18.76
Better	PMX	SCX	SCX	CX	SCX	GNX	SCX
ftv38	4.67	-9.28	-17.78	2.12	-4.38	0.82	-19.83
Better	PMX	SCX	SCX	CX	SCX	---	SCX
ftv44	0.62	-17.27	-32.05	-3.24	-10.11	-0.63	-29.88
Better	---	SCX	SCX	SCX	SCX	---	SCX
ft53	-10.68	-37.72	-56.71	-10.15	-27.40	-10.71	-10.14
Better	SCX	SCX	SCX	SCX	SCX	SCX	SCX
ftv64	-0.14	-15.45	-32.26	-1.20	-10.19	-3.63	-27.53
Better	---	SCX	SCX	---	SCX	SCX	SCX
ft70	-14.95	-57.48	-63.04	-16.90	-25.47	-13.09	-16.30
Better	SCX	SCX	SCX	SCX	SCX	SCX	SCX
ftv70	-0.58	-18.32	-31.86	-3.18	-9.29	-2.22	-37.34
Better	---	SCX	SCX	SCX	SCX	SCX	SCX
kro124p	-0.60	-32.86	-32.33	-1.76	-6.85	-1.83	-32.60
Better	---	SCX	SCX	---	SCX	---	SCX
ftv170	-57.23	-54.75	-18.71	-56.55	-70.18	-55.56	-104.43
Better	SCX	SCX	SCX	SCX	SCX	SCX	SCX

To rank the other crossover operators, the t-values against PMX is calculated and reported in Table VI. There is no significant difference found between PMX and GNX on five instances. Each of them performed better than the other one on two instances. There is no significant difference found between PMX and CX on five instances. On three instances PMX is found better than CX, whereas, CX is found better than PMX on one instance. It shows that PMX and GNX are sharing 2nd rank. We further carried out an adequate statistical analysis. The results of our hypotheses testing are summarized in Table VII. In the table, each row contains two columns, where the first lists a crossover operator and the second column lists its inferior crossover operators. Each crossover is ranked according to its number of inferior crossover operators. No significant difference is found between AEX and GX, and hence, they share the worst rank.

From the Table VIII, it is seen that the crossovers OX, AEX, CX and GX could not obtain either best solution or best average cost for any asymmetric instance. The crossover PMX and ERX obtain best average costs with lowest S.D. for the instances eil51 and dantzig42 respectively, whereas SCX obtains best lowest average costs with lowest S.D. for the remaining nine instances. So, the crossover SCX is found to be the best.

The results are shown in Fig. 3 that also shows the usefulness of SCX. The crossovers ERX, OX, CX and GNX are competing, and GX is the worst. Based on this study also one can say that SCX is the best and GX is the worst, and others are competing.

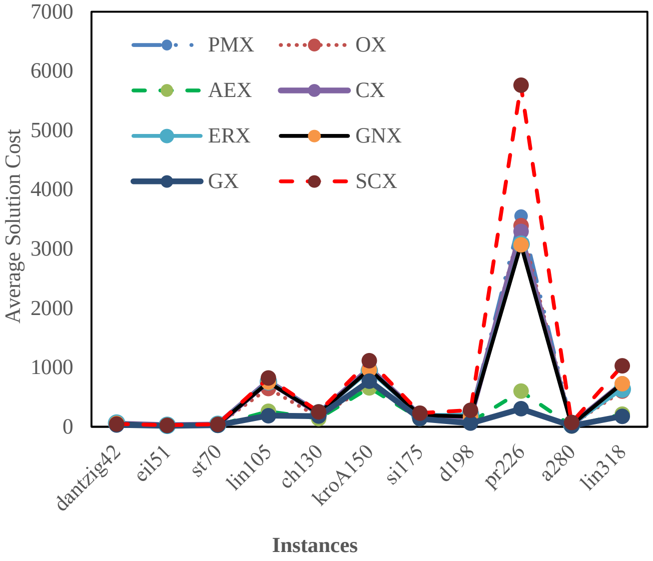


Fig. 3. Average Solution Cost by different GAs for Symmetric Instances.

For these symmetric instances also, to confirm whether SCX-based GA average solution is significantly different from the average solution found by other GAs, Student's t-test is performed, and the calculated t-values are reported in the Table IX.

On two instances there is no statistically significant difference between SCX and ERX. On eight instances SCX is better than ERX, whereas, ERX is better than SCX on one instance only. On one instance, there is no significant difference between SCX and (PMX, OX, CX and GNX). SCX performed better than PMX, CX and GNX on nine instances, whereas, PMX, CX and GNX are better than SCX on only one instance. Next, on one instance there is no statistically significant difference between SCX and OX. On remaining ten instances SCX is better than OX. From this study we can conclude that SCX is the best. However, to rank the other crossover operators, an adequate statistical analysis is carried out, and the results are summarized in Table X. The crossovers PMX, ERX, GNX, CX, OX, AEX and GX are placed in the 2nd, 3rd, 4th, 5th, 6th, 7th and worst rank, respectively. On both kind of problem instance, SCX is placed the 1st rank, PMX is in the 2nd rank and GX is in the worst rank.

TABLE VI. The T-Values Against PMX and the Information about Crossovers that Found Significantly better Solutions

Instance	**OX**	**AEX**	**CX**	**ERX**	**GNX**	**GX**
ftv33	-11.31	-14.13	-1.73	-9.50	0.92	-20.30
Better	PMX	PMX	----	PMX	---	PMX
ftv38	-13.92	-21.20	-3.07	-8.49	-2.95	-22.71
Better	PMX	PMX	PMX	PMX	PMX	PMX
ftv44	-16.66	-30.55	-3.52	-9.66	-1.09	-28.77
Better	PMX	PMX	PMX	PMX	---	PMX
ft53	-22.00	-38.03	0.31	-12.54	-0.46	-4.04
Better	PMX	PMX	---	PMX	---	PMX
ftv64	-13.27	-26.94	-0.91	-8.20	-3.04	-23.77
Better	PMX	PMX	---	PMX	PMX	PMX
ft70	-24.27	-30.63	1.43	-8.59	2.45	-1.53
Better	PMX	PMX	---	PMX	**GNX**	---
ftv70	-17.29	-30.73	-2.47	-8.37	-1.53	-35.78
Better	PMX	PMX	PMX	PMX	---	PMX
kro124p	-31.57	-31.11	-1.02	-6.18	-1.26	-31.40
Better	PMX	PMX	---	PMX	---	PMX
ftv170	-2.39	-1.70	8.60	-9.72	3.14	-59.72
Better	PMX	---	**CX**	PMX	**GNX**	PMX

TABLE VII. Results of Statistical Hypotheses Testing on Asymmetric Instances

Crossover	**Inferior crossovers**
SCX	PMX, OX, AEX, CX, ERX, GNX, GX
PMX	OX, AEX, CX, ERX, GX
GNX	OX, AEX, CX, ERX, GX
CX	OX, AEX, ERX, GX
ERX	OX, AEX, GX
OX	AEX, GX
AEX	----
GX	----

TABLE VIII. COMPARATIVE STUDY OF 8 CROSSOVER-BASED GAS FOR SYMMETRIC TSPLIB INSTANCES

Instance	n	Results	PMX	OX	AEX	CX	ERX	GNX	GX	SCX
dantzig42	42	Best Sol	65	54	50	62	**69**	62	41	52
		Avg. Sol	57.80	48.75	39.50	53.85	**61.30**	55.50	31.70	49.05
		S.D.	3.78	2.66	4.02	4.29	3.27	4.63	4.28	1.28
		Avg. Time	0.03	0.04	0.07	0.09	1.32	0.04	0.10	0.13
eil51	51	Best Sol	29	25	24	30	29	**31**	21	30
		Avg. Sol	**25.75**	22.10	20.90	25.25	26.00	25.60	18.40	25.60
		S.D.	2.02	1.26	2.12	1.89	1.26	2.40	1.62	1.88
		Avg. Time	0.04	0.05	0.08	0.10	1.47	0.05	0.17	0.13
st70	70	Best Sol	**48**	42	30	44	**48**	45	33	**48**
		Avg. Sol	41.00	32.70	25.55	37.20	41.90	38.55	27.00	**43.15**
		S.D.	3.91	2.81	2.60	2.62	3.69	3.38	2.68	2.65
		Avg. Time	0.06	0.07	0.15	0.23	3.29	0.07	0.29	0.37
lin105	105	Best Sol	906	736	385	850	857	832	282	**914**
		Avg. Sol	768.40	645.80	261.85	765.45	760.95	751.05	189.45	**822.30**
		S.D.	62.20	59.35	66.69	46.63	51.11	45.59	29.66	57.26
		Avg. Time	0.09	0.18	0.10	0.45	8.20	0.10	0.70	0.75
ch130	130	Best Sol	265	173	163	269	253	250	208	**273**
		Avg. Sol	236.00	160.35	140.80	228.60	234.40	227.65	180.35	**251.90**
		S.D.	21.79	7.09	11.45	19.10	13.23	13.20	19.18	12.94
		Avg. Time	0.11	0.20	0.44	0.69	12.10	0.13	1.22	1.19
kroA150	150	Best Sol	1147	791	719	1142	1053	1094	899	**1238**
		Avg. Sol	997.85	706.20	657.65	983.45	944.85	966.10	771.50	**1113.20**
		S.D.	73.05	40.69	39.32	65.35	56.88	65.94	71.82	67.41
		Avg. Time	0.13	0.29	0.51	0.94	16.64	0.15	1.42	1.45
si175	175	Best Sol	211	193	149	211	208	211	149	**248**
		Avg. Sol	196.15	186.60	136.85	189.95	197.10	195.20	136.85	**231.40**
		S.D.	9.93	3.44	5.46	11.74	7.39	10.49	5.46	10.73
		Avg. Time	0.13	0.31	0.14	0.81	15.24	0.13	0.09	1.81
d198	198	Best Sol	265	198	114	218	234	217	92	**309**
		Avg. Sol	191.70	168.15	75.10	165.75	207.70	173.15	62.50	**279.10**
		S.D.	39.29	13.54	16.60	24.58	17.31	28.52	11.77	21.89
		Avg. Time	0.15	0.36	0.06	1.50	23.23	0.16	0.16	1.90
pr226	226	Best Sol	4887	3640	1650	4014	3796	3790	522	**6540**
		Avg. Sol	3555.40	3394.15	600.75	3295.50	3080.35	3069.40	304.30	**5761.80**
		S.D.	595.65	133.00	352.07	389.34	427.27	588.77	83.88	319.06
		Avg. Time	0.18	0.73	0.29	1.89	33.15	0.21	0.25	1.89
a280	280	Best Sol	54	40	23	49	45	45	18	**88**
		Avg. Sol	39.50	34.50	16.50	32.65	35.70	32.85	15.25	**72.90**
		S.D.	6.84	2.01	2.67	7.70	4.06	6.51	1.95	5.44
		Avg. Time	0.20	0.63	0.13	2.35	34.78	0.19	0.09	3.70
lin318	318	Best Sol	860	721	346	860	850	813	202	**1151**
		Avg. Sol	742.15	602.8	213.5	723.6	635.75	726.6	176.3	**1027.6**
		S.D.	75.84	57.18	51.71	81.4	64.36	48.88	12.08	72.64
		Avg. Time	0.26	0.85	0.3	3.59	59.5	0.29	0.3	4.73

TABLE IX. THE T-VALUES AGAINST SCX AND THE INFORMATION ABOUT CROSSOVERS THAT FOUND SIGNIFICANTLY BETTER SOLUTIONS

Instance	**PMX**	**OX**	**AEX**	**CX**	**ERX**	**GNX**	**GX**
dantzig42	15.35	-0.71	-15.85	7.51	24.42	9.40	-27.19
Better	**PMX**	---	SCX	**CX**	**ERX**	**GNX**	SCX
eil51	0.38	-10.83	-11.61	-0.92	1.24	0.00	-20.31
Better	---	SCX	SCX	---	---	---	SCX
st70	-3.19	-18.94	-33.19	-11.18	-1.93	-7.50	-30.00
Better	SCX	SCX	SCX	SCX	---	SCX	SCX
lin105	-4.46	-14.98	-44.63	-5.39	-5.60	-6.81	-68.70
Better	SCX	SCX	SCX	SCX	SCX	SCX	SCX
ch130	-4.39	-43.43	-45.01	-7.07	-6.62	-9.18	-21.65
Better	SCX	SCX	SCX	SCX	SCX	SCX	SCX
kroA150	-8.12	-36.18	-40.86	-9.67	-13.36	-10.92	-24.28
Better	SCX	SCX	SCX	SCX	SCX	SCX	SCX
si175	-16.88	-27.83	-54.97	-18.24	-18.43	-16.89	-54.97
Better	SCX	SCX	SCX	SCX	SCX	SCX	SCX
d198	-13.60	-30.17	-51.98	-24.11	-17.91	-20.63	-61.01
Better	SCX	SCX	SCX	SCX	SCX	SCX	SCX
pr226	-22.86	-47.95	-76.04	-34.30	-35.20	-28.14	-115.80
Better	SCX	SCX	SCX	SCX	SCX	SCX	SCX
a280	-26.75	-46.35	-65.15	-29.88	-38.36	-33.05	-69.83
Better	SCX	SCX	SCX	SCX	SCX	SCX	SCX
lin318	-19.03	-32.17	-63.91	-19.51	-28.26	-24.06	-80.92
Better	SCX	SCX	SCX	SCX	SCX	SCX	SCX

TABLE X. RESULTS OF STATISTICAL HYPOTHESES TESTING ON SYMMETRIC INSTANCES

Crossover	**Inferior crossovers**
SCX	PMX, OX, AEX, CX, ERX, GNX, GX
PMX	OX, AEX, CX, ERX, GNX, GX
ERX	OX, AEX, CX, GNX, GX
GNX	OX, AEX, CX, GX
CX	OX, AEX, GX
OX	AEX, GX
AEX	GX

V. CONCLUSION AND FUTURE WORKS

Numerous crossover operators have been proposed for the TSP using GAs which can also be used for its variations. In this paper, eight simple GAs using eight different crossover operators, namely PMX, OX, AEX, CX, ERX, GNX, GX and SCX, have been developed for solving the MSTSP. We first applied these operators in manual experiment on two parent chromosomes to produce an offspring, for each crossover operator. We then run the algorithms run on TSPLIB instances of different types and sizes. We set highest crossover probability to show exact nature of crossover operators. We carried out comparative study of the GAs on nine asymmetric and eleven symmetric TSPLIB instances. In terms of solution quality, our comparative study showed that crossover operator SCX is the best, PMX is the second-best and GX is the worst. Our observation is confirmed using Student's t-test at 95% confidence level. Thus, SCX may be good crossover operator to obtain more accurate results, researchers may apply it for other related combinatorial optimization problems. However, it is seen that PMX is better than SCX for small-sized instances.

In this study, our aim was to compare the solution quality found using different crossover operators, neither to improve the solution quality nor to develop the most competitive algorithm for the MSTSP. So, neither any local search technique is used to improve the solution quality nor parallel version of algorithms is developed to find exact solution. Therefore, we have developed simple and pure GAs. Thus, modified SCX operators ([43]-[45]) can be used instead of SCX and then good local search and immigration procedures [46] can be incorporated to hybridize the algorithm to solve the instances more accurately, which is under our investigation.

ACKNOWLEDGMENT

The author is very much thankful to the honourable anonymous reviewers for their constructive comments and suggestions which helped the author to improve this paper. This research was supported by Deanery of Academic Research, Al-Imam Mohammad Ibn Saud Islamic University, Saudi Arabia vide Grant No. 18-11-09-010. The author is also thankful to the Deanery for its financial support.

REFERENCES

[1] E.M. Arkin, Y.-J. Chiang, J.S.B. Mitchell, S.S. Skiena, and T.-C. Yang, “On the maximum scatter traveling salesperson problem,” SIAM Journal of Computing, vol. 29, pp. 515–544, 1999.

[2] Z.H. Ahmed, “A hybrid genetic algorithm for the bottleneck traveling salesman problem,” ACM Transactions on Embedded Computing Systems, vol. 12, Art. No. 9, 2013.

[3] F. Scholz, “Coordination hole tolerance stacking,” Technical Report BCSTECH-93-048, Boeing Computer Services, November 1993.

[4] L.R. John, “The bottleneck traveling salesman problem and some variants,” Master of Science of Simon Fraser University, Canada, 2010.

[5] J. LaRusic and A.P. Punnen, “The asymmetric bottleneck traveling salesman problem: Algorithms, complexity and empirical analysis,” Computers & Operations Research, vol. 43, pp. 20–35, 2014.

[6] J.D.C. Little, K.G. Murthy, D.W. Sweeny, and C. Kare, “An algorithm for the travelling salesman problem,” Operations Research, vol. 11, pp. 972-989, 1963.

[7] S.N.N. Pandit, “The Loading Problem,” Operations Research, vol. 11, pp. 639-646, 1962.

[8] D. Applegate, R.E. Bixby, V. Chv´atal and W. Cook, “On the solution of traveling salesman problems,” Documenta Mathematica, Extra Vol. ICM III, pp. 645-656, 1998.

[9] D.S Johnson and L.A. McGeoch, “The traveling salesman problem: a case study,” in E. Aarts, J.K. Lenstra, eds. Local Search in Combinatorial Optimization. Wiley, Chichester, UK. Pp. 215-310, 1997.

[10] J.W. Ohlmann and B.W. Thomas, “A compressed-annealing heuristic for the traveling salesman problem with time windows,” INFORMS Journal of Computing, vol. 19, no. 1, pp. 80–90, 2007.

[11] W.B. Carlton and J.W. Barnes, “Solving the travelling salesman problem with time windows using tabu search,” IEE Transaction, vol. 28, pp. 617–629, 1996.

[12] M. Gendreau, A. Hertz, G. Laporte and M. Stan, “A generalized insertion heuristic for the traveling salesman problem with time windows,” Operations Research, vol. 46, no. 3, pp. 330–335, 1998.

[13] C.-B. Cheng and C.-P. Mao, “A modified ant colony system for solving the travelling salesman problem with time windows,” Mathematical Computer Modelling, vol. 46, pp. 1225–1235, 2007.

[14] D.E. Goldberg, “Genetic algorithms in search, optimization, and machine learning,” Addison-Wesley, New York, 1989.

[15] R.F. da Silva and S. Urrutia, “A general VNS heuristic for the traveling salesman problem with time windows,” Discrete Optimization, vol. 7, no. 4, pp. 203–211, 2010.

[16] Z.H. Ahmed, “Genetic algorithm for the traveling salesman problem using sequential constructive crossover operator,” International Journal of Biometrics & Bioinformatics, vol. 3, pp. 96-105, 2010.

[17] Yi-J. Chiang, “New approximation results for the maximum scatter TSP,” Algorithmica, vol. 41, pp. 309–341, 2005.

[18] S.N. Kabadi and A.P. Punnen, “The bottleneck TSP,” In The Traveling Salesman Problem and Its Variations, G. Gutin and A.P. Punnen (eds.), Chapter 15, Kluwer Academic, Dordrecht, 2002.

[19] I. Hoffmann, S. Kurz, and J. Rambau, “The maximum scatter TSP on a regular grid,” in Operations Research Proceedings 2015, Springer, 2015, pp. 63–70.

[20] G. Gutin and A.P. Punnen (eds.), “The Traveling Salesman Problem and Its Variations,” Kluwer Academic, Dordrecht, 2002.

[21] Z.H. Ahmed, “A lexisearch algorithm for the bottleneck travelling salesman problem,” International Journal of Computer Science and Security, vol. 3, no. 5, pp. 569-577, 2010.

[22] Z.H. Ahmed, “A data-guided lexisearch algorithm for the bottleneck travelling salesman problem,” International Journal of Operational Research, vol. 12, no. 1, pp. 20-33, 2011.

[23] Z.H. Ahmed, “A hybrid sequential constructive sampling algorithm for the bottleneck traveling salesman problem,” International Journal of Computational Intelligence Research, vol. 6, no. 3, pp. 475-484, 2010.

[24] Z.H. Ahmed, “A hybrid genetic algorithm for the bottleneck traveling salesman problem,” ACM Transactions on Embedded Computing Systems, vol. 12, Art. No. 9, 2013.

[25] A. Barvinok, S.P. Fekete, D.S. Johnson, A. Tamir, G.J. Woeginger and R. Woodroofe, “The geometric maximum traveling salesman problem,” Journal of the ACM, vol. 50, no. 5, pp. 641–664, 2003.

[26] Z.H. Ahmed, “An experimental study of a hybrid genetic algorithm for the maximum travelling salesman problem,” Mathematical Sciences, vol. 7, pp. 1-7, 2013.

[27] W. Dong, X. Dong and Y. Wang, “The improved genetic algorithms for multiple maximum scatter traveling salesperson problems,” In J. Li et al. (Eds.): CWSN 2017, CCIS 812, pp. 155–164, 2018.

[28] P. Venkatesh, A. Singh and R. Mallipeddi, “A multi-start iterated local search algorithm for the maximum scatter traveling salesman problem,” in 2019 IEEE Congress on Evolutionary Computation (CEC), Wellington, New Zealand, 2019, pp. 1390-1397.

[29] K. Deb, “Optimization for engineering design: algorithms and examples,” Prentice Hall of India Pvt. Ltd., New Delhi, India, 1995.

[30] D.E. Goldberg, and R. Lingle, “Alleles, loci and the travelling salesman problem,” In J.J. Grefenstette (ed.) Proceedings of the 1st International Conference on Genetic Algorithms and Their Applications. Lawrence Erlbaum Associates, Hilladale, NJ, 1985.

[31] L. Davis, “Job-shop scheduling with genetic algorithms,” Proceedings of an International Conference on Genetic Algorithms and Their Applications, pp. 136-140, 1985.

[32] J. Grefenstette, R. Gopal, B. Rosmaita, and D. Gucht, “Genetic algorithms for the traveling salesman problem,” In Proceedings of the First International Conference on Genetic Algorithms and Their Applications, (J. J. Grefenstette, Ed.), Lawrence Erlbaum Associates, Mahwah NJ, pp. 160–168, 1985.

[33] I.M. Oliver, D. J. Smith and J.R.C. Holland, “A study of permutation crossover operators on the travelling salesman problem,” In J.J. Grefenstette (ed.). Genetic Algorithms and Their Applications: Proceedings of the 2nd International Conference on Genetic Algorithms. Lawrence Erlbaum Associates, Hilladale, NJ, 1987.

[34] D. Whitley, T. Starkweather and D. Shaner, “The traveling salesman and sequence scheduling: quality solutions using genetic edge recombination,” In L. Davis (Ed.) Handbook of Genetic Algorithms. Van Nostrand Reinhold, New York, pp. 350-372, 1991.

[35] N.J. Radcliffe and P.D. Surry, “Formae and variance of fitness,” In D. Whitley and M. Vose (Eds.) Foundations of Genetic Algorithms 3, Morgan Kaufmann, San Mateo, CA, pp. 51-72, 1995.

[36] Z.H. Ahmed, “Improved genetic algorithms for the traveling salesman problem,” International Journal of Process Management and Benchmarking, vol. 4, no. 1, pp. 109-124, 2014.

[37] Z.H. Ahmed, “The ordered clustered travelling salesman problem: A hybrid genetic algorithm,” The Scientific World Journal, vol. 2014, Art ID 258207, 13 pages, 2014.

[38] Z.H. Ahmed, “A simple genetic algorithm using sequential constructive crossover for the quadratic assignment problem,” Journal of Scientific & Industrial Research, vol. 73, pp. 763-766, 2014.

[39] Z.H. Ahmed, “The minimum latency problem: a hybrid genetic algorithm,” IJCSNS International Journal of Computer Science and Network Security, vol. 18, no. 11, pp. 153-158, 2018.

[40] Z.H. Ahmed, “Performance analysis of hybrid genetic algorithms for the generalized assignment problem,” IJCSNS International Journal of Computer Science and Network Security, vol. 19, no. 9, pp. 216-222, 2019.

[41] G. Reinelt, TSPLIB, http://comopt.ifi.uni-heidelberg.de/ software/ TSPLIB95/

[42] M. Nikolić and D. Teodorović, "Empirical study of the bee colony optimization (BCO) algorithm," Expert Systems with Applications, vol. 40, pp. 4609–4620, 2013.

[43] Z.H. Ahmed, "Solving the traveling salesman problem using greedy sequential constructive crossover in a genetic algorithm," IJCSNS International Journal of Computer Science and Network Security, vol. 20, no. 2, pp. 99-112, 2020.

[44] Z.H. Ahmed, "Adaptive sequential constructive crossover operator in a genetic algorithm for solving the traveling salesman problem," IJACSA International Journal of Advanced Computer Science and Applications, vol. 11, no. 2, pp. 593-605, 2020.

[45] Z.H. Ahmed, "Genetic algorithm with comprehensive sequential constructive crossover for the travelling salesman problem," IJACSA International Journal of Advanced Computer Science and Applications, vol. 11, no. 5, pp. 245-254, 2020.

[46] Z.H. Ahmed, "A hybrid algorithm combining lexisearch and genetic algorithms for the quadratic assignment problem," Cogent Engineering, vol. 5, Article 1423743, 2018.

AUTHOR'S PROFILE

Zakir Hussain Ahmed is a Full Professor in the Department of Mathematics and Statistics at Al Imam Mohammad Ibn Saud Islamic University, Riyadh, Kingdom of Saudi Arabia. Till the end of 2019, he was in the Department of Computer Science at the same University. He obtained MSc in Mathematics (Gold Medalist), Diploma in Computer Application, MTech in Information Technology and PhD in Mathematical Sciences (Artificial Intelligence/Combinatorial Optimization) from Tezpur University (Central), Assam, India. Before joining the current position, he served in Tezpur University, Sikkim Manipal Institute of Technology, Asansol Engineering College and Jaypee Institute of Engineering and Technology, India. His research interests include artificial intelligence, combinatorial optimization, digital image processing and pattern recognition. He has several publications in the fields of artificial intelligence, combinatorial optimization and image processing.

Review on Personality Types and Learning Styles in Team-based Learning for Information Systems Students

Muhammad Zul Aiman Zulkifli[1], K.S. Savita[2], Noreen Izza Arshad[3]
Computer and Information Science Department
Universiti Teknologi PETRONAS
Perak, Malaysia

Abstract—**Team-based learning (TBL) has become a preferable method in learning approach at higher educational level. There are a lot of articles that discussed on the benefits and process of implementation of team-based learning but lack of studies that focus on the composition of members in team-based learning and effects of personality types and learning styles towards it. This article set out to analyze the existing literatures on team-based learning implementation at undergraduate and how personality types and learning styles affected the learning process plus exploring these topics in information systems field. Guided by Okoli systematic review method, a systematic review from Scopus, Web of Sciences and Association of Information Systems (AIS) databases has been conducted. Results shows that TBL received positive feedback from the scholars but only have issues on the implementation process consist of the usage of student's personality and learning styles, role of team members, TBL management in classroom, TBL is not "fit for all" and current studies about TBL. The usage of personality and learning style instruments is one of the suggested ways to improve it but there are no details guidelines available yet on how to use it. There is lack of studies about team-based learning in information systems field.**

Keywords—Team-based learning; personality type; learning style; undergraduate students

I. Introduction

Team-based learning is a highly designed teaching and learning strategy that maximises student preparation and participation, giving students' responsibility for their own learning before and during class session. Students spend time in class solving authentic problems in essentially self-managed, high-performing, permanent teams [1]. This approach requires students to apply learned knowledge to solve significant, authentic and complex scenarios individually and within a team [2]. Essential elements in TBL, consist of groups, accountability, feedback and assignment design [3]. The explanation for the elements are (1) groups - groups must be properly formed and managed, (2) accountability - students must be accountable for the quality of their individual and group work, (3) feedback - students must receive frequent and timely feedback, and (4) assignment design - group assignments must promote both learning and team development.

Team-based learning play a major role in promoting team learning among students. This can be seen with team learning become a demand within organization as the increasing global competition, consolidation, and innovation. Furthermore, team members affects the degree of the effectiveness of the team as it exhibits the expertise diversity and collective identification towards group performance [4]. In addition, working in group become more beneficial when the tasks and project requires a larger skill and commitment from the member to be completed especially in project involving completed and participation of multiple field [5].

This paper attempt to explore and understand the trends of the research's topics and identify the relationship between personality types and learning styles of undergraduate students within the process of team-based learning. Additionally, this study is vital because team learning becoming a demand within an organization as they requires the use of teams at all hierarchical level [6]. Working in team can be train before the students graduate into working field and higher education institution is one of the suitable medium to nurture it. Therefore, details on where past literatures has so far focused provide the opportunity in understanding on where the emphasis is and where the attention need to be placed. To construct a relevant systematic review, the current article is guided by the main research question – How learning styles and personality of undergraduate students affect team-based learning process? Focus is given on student's personality and learning style as these components may affects the group effectiveness when conducting team-based learning [7], [8]. Thus, this study attempts to investigate the effects of learning styles and personality types on team-based learning among undergraduate students across all fields with emphasize on information systems field.

This section explains the purpose of conducting systematic literature review while the second section details out the methodology process. The last section discussed on the current trends around TBL and its integration between personality and learning styles of learners.

II. Methodology

The author conducted a systematic review by following guideline by Okoli systematic review protocol [9]. A systematic review is a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyse data from the studies that are included in the review [9][10]. It follows a rigorous and scrupulous procedure to search and select the sample studies for coding and analysis. It is a methodical and meticulous process of collecting and collating the published empirical studies of acceptable quality with systematic criteria for selection to reduce researcher bias and provide transparency to the process.

A. *Systematic Review Protocol*

Protocol can be described as a document written before the start of a systematic review describing the rationale and intended purpose of the review, and the planned methodological and analytical approach [10]. According to Okoli review protocol [9], there are eight steps need to be followed when conducting systematic review. These steps are been followed in this paper's review protocol and the explanation are detailed in the following statements.

First step is to identify the purpose of this paper. Align with its objective, this paper reviews the current research on personality types and learning styles of undergraduate student within team-based learning implementation. The systematic review is focusing on domain review where it will highlight the empirical findings of the reviewed papers. All field of studies are included, and more focus is been given in information systems. The second step is drafting the searching protocol. Major search term is been extracted from the research question which are learning styles, personality types and team-based learning. Then, the author identifies the relevant terms, synonyms and alternative spelling that are used in the published literatures as shown in Table I.

Third, the steps continued with developing search strings to find the papers in the databases. The search strings are developed as stated below:

TITLE-ABS-KEY(("team-based learning" OR "team based learning" OR "team learning" OR "group-based learning" OR "group based learning" OR "group learning") AND (("learning style*" OR "learning strateg*" OR "learning approach" OR "learning method") OR ("personalit* type*" OR "personality")) AND ("undergraduate*" OR "under graduate*" OR "higher education" OR "university" OR "degree")).

These search strings will be used in the selected databases to find papers. The selected databases are Scopus, Web of Sciences and Association of Information Systems (AIS). Scopus and Web of Sciences are selected because it the database contain only high-indexed journal while AIS is selected as the focus is to explore more in information systems (IS) thus giving opportunity to the author to find the results of the articles in IS domain.

The fourth step is screening for inclusion. After getting the results from the databases, the articles will be filtered based on the criteria imposed by authors. The screening is important to make sure the papers get is within the topic's domain and sufficient while at the same time the excluded articles do not affect the quality content of this systematic review paper. The articles are selected for review if: (1) Using any of research keyword. (2) Using any of the research questions or attempt to describe its nature. (3) Published in or submitted to conferences or journals. (4) Written in English. (5) Published within 2015 to March 2019. (6) Related to the topics such as active learning and flipped classroom. On the contrary, the publication will exclude if they were: (1) Papers with no empirical findings (e.g. review paper). (2) Papers discussing completely different area from research topics. (3) Masters and PhD studies which are not published in any referred conferences or journals. (4) Informal literature survey (no defined search questions, no search process, no defined data extraction, or data analysis process). (5) Papers with no answer relevant to the research questions. The process of selecting the paper for review is shown at Fig. 1.

Fifth, after filtered all these articles, data extraction process is conducted. During this phase, author will review each article with its objective to answer these following questions: (1) How personality types and learning styles where been integrate in the study? (2) How the study was conducted? What are the results of the study? (3) What are the future recommendations from the study? (4) What are the limitations of the study? The results from data extraction phase are explained in results section.

Sixth, the next process is quality appraisal. After extracting the needed information from all the papers, the papers are going through quality appraisal process to prioritize the papers according to their quality and to exclude certain papers deemed to not useful due to inferior methodological quality. The components that are highlighted in this process are stated as follows: (1) What claims the papers make? (2) What are the evidences they provide to support the claims? (3) Are the evidences are warranted?

TABLE I. KEYWORDS SYNONYMS FOR SEARCH STRING

Keywords	Synonym
Team-based learning	team based learning, team learning, group- based learning, group-based learning, group learning
Personality type	personality
Learning style	learning approach, learning strategy, learning technique, study style, study approach, study strategy
Undergraduate	undergraduate, higher education, university, degree

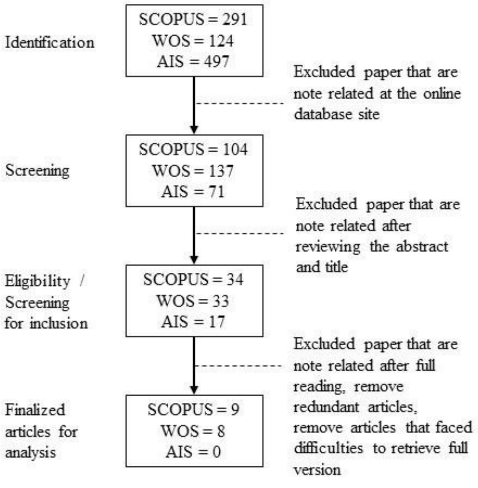

Identification
SCOPUS = 291
WOS = 124
AIS = 497
Excluded paper that are note related at the online database site
Screening
SCOPUS = 104
WOS = 137
AIS = 71
Excluded paper that are note related after reviewing the abstract and title
Eligibility / Screening for inclusion
SCOPUS = 34
WOS = 33
AIS = 17
Excluded paper that are note related after full reading, remove redundant articles, remove articles that faced difficulties to retrieve full version
Finalized articles for analysis
SCOPUS = 9
WOS = 8
AIS = 0

Fig. 1. Searching Protocol used to Select Articles.

Seventh, after conducting quality appraisal process, the data are synthesized using thematic analysis. The findings are been explained in the results section. The last process is to write the review which is the content of this paper itself.

III. Results and Discussion

Based on the selection process stated in methodology section, 17 articles are selected for this review as shown at Table II. From these articles, the content of the articles had been analyzed and the results are presented as follows.

TABLE II. SELECTED ARTICLES FOR REVIEW

Watkins et al. (2018)[11]	Salimath, Vijaylakshmi, & Shettar (2018) [12]
Jeno et al. (2017) [13]	Rezaee, Moadeb, & Shokrpour (2015) [14]
Nicole & Larson (2016) [15]	Kim, Song, Lindquist, & Kang (2016) [16]
Hettler (2015) [17]	Frame et al. (2015) [18]
Iramaneerat & Ba (2017) [19]	Miller, Khalil, Iskaros, & Van Amburgh (2017) [20]
Stepanova (2017) [21]	Obad et al. (2019) [22]
Behling, Murphy, & Lopez (2017) [23]	Ismail (2016) [24]
Frame et al. (2016) [25]	Kenny, Mclaren, Blissenden, & Villios (2015) [26]
Greetham & Ippolito (2018) [27]	

A. *Student's Personality and Learning Style in TBL*

From this search, there are two studies that mentioned the involvement of personality or learning style instrument in their researches: Myers-Briggs Type Indicator (MBTI) [18] and VARK [24]. For MBTI, study shows that this instrument help in creating the team for TBL. They suggest that the personality can become a key role in team success as student believe diversity in team contribute for better functioning group. Knowing team members personalities allows instructor to identify each individual their strengths and weaknesses and make them to be put in the right team that may give them opportunity to display their capabilities to others. At the same time, knowing personalities allows the members to distribute the tasks based on each capability.

Looking at Ismail [24], although the study does not mention VARK, they indirectly highlighted the component of VARK which are visual, audio, reading and kinesthetic and how these can help their students to incorporate their learning style in classroom. Understanding their learning styles help the teams to recognize each other capabilities and allows the distribution to be handle effectively. Other studies also mentioned that understanding student learning style will help to avoid clashing within the team. Student that know their preferences can use it to facilitate their learning for maximum results. From this, personality and learning styles will tackle the same issues in understanding the student's capabilities especially when working in team. Thus, using both personality and learning style instrument will add more value in getting the insights of each member's capabilities. Although the studies highlight the importance of personality and learning style, the are no details model available on how to form the group using these instruments. Further researches are needed for verification on the details of combination and synchronization between personality and learning style components.

B. *Team Members Play Role in Effective TBL*

In TBL, the composition of team members in a group plays major role in determining the successfulness of the learning method. Greetham [27] expressed that TBL help in maximizing learning outcome with better team dynamics, applying the knowledge into more technical tasks and enhance the process of subject mastery. In getting the team dynamics, the formation of the members is crucial. Stepanova [21] expressed that instructor should look on the formation of the team to make sure the team are productive for TBL successfulness. As TBL is known as learner centric approach, getting a right member for team is the first crucial step. There are few suggestions arise regarding the formation of the team. Frame [18] found that group formed with heterogonous members are more successful compare to homogenous members. The members are formed based on MBTI and the success are based on the student's survey. They also emphasized on understanding the student's different personality and their learning style approaches and use it to form the group.

C. *Managing TBL in Classroom*

One of the uniqueness of TBL is its systematic approach compromise of preparation before class, readiness assurance test, application exercises and getting feedback [1]. This approach also means that managing TBL is important because several processes need to be done within limited time. Ismail [24] noted that time management must be properly restructured in TBL in ensuring the learning is time efficient and productive. Supporting this, Watkins [11] also listed time pressures as the challenges face when implementing TBL. They expressed that students faced the pressure of time when they have poor time management, multiple tasks and lack of commitment by other members when conducting the project. Thus, time allocation become an important factor in in handling TBL classroom because every process of TBL need to be executed to get the desired results. Plus, getting students opinions on time allocation of TBL in classroom also vital to know either TBL help them or make them feel overwhelm. These issues need to be highlighted as it happens commonly not only when using TBL but also in other active learning method and traditional classroom session.

D. *TBL is not 'Fit for All'*

Using TBL as the teaching strategy in the classroom does get variety of results from all type of students. Miller [20] and Behling [23] highlight that TBL are more appreciated by senior and older students compare to young and junior students. Senior students seem to more appreciated TBL method due to their understanding of past knowledge, experience in the real-life application and familiarity with peers. Meanwhile, Salimath [12] found that TBL is more suited in course that emphasizes in application in real world by performing course projects while non-TBL method is more suitable for students whose individual learning path are innovative and exposed to different learning environment. From this, it can be found that TBL is more suited to the courses that involve practicality and application towards industry and society while non-TBL session is more suited towards new students that still not familiar with higher education learning environment and theoretical courses.

Kim [16] emphasized that TBL cannot be seen as the panacea that suit all students. He pressed that as the teaching method, TBL should be evaluated from other various factors and more researches should be conducted pertaining the other factor as that may affects the TBL itself. Further researches are needed on TBL implementation in various field of studies and various educational level to know each field compatibility in using TBL method itself.

E. Currents Studies on TBL

From these articles, the distribution of the field of study the researches are conducted as follows: 10 medical field, 2 from engineering, 1 from computer sciences, 1 from social science, 1 from tax and accounting, 1 from business English and 1 from economics. This data shows that most of TBL researches are conducted around medical field. Meanwhile, there are study that came from other fields. This highlight that there is opportunity to conduct researches on TBL for others field as current researches shows that TBL may be fit in all field of studies, but further verification is needed.

For research designs, 12 studies are using quantitative methods, 2 studies using qualitative, 1 study using mixed methods and 2 studies are not clearly stated their research designs. Paper that using qualitative are mostly emphasizing on getting the effectiveness of TBL using the student's marks and feedbacks forms. For studies that using qualitative [11], [26], their focus more on understanding on how TBL works on their field as not much paper can be found within their related field. Study that use mixed method [21] doing both discovering and verification of TBL implementation of TBL in the classroom. For future studies, using qualitative method is preferably when researching TBL in new field of study such as information systems, information technology and other fields that still not familiar yet with TBL implementation. Using quantitative is recommended when doing verification and enhancement of TBL in already familiar field.

Looking at the information systems, this field is not yet applying team-based learning in the teaching class. This may due to the nature of the field that related with information technology (IT) and computer sciences that focus on technical and programming knowledge. Although this is true, information systems field is more towards bridging the technical and IT related entities with businesses application [28]. Thus, capability of information systems graduates to work in team is inevitable, making team-based learning method as a suitable approach in enhancing students' capabilities.

Nevertheless, more studies need to be done in measuring the successfulness of this approach in this field.

IV. Conclusion

Using the approach of systematic review, the author managed to conduct a literature study on team-based learning and examining its relationship with student's personality and learning styles. Relevant literatures have been chosen for reviews and several aspects have been identified that directly involved in this topic. Overall, team-based learning (TBL) shows positive outcome in enhancing students learning outcome in the higher education. Most studies agree that TBL is a strategic learning approach that give benefits in enhancing students learning capabilities and working in the team. However, there still arising the issues are on the implementation of TBL itself. Here, the highlighted issues are on the usage of student's personality and learning styles, role of team members, TBL management in classroom, TBL is not "fit for all" and current studies about TBL.

Notably, there are lack of studies regarding these topics on information systems. It is clearly stated that TBL enhance the team learning among the students and information systems field also required the involvement of team either in learning or projects. Thus, it opens the opportunity for this field to explore this topic as different fields give different views and insights. From this study, there are several recommendation for future studies which are: (1) how to form the right group with the right mixture members to ensure the success of TBL implementation, (2) how to manage time effectively in TBL, (3) how other fields been affected when using TBL in their courses, (4) how far TBL can be implemented in information system field?

References

[1] L. K. Michaelsen, M. Sweet, and B. C. Kelley, "Team-Based Learning," in New Directions for Teaching and Learning, no. 128, Wiley Online Library, 2011, pp. 35–41.

[2] J. Currey, S. K. Sprogis, G. Burdeu, J. Considine, J. A. Allen, and E. Oldland, "Students perceive Team-Based Learning facilitates development of graduate learning outcomes and professional skills,"J. Teach. Learn. Grad. Employab., vol. 9, no. 1, pp. 93–113, 2018.

[3] L. K. Michaelsen, M. Sweet, and D. Parmalee, "The essential elements of team-based learning," in New Directions for Teaching & Learning, vol. 2008, no. 116, 2008, pp. 7–27.

[4] S. Kozlowski and B. Bell, "Work groups and teams in organizations: Review update," Handb. Psychol., vol. 12, pp. 412–469, 2013.

[5] S. Lavy, "Who benefits from group work in higher education? An attachment theory perspective," High. Educ., vol. 73, no. 2, pp. 175–187, 2016.

[6] J. Mesmer-Magnus, A. A. Niler, G. Plummer, L. E. Larson, and L. A. DeChurch, "The cognitive underpinnings of effective teamwork: A Meta-Analysis," Career Dev. Int., vol. 22, no. 5, pp. 507–519, 2017.

[7] T. F. Hawk and A. J. Shah, "Using Learning Style Instruments to Enhance Student Learning," Decis. Sci. J. Innov. Educ., vol. 5, no. 1, pp. 1–19, 2007.

[8] M. Khatibi and F. Khormaei, "Learning and personality: A review,"J. Educ. Manag. Stud., vol. 6, no. 4, pp. 89–97, 2016.

[9] C. Okoli, "A Guide to Conducting a Standalone Systematic Literature Review," Commun. Assoc. Inf. Syst., vol. 37, pp. 879–910, 2015.

[10] D. Moher et al., "Preferred reporting items for systematic review and and explanation meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation," Rev. Esp. Nutr. Humana y Diet., vol. 20, no. 2, pp. 148–160, 2015.

[11] K. Watkins, N. Forge, T. Lewinson, B. Garner, L. D. Carter, and L. Greenwald, "Undergraduate Social Work Students' Perceptions of a Team-Based Learning Approach to Exploring Adult Development," J. Teach. Soc. Work, vol. 38, no. 2, pp. 214–234, 2018.

[12] S. Salimath, M. Vijaylakshmi, and A. S. Shettar, "A Comparative Study of Team Based Learning and Individual Learning," 2018.

[13] L. M. Jeno et al., "The relative effect of team-based learning on motivation and learning: A self-determination theory perspective," CBE Life Sci. Educ., vol. 16, no. 4, pp. 1–12, 2017.

[14] R. Rezaee, N. Moadeb, and N. Shokrpour, "Team-Based Learning : A

New Approach Toward Improving Education," no. 5, 2015.

[15] M. Nicole and L. Larson, "Team Learning Behaviors : Supporting Team-Based Learning in a First-Year Design and Communications Course," 2016.

[16] H. Kim, Y. Song, R. Lindquist, and H. Kang, "Effects of team-based learning on problem-solving , knowledge and clinical performance of Korean nursing students," Nurse Educ. Today, vol. 38, pp. 115–118, 2016.

[17] P. L. Hettler, "Student Demographics and the Impact of Team-Based Learning," Int. Adv. Econ. Res., vol. 21, no. 4, pp. 413–422, 2015.

[18] T. R. Frame, S. M. Cailor, R. J. Gryka, A. M. Chen, M. E. Kiersma, and L. Sheppard, "Student perceptions of team-based learning vs traditional lecture-based learning," Am. J. Pharm. Educ., 2015.

[19] C. Iramaneerat and O. S. Ba, "Assessing the Outcomes of Team-Based Learning in Surgery," vol. 100, pp. 24–32, 2017.

[20] D. M. Miller, K. Khalil, O. Iskaros, and J. A. Van Amburgh, "Professional and pre-professional pharmacy students' perceptions of team based learning (TBL) at a private research-intensive university," Curr. Pharm. Teach. Learn., vol. 9, no. 4, pp. 666–670, 2017.

[21] J. Stepanova, "Team-Based Learning in Business English," pp. 12– 13, 2017.

[22] A. S. Obad et al., "Assessment of first-year medical students ' perceptions of teaching and learning through team-based learning sessions," pp. 536–542, 2019.

[23] K. C. Behling, M. M. Murphy, and O. J. Lopez, "Team-Based Learning in a Pipeline Course in Medical Microbiology for Under-Represented Student Populations in Medicine Improves Learning of Microbiology Concepts †," J. Microbiol. Biol. Educ., vol. 17, no. 3, pp. 370–379, 2017.

[24] N. A. S. Ismail, "Effectiveness of Team-Based Learning in teaching Medical Genetics to Medical Undergraduates," Malays J Med Sci, vol. 23, no. 2, pp. 73–77, 2016.

[25] T. R. Frame, R. Gryka, M. E. Kiersma, A. L. Todt, S. M. Cailor, and A. M. H. Chen, "Student Perceptions of and Confidence in Self-Care Course Concepts Using Team-based Learning," vol. 80, no. 3, 2016.

[26] P. Kenny, H. Mclaren, M. Blissenden, and S. Villios, "Improving the Students Tax Experience: A Team-based Learning Approach for Undergraduate Accouting Students," vol. 10, no. 1, pp. 43–65, 2015.

[27] M. Greetham and K. Ippolito, "Instilling collaborative and reflective practice in engineers: using a team-based learning strategy to prepare students for working in project teams," High. Educ. Pedagog., vol. 3, no. 1, pp. 510–521, 2018.

[28] H. Topi and R. T. Wright, "Differentiating Information Systems and Information Technology as Fields of Study: An Evaluation of Model Curricula," in Siged 2013, 2013, pp. 1–13.

Data Fusion-Link Prediction for Evolutionary Network with Deep Reinforcement Learning

Marcus Lim[1], Azween Abdullah[2], NZ Jhanjhi[3]
School of Computer Science and Engineering (SCE)
Taylor's University, Selangor, Malaysia

Abstract—**The sophistication of covert activities employed by criminal networks with technology has been proven to be very challenging for criminal enforcement fraternity to cripple their activities. In view of this, law enforcement agencies need to be equipped with criminal network analysis (CNA) technology which can provide advanced and comprehensive intelligence to uncover the primary members (nodes) and associations (links) within the network. The design of tools to predict links between members mainly rely on Social Network Analysis (SNA) models and machine learning (ML) techniques to improve the precision of the model. The primary challenge of constructing classical ML models such as random forest (RF) with an acceptable level of accuracy is to obtain a large enough dataset to train the model. Obtaining a large enough dataset in the domain of criminal networks is a significant problem due to the stealthy and covert nature of their activities compared to social networks. The main objective of this research is to demonstrate that a link prediction model constructed with a relatively small dataset and dataset generated through self-simulation by leveraging on deep reinforcement learning (DRL) can contribute towards higher precision in predicting links. The training of the model was further fused with metadata (i.e. environment attributes such as criminal records, education level, age and police station proximity) in order to capture the real-life attributes of organised crimes which is expected to improve the performance of the model. Therefore, to validate the results, a baseline model designed without incorporating metadata (CNA-DRL) was compared with a model incorporating metadata (MCNA-DRL).**

Keywords*—*Metadata; time-series network; social network analysis; criminal network; deep reinforcement learning

I. Introduction

Currently, members of organised crimes often work together to form a resilient and flexible structure to execute their covert and stealthy activities [1]. The CNA tools are mainly constructed based on social network analysis (SNA) models and metrics [2]. SNA which combines knowledge of graph theory and the discipline of social science [3] is a common method employed to analyse the criminal network to uncover hidden structural relationships and key players in criminal syndicates [4,5]. These SNA applications also have a graphical interface that provides a comprehensive visual topological analysis of domain with network orientated dataset [6]. Most social media e.g. Snapchat, Twitter and Facebook recommend relationships using the SNA models based on the likelihood of associations or links of common interest [7].

In the topological analysis of criminal network, environmental factors that can affect the evolving formations of links between participants of the network have to be taken into consideration [8]. These factors such as criminal records, education level, age and family background (Fig. 1) are known as metadata. They provide further circumstantial information that may affect structural patterns of a criminal network over a period of time [9].

Criminal networks tend to exhibit a high likelihood of having unknown links or relationships as criminal activities usually operate in covert and stealthy nature [11]. The incomplete and inconsistent database captured during law enforcement procedures could have been done deliberately by criminals or unintentional human mistakes [12]. In such circumstances, the SNA methods of predicting probable existence or non-existence of links or relationships in criminal networks provide critical information that determines whether attempts to disrupt criminal activities can be successful (Fig. 1).

The prediction of links using SNA metrics is usually based on the structure of the network and supported by information on the contents of the nodes [13].

The link prediction model developed in this research (Fig. 1) leverages on DRL to achieve self-learning and generated datasets which can be combined with smaller domain datasets [14, 15] for ML training. The deep learning (DL) algorithm within the model would reduce reliance on specific human programmed algorithms when formulating an ML function [16-18]. The self-learning ability leverages on reinforcement learning (RL) whereby ML is achieved through recursive trial and error based on a system of rules operating in a domain where points are awarded when each task is successfully completed and deducted as punishments for failure [19].

The model developed in this research may encounter certain limitations in that it is constructed based on relatively small dataset which is a common attribute of criminal or terrorist networks compared to social networks such as Twitter, Instagram and Pinterest. The relatively small dataset may have an impact on the predictive performance of some classical supervised machine learning models being trained.

A. Structure of Paper

Our research paper consists of six sections with Section I is the introduction, Section II is a review of other research works involving evolutionary social network, ML models and implications of metadata fusion. In Section III, an explanation of the proposed and baseline models developed together with the methodology of training is provided. Section IV provides information on the properties of the dataset, set-up of the experiments and a discussion of the results of the experiments. The research conclusion is discussed in Section V and Section VI explores the trajectory for subsequent research work.

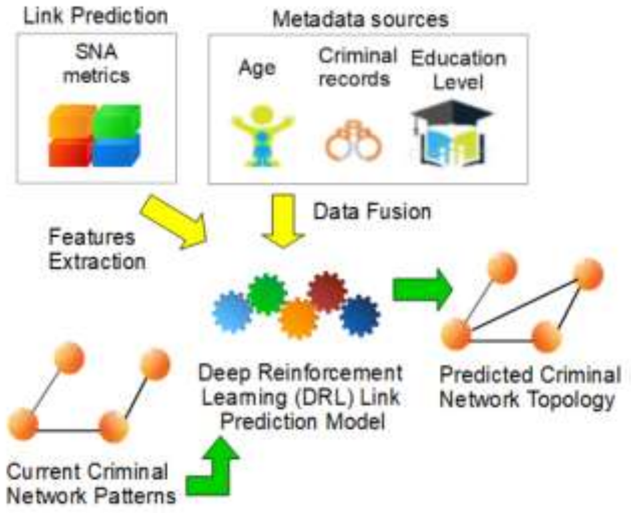

Link Prediction
SNA metrics
Metadata sources
Age
Criminal records
Education Level
Data Fusion
Features Extraction
Deep Reinforcement Learning (DRL) Link Prediction Model
Predicted Criminal Network Topology
Current Criminal Network Patterns

Fig. 1. MCNA-DRL Model to Predict Links for Evolutionary Criminal Network [10].

II. Related Work

In [13] Budur E. et al. developed an SNA model to predict hidden or missing links in a criminal network. The authors explained that the problem with real-world criminal dataset is that it is relatively small compared to social network dataset to effectively trained classical ML models. They leveraged upon the gradient boosting machine (GBM) ML algorithm and trained with a large dataset of about 1.5 million nodes and 4 million links by combining a few datasets. The large dataset is expected to capture the real-world nature of the criminal network with better precision. In their experiment, their proposed model managed to perform with higher predictive accuracy compared with models developed in prior works which performed with a smaller dataset. The authors did state that predictive accuracy could be improved if a time series dataset was used as it will better represent the properties of real-world criminal syndicates [20].

In [21], S. K. Dash et al. developed an SNA model to predict hotspot crime location by fusing crime data with metadata attributes such as police station proximity, education quality and emergency service call to improve the quality of the prediction. The model leveraged upon the supervised machine technique of support vector machine (SVM) which was trained on a dataset with feature extracted from environmental, social factors related to crime prediction. The data fusion was found to have improved the prediction quality of the model.

In [22], Bliss C. A. et al. proposed a link prediction model with Covariance Matrix Adaptation Evolution Strategy as link weights which ensemble 16 neighbourhood and similarity indices and leveraged on an evolutionary computing algorithm. The link prediction model, combining 16 SNA metrics, leveraged on evolutionary computing and trained with Twitter reciprocal reply network (RNN) dataset, was found to perform better than other supervised learning models such as SVM constructed from SNA metrics derived independently and in isolation. They also suggested future research to include geospatial features and community structure in a time-evolving network,

Sarvari, H. et al. [23] used SNA techniques such as centrality measures, PageRank and clustering coefficient to gain insight into the organisation of criminal community by constructing a large scale graph from a smaller dataset, for example, email address of criminals. The techniques of SNA analysis of large constructed social graph information were are able to provide a more detail profiling of criminals. Further research was suggested to incorporate the profile linking social graph from Facebook to other social networking media, e.g. Google+ and Twitter to derive a considerably complete profile of the scammers.

A recent breakthrough was achieved by Silver et al. [24] where an ML program that they developed with DRL, AlphaGo, demonstrated super-human performance by defeating the world's top grandmaster from China in the board game of Go, which has more permutation possibilities than all the atoms in the known universe. The feat was achieved with a combination of Monte Carlo tree search (MCTS) algorithm, which replicated the intuitive judgement capability of human to narrow the search scope to board patterns with the highest likelihood of success.

In the subsequent trajectory of their research on DRL, Silver et al. enhanced their AlphaGo program by developing AlphaGo Zero which was able to self-learn using dataset generated via self-play against prior versions of itself [25]. AlphaGo Zero was provided only with the basic domain rules of the game and was able to defeat AlphaGo after 3 days of self-learning. The DRL algorithm developed had reduced the reliance on incorporating human-crafted domain knowledge to achieve predictive performance. Therefore the DRL model had opened up possibilities of applying the algorithm to train other ML models with a relatively small real-world dataset.

In [26]. Lim, Marcus et al. have incorporated findings from the research work of Silver et al., and leveraged upon the algorithm of DRL to construct a link prediction in the domain of criminal network with a relatively small snapshot dataset combined with a self-generated dataset. The research yielded some positive results indicating DRL algorithm could be used to construct predictive models with adequate precision when trained with self-generated dataset in accordance with domain rules.

From the related works reviewed, there is little evidence that both DRL and metadata fusion technique have been incorporated into the field of link prediction for a dynamic criminal network structure which changes over time. This research is expected to fill the gaps by investigating the manner DRL and metadata fusion can be integrated to train a model to predict with better precision on a more diverse evolutionary graph-based dataset such as a terrorist network.

III. Models and Methodology

A. Proposed MCNA-DRL Model

The problem of predicting the formation or disappearance of edges in a network is treated as a binary classification task in ML modelling process.The data fusion DRL link prediction CNA (MCNA-DRL) model (Fig. 2) was developed as an extension of the research done by Marcus Lim et al. [27] using the MCTS model in the link prediction process.

The DL algorithm of the MCNA-DRL model has a significant influence on the overall performance of the model as it relies on the optimisation of parallel processing with graphics processing unit (GPU). The predictive accuracy of the MCNA-DRL model is assessed with the area under curve (AUC) scores [28].

The DL which functions as the value network for RL approximates a vector of a probability distribution (Fig. 3), computed from the SNA metrics as weights based on structural features of the nodes (vertex) and links (edges). In the formulation of the feature matrix (Fig. 2), metadata such as the count of criminal records, age and education levels are derived as weights in the training of the neural network. The metadata formulated weighted edges will approximate the output values provided by the neural network to rank node pairs based on the likelihood that links (edges) are predicted to form or disappear. The MCTS algorithm will perform the tree traversal commencing from the edges that achieved the top raking scores. The aggregated predicted scores computed from every state of a completed network traversal is then fed back to the value network to recalibrate the model's hyper-parameters to improve the precision of the prediction in the next iteration (Fig. 2).

B. Methodology

The classical SNA metrics (Table I) to predict the links are calculated for every pair of nodes and formulated as a feature matrix for the purpose of training the link prediction model [29]. During the features learning process, the DL (neural network) algorithms are trained to predict the probable edges as a classification of positive or negative edges (Fig. 2). An edge that is predicted to form in the next instance is tagged as a positive label or is tagged as a negative label if it disappears.

The SNA metrics are computed for each edge of the node pair of the criminal network where $\varphi(i)$ represents the neighbouring nodes in the network of node i. k_i refers to the degree of node i. $n_{ij}^{(t)}$ represents the number of walks of length t for each pair of nodes i and j. β denotes the discount factor for the computation of walks of longer length.

During testing, the prediction of links is made for every sample node pair based on the score aggregated from an array with multiple SNA feature metrics (Fig. 2).

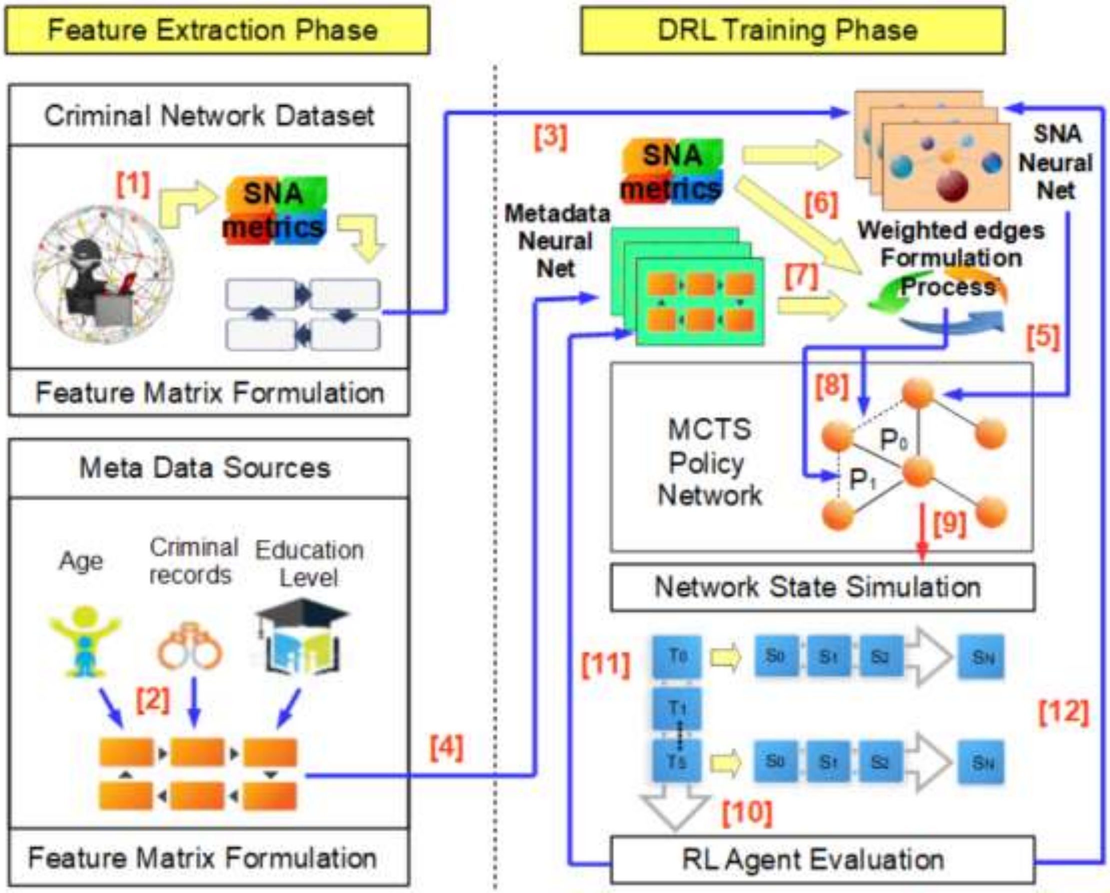


Fig. 2. Proposed MCNA-DRL Model for Link Prediction [10].

TABLE I. SNA METRICS FOR LINK PREDICTION [13]

Metrics	Definition
Common neighbour	$S_{xy} = \lvert \varphi(x) \cap \varphi(y) \rvert$
Jaccard Index	$S_{xy} = \frac{\lvert \varphi(x) \cap \varphi(y) \rvert}{\lvert \varphi(x) \cup \varphi(y) \rvert}$
Hub Index	$S_{xy} = \frac{\lvert \varphi(x) \cap \varphi(y) \rvert}{\min(k_{x,} k_y)}$
Preferential Attachment index	$S_{xy} = k_x \text{ x } k_y$
Adamic-Adar Index	$S_{xy} = \sum_{z \in \varphi(x) \cap \varphi(y)} \frac{1}{\log k_z}$
Katz	$S_{xy} = \sum_{t=1}^{\infty} \beta^t . n_{xy}^{(t)}$

In the MCNA-DRL model, the DL algorithm is an approximation function that is received as inputs of the first state of the network, S_0, and it computes the vectors indicating the likelihood of the existence or non-existence of the edges. The probability values derive from these vectors serve as weights of these edges in the link prediction process.

The SNA neural network (Fig. 2) learns from the values of these weights, and the value network generates the estimated measures via self-simulation using the SNA metric scoring by leveraging on the RL technique.

The MCTS network traversal commence at the root node, and the traversal to the following node creates a new state from the present network state based on the likelihood of an edge being formed or removed. A probable edge identified from the present state, S_1 to the subsequent state, S_2 is due to any action taken by the agent is based on the binary classification rules of evaluation will then be fed back to both the value and policy networks where a cost function will then calibrate the hyper-parameters again to enhance the predictive performance in the next iteration.

Notes (Fig. 2):

1) The Criminal Network dataset is mapped into SNA feature matrix for link prediction.

2) Metadata features is to a multi-dimensional feature matrix.

3) The SNA feature matrix of the Common Neighbour, Jaccard, Adamic-Adar Metrics, serve as the input layer of the value network.

4) The metadata feature matrix input such as the score of crime records, age and education level are processed to be processed by metadata data fusion value network. The SNA metrics of Hub Index and Preferential Attachment index functions as weights for the hidden layer 1 and hidden layer 2 respectively of the function approximator value Network.

5) The SNA metric function approximator identifies node-pairs with the highest likelihood of link formation or destruction.

6) The SNA feature matrix will also be factored in the data fusion weighted edges formulation process.

7) The output from .metadata fusion neural net is processed data fusion weighted edges formulation algorithm.

8) MCTS module simulates the network instance generation commencing on random node-pairs sorted by the links most likely to form or disappear (P_0, P_1) derived by The SNA metrics weighted edges formulation process.

9) The States, S_0 to S_N denote networks reconstructed with the identified hidden links at the end of each simulated link prediction rollout. The States generated are evaluated against the 5 test dataset instances (T_0 to T_5) to measure the degree of success in the link prediction.

*10)*The evaluation score from a prior instance is feedback to recalibrate the policy and value network to reduce errors in the next iteration.

*11)*The predictive performance evaluation score from time-elapsed training dataset (T_0 to T_5) by the RL is used to recalibrate the metadata fusion neural network.

*12)*The predictive performance evaluation score from time-elapsed training dataset (T_0 to T_5) by the RL is used to recalibrate the SNA neural network function approximator.

The AUC metric is used to evaluate both the MCNA-DRL and baseline CNA-DRL models. The AUC metric may have values from 0 to 1, where a score of 1 achieved by a model represents the best predictive precision.

C. *Time-Evolving Network*

In this research, both the MCNA-DRL model and the baseline CNA-DRL model is trained using the Madrid bombing time-series dataset based on the Rooted PageRank [28] algorithm. Every node pair is ranked based on the weights derived in accordance to the elapsed time between the present instance and the next instance of prediction process, Given a pair of nodes x and y with a common node, z that may exist between these nodes, the probability of a traversal commencing from x to y is represented as [30]:

$$P(x,y) = (1-\alpha)\frac{w(x,y)}{\sum_{z \in \varphi(x)} w(x,z)} + \begin{cases} \alpha \text{ if } y \text{ is the central node} \\ 0 \text{ if otherwise} \end{cases} \qquad (1)$$

The time factor is formulated in Rooted PageRank as a weight with the time interval being a probability scaled in accordance with the distance between a pair of nodes.

The time-evolving network can be used to model social groups with structural configurations that change over time [30]. The structural configuration that varies over time may be caused by actors (nodes) joining or dropping out of the network as time passes.

D. *Metadata Fusion*

The fusion of metadata is the technique of combining various data sources derived from the external factors of the environment, which may have an impact on the features extracted to train a predictive model. Metadata in the context of a criminal network that could have an impact on the behaviour of actors (nodes) to participate or exit from the network over time are criminal records, age, education level and family background [31]. In the construction of the MCNA-DRL model (Fig. 2), the number of criminal records, age and education level are factored as weights to train the metadata fusion DL value network. The value computed by the metadata fusion DL is factored in the calculation of the weights to rank the edges based on the likelihood to change in the next instance. The feature matrix extracted from the metadata and factored as weight is computed as follows [31]:

$$v_k = \sum_i w_{ki} x_i + b_i \qquad (2)$$

with *v* representing the feature vector of the DL layers, *w* refers to the weight of every time-elapsed, *k*, for node *i* with *b* indicating the related bias. The bias is recalibrated at the completion of every training cycle.

The SNA metrics is combined linearly with the metadata weight index (2) in the formulation of the feature matrix for training the DL. Linear combination is used to simplify the resource hungry computation process. The combined weights index are used for making the actual prediction during the prediction process period. The combined index computed for every node pair would allow prediction to be made using alternative combination of parameters while reducing greatly the computation resource required by the technique. The first set of model parameters used as input to the DL is derived in random from parameter space for every prediction iteration.

IV. Experiments and Results

The terrorist network dataset of the Madrid train bombing in 2004 is a time-series dataset containing 20 time periods from the years 1985 to 2006 involving some 55 nodes (actors) [32] (Fig. 3, 4). The proposed MCNA-DRL model and CNA-DRL models are evaluated based on the AUC score which is a typical technique adopted to evaluate the precision of the classification models [13].

For the purpose of this experiment, only the dataset from the years 1998 to 2003 was used before the 2004 bombing event which was an exceptional event not reflective of the normal factors affecting the structural changes of the network.

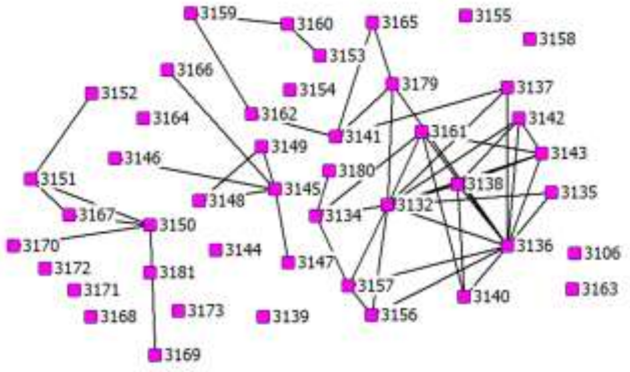


Fig. 3. Actual Criminal Network at Time-Stamp $T_{2002.}$

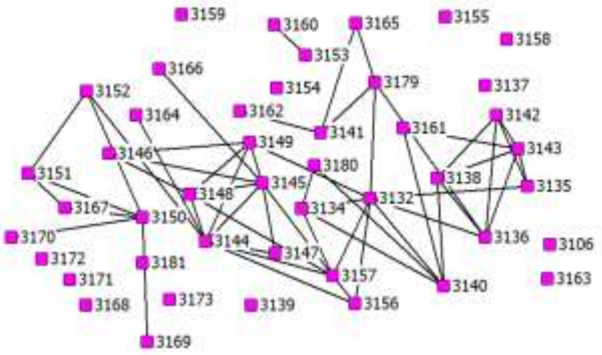


Fig. 4. Actual Criminal Network at Time-Stamp $T_{2003.}$

A. Experiment Set-up

To train the CNA-DRL and MCNA-DRL models, the dataset is formulated into a feature matrix whereby each state of the network represents the formation or cessation of an edge. The original node pair edge at each state is mapped onto a feature matrix with values from a prior time snapshot where a criminal link comes into existence or disappear (Fig. 2).

The Madrid bombing dataset segregated randomly into two (2) subset with a ratio of 75%:25% was used for training and testing respectively. The training set is extracted from the years 1998 to 2002 to build the feature matrix and used for training both the models. The number of positive links denoting the formation of new links in the next time step is obtained. The negative edges denoting cessation of the existing links in the next time step are then randomly chosen to match the number of positive links.

The performance evaluation score from the time-elapsed training dataset is fed back to the neural network to recalibrate the hyper-parameters using a cost function to minimise the error in prediction by both models in the next instance. The prediction of links is then simulated with the trained models which have been recalibrated on the test dataset (Fig. 5 and 6).

B. Results and Discussion

The MCNA-DRL model was able to correctly predict more edges (Fig. 5) that were supposed to appear in the topology of the year 2003 network than the CNA-DRL model (Fig. 6) when compared to the original terrorist network topology at T_{2003} (Fig. 4).

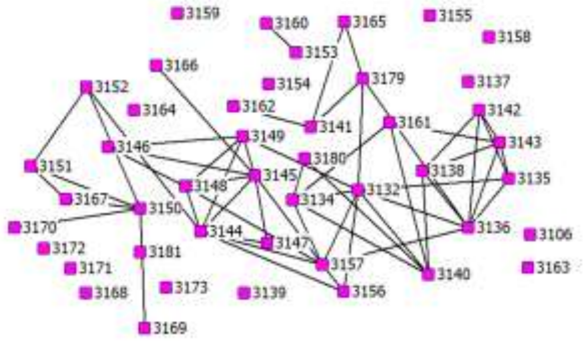


Fig. 5. Predicted Network by MCNA-DRL Model at Time-Stamp $T_{2003.}$

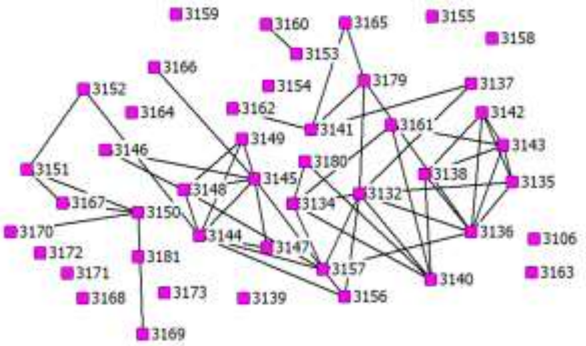


Fig. 6. Predicted Network by CNA-DRL Model at Time-Stamp T_{2003}.

The CNA-DRL model did not predict four new edges, i.e. node pairs (3146, 3149), (3144, 3164), (3132, 3149) and (3150, 3152) correctly. The CNA-DRL model did not correctly predict four edges that should have disappeared, i.e. node pairs (3134, 3161), (3132, 3137), (3137, 3141) and (3136, 3157) (Fig. 6) compared to the actual network in year 2002 (Fig. 3).

The MCNA-DRL model did not correctly predict one new edge, i.e. node pairs (3144, 3164) and two edges that should have disappeared, i.e. node pairs (3134, 3161)(3136, 3157) (Fig. 6) compared to the actual network at the year 2002 (Fig. 3).

Comparing the predicted terrorist network structure the year 2003, the results of the experiment indicate that the MCNA-DRL model (Fig. 6) which incorporates weights derived from the metadata incorrectly predicted five edges less than the CNA-DRL model (Fig. 5). Therefore, the features of metadata data sources factored as weights, attributed by the metadata formulation process seem to have contributed to the higher predictive precision of the MCNA-DRL model. This could be because of the incorporation of the metadata which captures the real-life environmental features of the terrorist network.

The AUC scores of the MCNA-DRL prediction model (Fig. 7) that factor in the metadata as weights achieved a higher AUC score than the CNA-DRL prediction model which did not incorporate metadata fusion by a score of 0.09 (Tables II).

The overall better performance of the MCNA-DRL model could be attributed to the fact that metadata provides other co-related environmental information that may strengthen or weaken the relationships between the nodes over time. This information improves the likelihood of identifying edges which can reduce the scope of the search performed by the MCTS algorithm.

The results also demonstrated that both models, constructed by leveraging on DRL, achieved predictive precision with the AUC scores above 0.5 (Fig. 8). This predictive precision was achieved despite the original dataset being relatively small compared to the most social networks as the models were further trained with self-simulated instances by RL.

The results of the experiment conducted are consistent with the investigation on DRL by Lim, Marcus et al. who constructed a criminal network link prediction model and trained on a snapshot dataset [27]. The current research represents an extension of the work done by the same research team [27] which made comparison of the DRL technique with classical GBM, SVM and RF techniques for link prediction models that were also trained on relatively small dataset which is characteristics of most criminal network (Table III). The comparisons made indicate that the classical models generally need to be trained on relatively larger dataset to achieve a better predictive accuracy than the DRL model that could be trained with the domain dataset and self-generated dataset.

Comparisons are also made with the time-evolving link prediction model (TDRL-CNA) model by Lim, Marcus et al. [28] which did not incorporate metadata fusion (Table III). While the TDRL-CNA model performance seems to peak after 1500 iterations, the MCNA-DRL model still managed to achieve a marginal improvement in the predictive accuracy by incorporating metadata after extended training iterations.

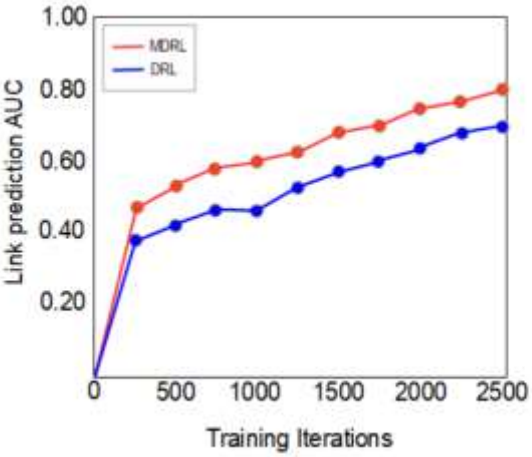


Fig. 7. AUC Score for the MCNA-DRL and CNA-DRL Link Prediction Models for Madrid Bombing Terrorist Network.

TABLE II. AUC SCORES OF MCNA-DRL LINK PREDICTION MODEL AND CNA-DRL MODELS

Dataset	AUC	Time-score(Hr)	Iterations
MCNA-DRL	0.79	4.3	2500
CNA-DRL	0.70	3.9	2500

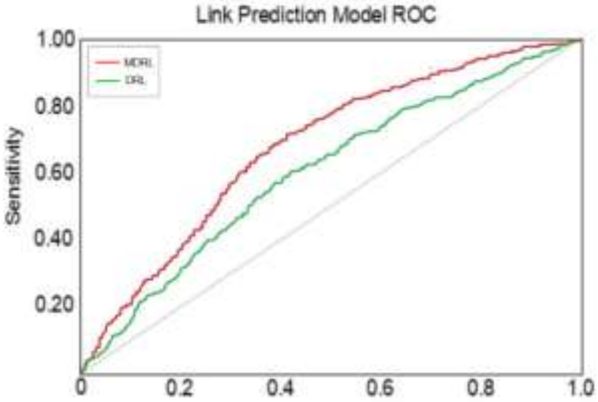


Fig. 8. ROC Curve of Link Prediction Model.

TABLE III. COMPARISON OF DRL LINK PREDICTION MODELS FROM RELATED RESEARCH WORKS

Model	**MCNA-DRL**	**TDRL-CNA**	**DRL-CNA**
ML technique	DRL with metadata fusion	DRL	DRL
Tree search ranking algorithm	MCTS	Breadth first search	Depth first search
SNA metrics	classical	classical	classical
Dataset	20 time-periods	11 time-periods	snapshot
Maximum nodes	55	27	62
Training time-score (hour)	4.3	Not available	1.2
Training iterations	2500	1500	1500
AUC Score	0.79	0.78	0.73
Authors	Current work	[28]	[27]

V. CONCLUSION

The results from this research was able to demonstrate that a model can be trained for the purpose of link prediction with a combination of metadata, relatively smaller dataset and self-generated dataset by leveraging on DRL. These results are evidenced by the AUC score of 0.79 and 0.70 achieved respectively by the MCNA-DRL and CNA-DRL models (Tables II). However, further experiments may need to be conducted to confirm if models constructed with DRL can achieve a better predictive performance than classical supervised ML models if a large scale dataset is used.

VI. FUTURE WORK

The future direction of this research will consider developing a new SNA metric and network search algorithm based on evolutionary computing to further improve the precision of the MCNA-DRL model. The performance of the search algorithm based on evolutionary computing will be compared with the MCTS model. The new SNA metric, will be indexed with metadata weights and is expected to further enhance predictive precision of the model as current findings indicate that models incorporating metadata are more reflective of real-world characteristics.

REFERENCES

[1] Duijn, Paul AC, Victor Kashirin, and Peter MA Sloot. "The relative ineffectiveness of criminal network disruption." Scientific reports 4: 4238, 2014.

[2] Taha, Kamal, and Paul D. Yoo. "A system for analyzing criminal social networks." In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 1017-1023. 2015.

[3] Qazi, Nadeem, and BL William Wong. "Behavioural & tempo-spatial knowledge graph for crime matching through graph theory." In 2017 European Intelligence and Security Informatics Conference (EISIC), pp. 143-146. IEEE, 2017.

[4] Kumar, Arunima S., and Raju K. Gopal. "Data mining based crime investigation systems: Taxonomy and relevance." In 2015 Global Conference on Communication Technologies (GCCT), pp. 850-853. IEEE, 2015.

[5] Giorgos Cheliotis, Associate Professor National University of Singapore. “Social Network Analysis”, 2014.

[6] Thangamuthu, Mr AP, Mr G. Vadivel, and Mrs A. Priyadharshini. "Detecting Criminal Method using Data Mining.", 2019.

[7] Campana, Mattia G., and Franca Delmastro. "Recommender systems for online and mobile social networks: A survey." Online Social Networks and Media 3 : 75-97, 2017.

[8] Potgieter, Anet, Kurt A. April, Richard JE Cooke, and Isaac O. Osunmakinde. "Temporality in link prediction: Understanding social complexity." Emergence: Complexity & Organization (E: CO) 11, no. 1 : 69-83, 2009.

[9] Huang, Zan, and Dennis KJ Lin. "The time-series link prediction problem with applications in communication surveillance." INFORMS Journal on Computing 21, no. 2 : 286-303, 2009.

[10] Lim, Marcus, Azween Abdullah, N. Z. Jhanjhi, and Muhammad Khurram Khan. "Situation-Aware Deep Reinforcement Learning Link Prediction Model for Evolving Criminal Networks." IEEE Access 8 : 16550-16559, 2019.

[11] Dijkstra, L. J., Andrei V. Yakushev, P. A. C. Duijn, A. V. Boukhanovsky, and Peter MA Sloot. "Inference of the Russian drug community from one of the largest social networks in the Russian Federation." Quality & Quantity 48, no. 5 : 2739-2755, 2014.

[12] Spapens, Toine. "Macro networks, collectives, and business processes: An integrated approach to organized crime." European Journal of Crime, Criminal Law and Criminal Justice 18, no. 2 : 185-215, 2010.

[13] Budur, Emrah, Seungmin Lee, and Vein S. Kong. "Structural analysis of criminal network and predicting hidden links using machine learning." arXiv preprint arXiv:1507.05739, 2015.

[14] Li, Haoqi, Naveen Kumar, Ruxin Chen, and Panayiotis Georgiou. "A deep reinforcement learning framework for Identifying funny scenes in movies." In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3116-3120. IEEE, 2018.

[15] Bahdanau, Dzmitry, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. "End-to-end attention-based large vocabulary speech recognition." In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4945-4949. IEEE, 2016.

[16] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." In European conference on computer vision, pp. 818-833. Springer, Cham, 2014.

[17] Chen, Xi-liang, Lei Cao, Chen-xi Li, Zhi-xiong Xu, and Jun Lai. "Ensemble network architecture for deep reinforcement learning." Mathematical Problems in Engineering, 2018.

[18] Duan, Yan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. "Benchmarking deep reinforcement learning for continuous control." In International Conference on Machine Learning, pp. 1329-1338. 2016.

[19] Yao, Kaisheng, Geoffrey Zweig, Mei-Yuh Hwang, Yangyang Shi, and Dong Yu. "Recurrent neural networks for language understanding." In Interspeech, pp. 2524-2528. 2013.

[20] Sharan, Umang, and Jennifer Neville. "Exploiting time-varying relationships in statistical relational models." In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pp. 9-15. 2007.

[21] Dash, Saroj Kumar, Ilya Safro, and Ravisutha Sakrepatna Srinivasamurthy. "Spatio-temporal prediction of crimes using network analytic approach." In 2018 IEEE International Conference on Big Data (Big Data), pp. 1912-1917. IEEE, 2018.

[22] Bliss, Catherine A., Morgan R. Frank, Christopher M. Danforth, and Peter Sheridan Dodds. "An evolutionary algorithm approach to link prediction in dynamic social networks." Journal of Computational Science 5, no. 5 : 750-764. 2014.

[23] Sarvari, Hamed, Ehab Abozinadah, Alex Mbaziira, and Damon Mccoy. "Constructing and analyzing criminal networks." In 2014 IEEE Security and Privacy Workshops, pp. 84-91. IEEE, 2014.

[24] Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert et al. "Mastering the game of go without human knowledge." nature 550, no. 7676 : 354-359. 2017.

[25] Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." Science 362, no. 6419 : 1140-1144. 2018.

[26] Lim, Marcus, Azween Abdullah, N. Z. Jhanjhi, and Mahadevan Supramaniam. "Hidden link prediction in criminal networks using the deep reinforcement learning technique." Computers 8, no. 1 : 8. 2019.

[27] Lim, Marcus, Azween Abdullah, and N. Z. Jhanjhi. "Performance optimization of criminal network hidden link prediction model with deep reinforcement learning." Journal of King Saud University-Computer and Information Sciences, 2019.

[28] Lim, Marcus, Azween Abdullah, N. Z. Jhanjhi, Muhammad Khurram Khan, and Mahadevan Supramaniam. "Link Prediction in Time-Evolving Criminal Network With Deep Reinforcement Learning Technique." IEEE Access 7 184797-184807. 2019.

[29] Özcan, Alper, and Şule Gündüz Öğüdücü. "Multivariate temporal link prediction in evolving social networks." In 2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), pp. 185-190. IEEE, 2015.

[30] Casteigts, Arnaud, Paola Flocchini, Walter Quattrociocchi, and Nicola Santoro. "Time-varying graphs and dynamic networks." International Journal of Parallel, Emergent and Distributed Systems 27, no. 5 : 387-408. 2012.

[31] Satpathy, S., and A. Mohapatra. "A data fusion based digital investigation model as an effective forensic tool in the risk assessment and management of cyber security systems." In The 7th international conference on computing, communications and control technologies. 2009.

[32] Borgatti, S.P., Everett, M.G. and Freeman, L.C. 2002. Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.Vol.1.2002.

Underwater Wireless Sensor Network Route Optimization using BIHH Technique

Turki Ali Alghamdi
Department of Computer Science
College of Computer and Information Systems
Umm Al-Qura University, Makkah. Saudi Arabia

Abstract—**Underwater wireless sensor network (UWSN) is established in water bodies such as oceans, seas and rivers to observe the activity of military, to perform rescue operations and to do mining activity of resources. The sensor nodes communicate through acoustic channels. These nodes have limited battery life (energy), narrow bandwidth and a channel is incurred with delays and noise posing security thrust. The art of work presented different routing protocols in this era to utilize energy and bandwidth efficiently with less delay and to provide the security against black hole attack. However, these methods do not show an appropriate enhancement in the security and to utilize the bandwidth efficiently due to mobile environment. As a result of which, the delay also increases. In this paper a secured and bandwidth utilization path is enhanced using Bellman Inora Hex Hamming technique (BIHH), which not only improves the performance of the routing but also saves the energy. The presented approach is validated with network simulator.**

Keywords—Sensor nodes; energy; routing; black hole; hamming code; hex code

I. Introduction

Underwater communication becomes necessary to the entire world to obtain necessary information about the underwater resources (e.g. minerals), to check the occurrence and frequencies of tsunamis and to monitor the warfare environment. These issues change the mindset of the research community to work in this era. The method is about transmitting and getting communication under the exploitation of sound transmission in underwater atmosphere and it is acknowledged as an acoustical or an audio communication. The Underwater sensor networks (UWSNs) comprise numerous sensors and vehicles to be organized in a particular area, to complete joint observation and to gather the data tasks [1]. Conventionally for observation of sea base, the ocean sensors are used to monitor the information at a fixed position and recuperate the equipment's at the end of assignment. The drawback of the conventional technique is that it lacks the cooperation in the announcement among dissimilar ends, the monitored data is of no use and in the case of any collapse, and the monitored information will be damaged.

The main key issue of UWSNs is node mobility (i.e. mobility of node is almost 2-3m/sec. due to water currents [2]) and energy of the nodes. Several techniques have been proposed to address these issues. However, few gaps still exist (i.e. usage of bandwidth, security of the network and delay in the network). Therefore, an efficient routing protocol is needed to UWSNs which can improve the energy, bandwidth, and security and reduces the end to end delay.

In this paper, an efficient path estimation technique is being presented using Bellman Inora Hex Hamming (BIHH) technique. This protocol also provides security to UWSNs against well-known attack called black hole attack which is very common type of attack to mobile sensor network.

The remaining part of the paper is discussed as Section 2 represents related work, Section 3 represents the proposed technique, and Sections 4 and 5 represents the simulation results and conclusion of the paper, respectively.

II. Related Work

Cinar and M. Bulent Orencik [3] proved that sensor networks in underwater have wide range of applications such as pollution check in sea and rivers, monitoring wind pressure, aquatic surveillance and can also be used to monitor the warfare environment. Underwater Sensor Network (UWSN) with the acoustic channel is the only technique that can be used to measure various network parameters in the sea [4, 5], as it is considered that the velocity of sound is constant in underwater. But the velocity is changed by temperature difference, salinity and depth of the water. As a result of which sound in underwater environment varies [6]. Due to characteristics change in underwater environment, it becomes challenging to utilize the acoustic channel [7] (e.g. multipath environment effects the phase and fading fluctuations, Doppler Effect is present at both source and destination nodes). In [8-10] comparison of the weaknesses and strengths of MAC layer protocols for both single and multipath environment in underwater sensor networks is presented. The three-dimensional network arrangements and the ground level of the ocean are observed by anchoring sensor nodes [11-13].

In [14, 15] the authors presented routing protocols for UWSNs in which network lifetime has been improved. To achieve this goal the authors used autonomous vehicle to collect the data from gateways and used shortest path to transfer the data from sensor nodes to gateways, by minimizing the associated nodes in the network. However, minimization of nodes in the path increases delay and packet delivery ratio is reduced. Jing Li et al. [16] presented an energy efficient protocol for UWSNs, in which packet delivery ratio is improved by managing the energy and power allocation. However, the given algorithm is more complex and enhances the delay which in turn reduces the overall output and may not be suitable for real time applications. Faheem. M et al. [9] presented a quality of service (QoS) routing protocol for UWSNs based on clustering technique. But due to autonomous structure it is difficult to maintain cluster head for long period of time and therefore reduces the routing performance.

Zhiping Wan et al. [17] proposed an energy efficient multilevel routing algorithm for UWSNs. In this algorithm a hierarchical structure based on residual energy was designed to calculate the competition radius size. However, checking the residual energy continuously decreases the packet delivery ration. Meiju Li et al. [18] presented a shortest path routing technique for underground water acoustic networks based on vertical angle. In this algorithm prioritization concept is used to check the vertical angle at every anchoring sensor node till it reaches to destination which increases the complexity and reduces throughput. M. Awais et al. [19] presented an energy efficient algorithm using void-hole alleviation technique. In this method the forwarding node determines the next hop. However, two same hops with different link weights may have different delays (i.e. more is the link weight more will be the delay) so cannot be treated as optimum route from source to destination. Adil et al. [20] proposed an energy efficient method for UWSNs using EH-ARCUN technique. This method entirely depends on cooperation of sensor nodes within the network, so may not be suitable for heterogeneous networks. In [21, 22] two different energy efficient routing protocols have been proposed using IoT, however both the two techniques maximize the delay as number of IoT sensors are employed in the existing network. So, may not be the optimum for real time applications. UWSN has several challenges i.e. bandwidth, error rate and failure of the route. This includes mobility of 2-3 m/sec. at water current [23].

From the above theories and models, it has been observed that path has been optimized either by energy (i.e. using clustering technique, minimization of nodes etc. and inserting number of sensor nodes in the network) or choosing the shortest path between end nodes. However, both the two may not be the optimum as the delay parameter is increased with these approaches, which in turn maximizes the consumption of energy and reduces the overall throughput and packet delivery ratio, so may not be optimum for real time applications. Also, along with the above mentioned issues this security against malicious node attack must be enhanced, because these attacks have been addressed in the UWSN using key distribution which may not always true and optimum due to an autonomous structure. For route optimization, a model must be proposed that can overcome these issues.

In this paper, a secured and efficient path optimization technique is being presented which improves the throughput, utilizes the bandwidth efficiently and enhances the energy. Also, the paths from source to destination are being ranked based on link cost and its energy. In addition, security is also been provided to the network, unlike the conventional techniques which considers only energy or only the shortest distance of the links between the nodes.

III. Proposed Approach

Under water wireless sensor network have numerous challenges, such as consumption of energy, optimization of path from source to destination, utilization of bandwidth and security. In this paper, the main focus is given towards the path estimation and security against black hole attack, which is common attack in this network. To achieve this goal, the paper has three folds.

- Estimation of paths from source to destination and rank them.
- Establishment of alternate route in case of failure based on the rank of the path (from source to destination).
- Security against black hole attack.

A. Estimation of Path

To estimate the path from source to destination, Bellman Ford technique is used to obtain the least cost path between end nodes. Initially the approach estimates calculation of ‘m’ nodes that involves the cost of each of its neighboring node links from a definite source node ‘S_N’ ($d_i^{(H)}$) where ‘H’ is the hop count. It is assumed that each node has a link cost and paths for other nodes, which are available in the network. Also, the information which is available with the node can be exchanged directly to its neighboring nodes in regular time intervals. Based on this information it updates the link cost and the available paths. The notations used to describe this technique can be represented as

M = No of nodes

S_N = Source node

N = Number of nodes which are incorporated within this approach

lc_{ij} = Cost of the link from ith node to jth node

However, if the node are not connected directly then ‘$lc_{ii} = 0$’ and ‘ $lc_{ij} = \infty$’. Whereas if ‘$lc_{ij} \geq 0$’, then nodes are connected directly.

$P_{lc(m)}$ = least cost path from ‘S to m’, under the limit that the links should not more than ‘p’.

p = max. links in the path

Algorithm

The following are the steps of algorithm to find the shortest path, however step 2 repeats the link cost change.

1. Initialize
$p_{lc(m)}^{0} = \infty, \forall\ m \neq S_N$
$p_{lc(S)}^{(P)} = 0, \forall\ P$

2. For every consecutive ‘$P \geq 0$
$p_{lc(m)}^{(P+1)} = {}^{min}{}_{j}[p_{lc(j)}^{(P)} + p_{lc(jm)}]$

The route from source node ‘S_N’ to ith node stops with the cost of link from node ‘j’ to node ‘i’.

For step 2 iteration with ‘H = Q’, and for every sink node ‘m’ this technique analyzes and compares the routes from ‘S to m’ of length ‘Q + 1’ with the obtained route in pervious iteration. If the previous route is shorter, it retains this path as the path with low cost. Else the new obtained path i.e. ‘Q+1’ is employed from source to destination. Thus, this route is of distance ‘Q’ from ‘S’ to other node say ‘j’ with addition of a hop directed from ‘j’ to node ‘m’. Also, it will maintain the route information with the source node till route from source to destination is finalized.

Example:

Fig. 1 represents an underwater wireless sensor network deployed in a certain geographical area, in which a source node ‘S_N’ and destination node ‘D_N’ is deployed at ground. However intermediate nodes i.e. A, B, C and E are deployed in under water.

Let the source node wants to communicate with the destination node through these underwater intermediate sensor nodes. Initially, the source node interacts with its neighboring nodes which are one hop away from it and makes that path permanent whose link cost is low in comparison with the other neighboring nodes at hop 1. However, the source has the node and link information of all its neighboring nodes at hop 1. This process of finalizing the path at various hops will continue till completion of communication between end nodes.

Operation at hop 1: As per the model description there are two neighboring nodes (i.e. A and B) of source node ‘S_N’, Table I represent the path and link cost at hop 1. The other nodes at hop 1 are not accessible, so can be represented by ‘∞’.

The link cost of node ‘B’ is less as compared with node ‘A’. Hence node ‘B’ will be treated as permanent node for path calculation. However, the path and link cost information will be available with the source node till finalization of the path is done between end nodes.

Operation at hop 2: Table II represent the path and link cost at hop 2. In this stage there are three neighboring nodes (i.e. B, E and C), so the possible paths at this hop will be three (see Table II).

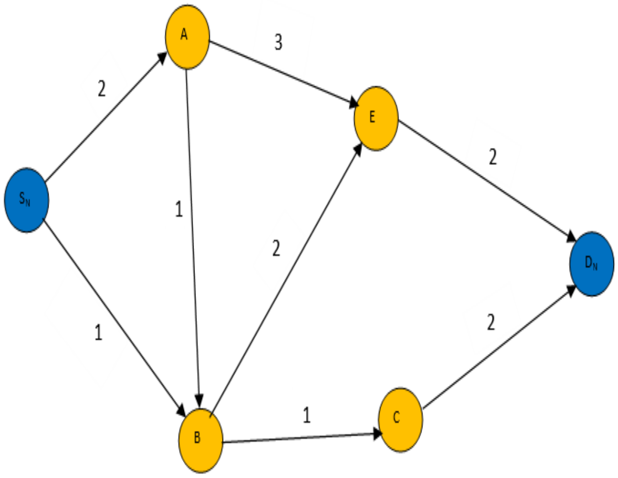


Fig. 1. Wireless Sensor Network.

TABLE I. PATH AND LINK COST AT HOP 1

S. No	Path	Link Cost
1	S_N- A	2
2	S_N- B	1

TABLE II. PATH AND LINK COST AT HOP 2

S. No		Path	Link Cost
1		S_N- A- B	3
2		S_N- A - E	5
3		S_N- B - C	2
4		S_N- B - E	3

Out of these three paths the proposed technique will select S_N-B-C path and make the node ‘C’ as permanent node, as the link cost of this path is having less value as compared to the other two. Also, the other routes information will be available with the source node. Again, the destination node is not accessible hence can be represented by ‘∞’.

Similarly, the operation at hop 3 and hop 4 can be represented in Tables III and IV, respectively.

At this stage (i.e. at hop 3 as per the Table III) the given technique will choose ‘S_N- B – C – D_N’ path, as destination is achieved directly. Though the destination node is reached in hop 3 the other hops in the path estimation is evaluated so as to give the rank to all the possible paths from source to destination.

Again, the pervious path will be chosen as the optimum path because the link cost is less between end nodes. However, in case of failure of path due to unavailability of bandwidth or occurrence of congestion at the node, it will choose the route which will have the next least link cost. If the two paths have same link cost, it will choose the route which has less involvement of nodes as it will have less nodal delay. So due to less involvement of nodes and choose of least cast path between end nodes, the consumed energy is reduced. Also, in case of failure of the route the next optimized path information is available with the preceding node that reduces further nodal time (i.e. propagation, queuing, transmission and processing time). Which intern saves energy of the network and enhances the lifetime of the node in terms of energy consumption. The order (prioritization) of the paths from source to destination is represented in Table V.

The estimation of congestion occurrence and bandwidth unavailability is discussed in next section and accordingly the path from source to destination is finalized.

TABLE III. PATH AND LINK COST AT HOP 3

S. No	Path	Link Cost
1	S_N- A- B - E	5
2	S_N- A – E – D_N	7
3	S_N- A- B - C	4
4	S_N- B – C – D_N	4
5	S_N- B – E –D_N	5

TABLE IV. PATH AND LINK COST AT HOP 4

S. No	Path	Link Cost
1	S_N- A- B – E – D_N	7
2	S_N- A – B - C – D_N	6

TABLE V. PRIORITIZATION OF PATHS

S. No	Priority	Link Cost
1	S_N- B – C – D_N	4
2	S_N- B – E –D_N	5
3	S_N- A – B - C – D_N	6
4	S_N- A – E – D_N	7
5	S_N- A- B – E – D_N	7

B. Establishment of Alternate Route in Case of Failure

To get an alternate route Dharmaraju et al. [24] proposed a framework to guarantee the QoS (quality of service) routing. This framework makes use of INSIGNIA and TORA [25] to obtain multiple routes between end users. To get the QoS routing the work is subdivided into two types which are:

- Feedback based on coarse method
- Feedback based on class method

1) Coarse method: This method fails to provide the QoS, if a node has insufficient bandwidth available to transfer the information between end nodes or due to occurrence of congestion at a node. In this case a given node sends admission control failure (ACF) information to the upstream node. This node (upstream node) then selects the next optimum route to guarantee the QoS. The operation of this method is explained by considering the following example.

Example:

Fig. 2 represents the application of coarse feedback technique in wireless sensor network. As per the given method let the route created by directed acyclic graph (DAG) available be the shortest path i.e. S_N – B- C- D_N.

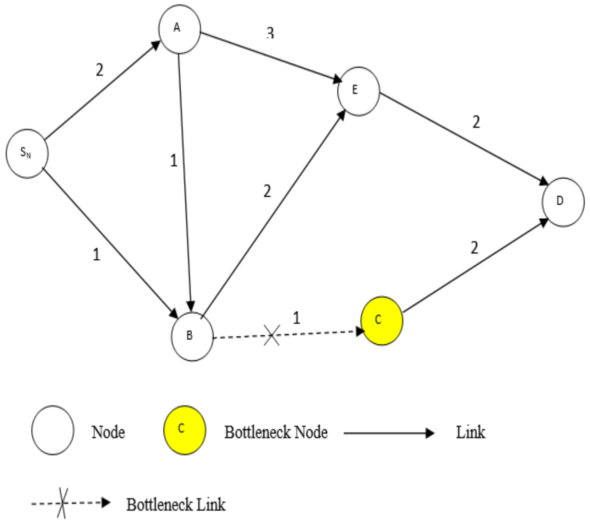


Fig. 2. Coarse Feedback.

Assume node ‘C’ may not be able to admit the data flow due to unavailability of resources. So, it sends an ACF (Authentication control function) from node ‘C’ to node ‘B’, then node ‘B’ checks the feasible path among its neighbors. Here node ‘E’ is the only available node that can forward the data flow towards the destination node. Thus, the feasible path available which can guarantee the QoS is ‘S_N– B – E – D_N’.

Let node ‘E’ fails to receive the data flow, it will also send an ACF information to node ‘B’ and it will send ACF message to source node ‘S_N’ and source node makes use of another neighboring node and try to finalize the route. If source node will not get any path that can guarantee the QoS, it will simply reject the flow.

2) Class method: In this type the period between $(Min)_{B.W}$ and $(Max)_{B.W}$ is divided into ‘X’ classes, where $(Min)_{B.W}$ and $(Max)_{B.W}$ are the bandwidths required to generate the flow. Let the source node is ready to transmit the information towards sink node and the transmitted data flow is of class ‘r (r< X)’. Consider a wireless sensor network as shown in Fig. 3 and the path created by DAG is ‘S_N – B- C- D_N’.

The node ‘B’ accepts the data flow with ‘r’ class effectively and node ‘C’ accepts the data flow whose bandwidth lies in class ‘p’ (p < r) only. At this stage node ‘C’ transmits Admission Report information (AR (p)) to upstream node i.e. ‘B’ to indicate that it has the ability to consider only ‘p’ bandwidth that can be accepted by it. To solve this issue node ‘B’ divides the data flow at a ratio of ‘p to r – p’ and transmits the flow to node ‘C’ and node ‘E’ in the given ratio. The ‘r’ class node is divided into two flows, if ‘E’ node will give the class ‘r – p’ as requested by node ‘B’. The two flows, one with the bandwidth of ‘p’ class having path ‘S_N – B- C- D’. However, if node ‘E’ accepts only class ‘h (h < r- p)’, it transmits an AR (h) information to node ‘B’.

If the other neighbors of node ‘B’ are not able to provide the ‘r’ class facilities, it sends AR (p+h) to source node as it has the ability to provide class services of (p + r). Then the source node finds another anchoring node which can provide the facility to accept the flow of class (r – (p +h)), however if source node will not find any such node, it simply rejects the flow.

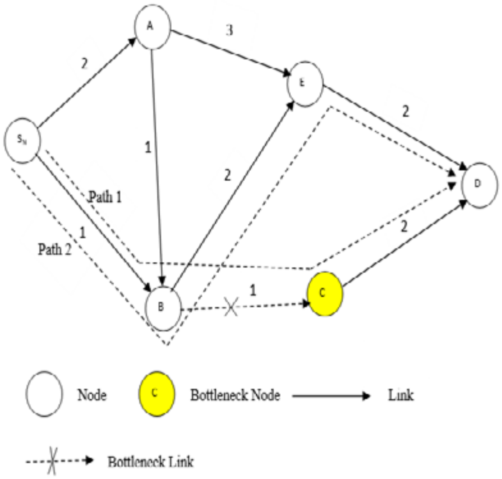


Fig. 3. Class Feedback.

C. Security Against Black Attacks

Due to autonomous structure sensor, the network possesses a well-known security threat called Black Hole Attack. Several models have already been proposed e.g. [26, 27] to detect this attack using distribution of keys. However, due to autonomous structure key distribution is difficult as nodes can move in and out of the network at any time instant. Assignment of keys is favorable in the network with static nodes and may not be effective for the network with mobile nodes. In this paper, a black hole attack is detected using hamming and hex coding technique and is excluded from any route (from source to destination) at any hop in the network. As the coding is kept simple it will not increase the complexity and reduce delay. Let Fig. 4 represents the wireless sensor environment with a black hole in the network.

In the presented approach the code at source is binary hex value of decimal '1' (i.e. 1= 0001). Therefore, the hamming bit positions will be at '2^{n}' where 'n = 0,1 and 2 in this case', as we are considering only 4-bit code at source.

Hamming bit positions (HP) P1 P2 D P4

Source code (SC) 0 0 0 1

However, if one can increase the code length, hamming bits will also increase. The security code at various hops can be generated using compliment to hamming bits with respect to hop count. So, the compliment bits at various hops can be represented as

$$B_{ch} = \bar{P}_{(2)^n} \qquad (1)$$

Where ch = hop number and 'n =0, 1and 2' for hop 1, 2 and 3 respectively. So, the security code-word at hop '1' will be "1001". Similarly, the security codeword bits at hop 2 and 3 are generated by complimenting remaining parity bits (one at each hop) as given in eq. (1). Table VI shows the security codes of hops 1, 2 and 3. After '3^{rd}' hop the security code repeats.

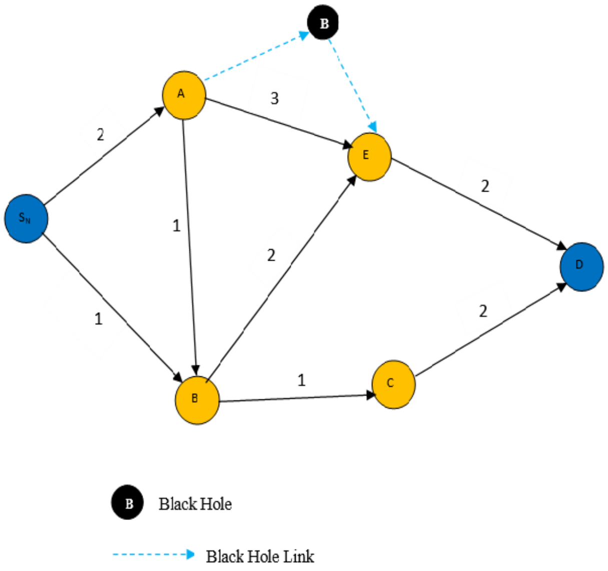


Fig. 4. Wireless Sensor Network with Black Hole Attack.

TABLE VI. SECURITY CODE AT VARIOUS HOPS

Source code		0 0 0 1			
Hamming Parity bits		P_1 P_2 D P_4			
S. No	Hop count	Security code			
1	1	1	0	0	1
2	2	1	1	0	1
3	3	1	1	0	0

1) *Node matching process:* Let the source node is ready to transmit the information to destination node through intermediate nodes. So at hop '1' node 'A' and node 'B' are the only nodes which can take part in the network, as both the two nodes are active and knows the security operation at hop '1', therefore can be able to generate the security code at this hop and match with the source node and source node can transmit the information to both the two nodes.

Now at hop '2' nodes 'C, B and E can take part in the network. Out of the three nodes, node 'B' cannot judge the security code at this hop because it doesn't know the security operation at this particular hop and cannot match with the code word of the upstream node. Therefore, can be easily detected and will not allow taking part with the network nodes. Similarly, at other hops this process of node matching and removal of black hole will continue till the information reaches the destination.

IV. Simulation Results

The stimulation tool which is used to validate the proposed approach is Network Simulator. The network used is multi-hop, protocol used is MAC and the number of nodes considered in the network is 250 with node mobility above 0.2 m/s. The simulation parameters are briefly discussed in Table VII. The presented approach is compared with the Faheem, Zhiping and Meiju approaches as they are the recent methods proposed in this field.

Fig. 5 represents the variation of energy with respect to the nodes. From the figure, it is observed that the overall energy consumed by the presented approach is less when compared with the other techniques as only the QoS route is considered which minimizes the energy consumed, hence optimizes the energy of the nodes that can take part in the route.

Fig. 6 represents the variation of loss rate with respect to the no of packets sent. It is clear from the figure that the presented approach has less packet loss rate. As it utilizes the bandwidth efficiently to avoid the path in which congestion may occur. Because it selects the next alternate optimum path between end nodes to avoid packet loss.

The variation of overhead with respect to the number of hops is represented in Fig. 7 with least packet overhead of the presented approach in comparison with the conventional techniques as the congestion is avoided which minimizes the overhead.

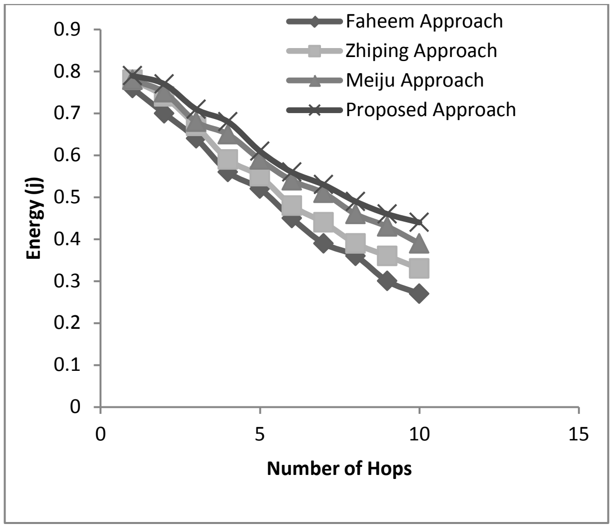


Fig. 5. Energy Saved Versus Number of HOPS.

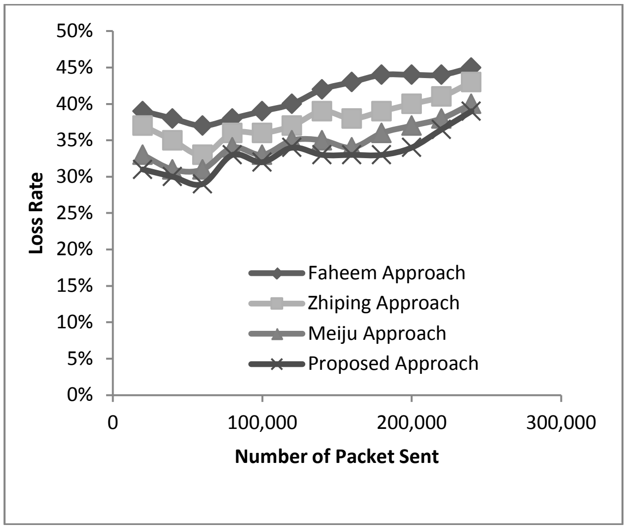


Fig. 6. Loss Rate Versus Packet Sent.

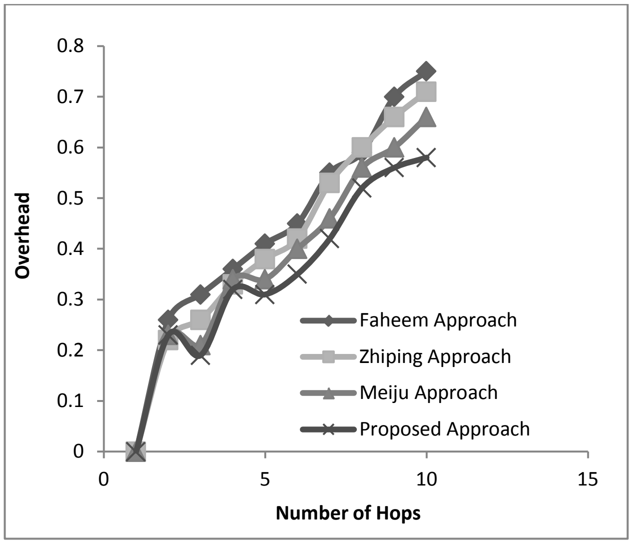


Fig. 7. Overhead Versus Number of HOPS.

TABLE VII. SIMULATION PARAMETERS

S.No	Parameters	Value
1	Underwater Monitoring Area	400m x 400m x400m
2	Maximum speed of node mobility	3m/s
3	Minimum speed of node mobility	0.2m/s
4	MAC	8.11
5	Number of black holes	23
6	Node communication radius (mts)	230
7	Packet size (bits)	512
8	Packet header (bits)	100
9	Initial energy of sensor node (j)	10

V. CONCLUSION

The presented approach is simple and effective for the research community to enhance their work for underground water sensor network. The presented approach considers multiple QoS parameters for choosing a route from source to destination. This approach also ranks the paths in case of any failure occurred due to unavailability of bandwidth or congestion at any node. Security against black hole attack is also an enhancement to the proposed approach.

REFERENCES

[1] N. Ismail, L. A. Hussein, and S. H. S. Ariffin, “Analyzing the performance of acoustic channel in underwater wireless sensor network (UWSN)”, in Proceedings of the Asia Modeling Symposium 2010: 4thInternational Conference on Mathematical Modelling and Computer Simulation, AMS2010, pp. 550–555, Malaysia, May 2010.

[2] K. Awan, P. Shah, K. Iqbal, S. Gillani, W. Ahmad, and Y. Nam, “Underwater Wireless Sensor Networks: A Review of Recent Issues and Challenges”, Wireless Communications and Mobile Computing Volume 2019, Article ID 6470359, pp 1-20 https://doi.org/10.1155/2019 /6470359

[3] T. Cinar and M. Orencik, “An underwater acoustic channel model using ray tracing in ns-2”, in Proceedings of the 2009 2nd IFIP Wireless Days (WD 2009), pp. 1–6, Paris, December 2009.

[4] M. Ayaz and A. Abdullah, “Underwater wireless sensor networks: Routing issues and future challenges”, in Proceedings of 7thInternational Conference on Advances in Mobile Computing and Multimedia, MoMM2009, pp. 370–375, Malaysia, December 2009.

[5] L. Liu, S. Zhou, and J. H. Cui, “Prospects and problems of wireless communication for underwater sensor networks”, Wireless Communications and Mobile Computing, vol. 8, no. 8, pp.977–994, 2008.

[6] G. Zaibi, N. Nasri, A. Kacouri, and M. Samet, “Survey of temperature variation effect on underwater acoustic wireless transmission”, ICGST-CNIR Journal, vol.9, pp 1-6, 2009.

[7] X.-P.Gu,Y.Yang and R.L.Hu, “Analyzing the performance of channel in Underwater Wireless Sensor Networks (UWSN)”, in Proceedings of the 2011 International Conference on Advanced in Control Engineering and Information Science, CEIS2011, pp. 95– 99, China, August 2011.

[8] N. Li, J. Martínez, J. Chaus, and M. Eckert, “A survey on under water acoustic sensor network routing protocols”, Sensors, vol.16, no.3,414,2016.

[9] M. Faheem, G. Tuna, and V. Gungor, “QERP: Quality of Service (QoS) Aware Evolutionary Routing Protocol for Underwater Wireless Sensor

Networks", IEEE Systems Journal, vol 12, issue 3, pp 2066-2073, sept. 2018.

[10] C. Zidi, F. Bouabdallah, and R. Boutaba, "Routing design avoiding energy holes in underwater acoustic sensor networks", Wireless Communications and Mobile Computing, vol. 16, no.14, pp.2035–2051, 2016.

[11] I. Akyildiz, D. Pompili, and T. Melodia, "Underwater acoustic sensor networks: research challenges", AdHoc Networks, vol.3, no.3, pp. 257–279, 2005.

[12] I. Akyildiz, D.Pompili, and T. Melodia, "Challenges for efficient communication in underwater acoustic sensor networks", SIGBED Review, vol.1, no.2, pp.3–8, 2004.

[13] C. Peach and A. Yarali, "An Overview of Underwater Sensor Networks", in Proceedings of the routing techniques regarding network layer, pp.31–36,2013.

[14] N. Javaid., N. Ilyas., A. Ahmad, et al. "An efficient data gathering routing protocol for underwater wireless sensor networks", Sensors. Vol 15(11), pp 29149–29181 (2015).

[15] N. Ilyas, T. Alghamdi, M. Farooq, B. Mehboob, A. Sadiq, U. Qasim, Z. Khan and N. Javaid, "AEDG: AUV-aided Efficient Data Gathering Routing Protocol for Underwater Wireless Sensor Networks,", Procedia Computer Science, vol. 52, no. 1, pp. 568-575, Jun 2015.

[16] Jing, L., He, C., Huang, J., et al.: Energy management and power allocation for underwater acoustic sensor network. IEEE Sens. J. vol 17(19), pp 6451–6462 (2017).

[17] Zhiping Wan. Shao jiang Liu. Weichuan Ni. Zhiming Xu' "An energy-efficient multi-level adaptive clustering routing algorithm for underwater wireless sensor networks," clustering computing, vol 22, pp 14651-14660, March 2018.

[18] Meiju Li, Xiujuan Du, Xin Liu, and Chong Li, "Shortest Path Routing Protocol Based on the Vertical Angle for Underwater Acoustic Networks," Journal of Sensors Volume 2019, Article ID 9145675, pp 1-14, https://doi.org/10.1155/2019/9145675.

[19] M. Awais, I. Ali, T. Alghamdi, M. Ramzan, M. Tahir,M. Akbar, N.Javaid, "Towards Void Hole Alleviation: Enhanced GEographic and Opportunistic Routing Protocols in Harsh Underwater WSNs", in IEEE Access, vol. 8, pp. 96592-96605, 2020, doi: 10.1109/ACCESS. 2020.2996367.

[20] Adil Khan,Mukhtaj Khan, Sheeraz Ahmed, MohdAmiruddin Abd Rahman, Mushtaq Khan, "Energy harvesting based routing protocol for underwater sensor networks," PLOS ONE | vol 14(7) https://doi.org/10.1371/journal.pone.0219459 July 17, 2019.

[21] Pan Feng, Danyang Qin, Ping Ji, Min Zhao, Ruolin Guo and Teklu Merhawit Berhane, "Improved energy-balanced algorithm for underwater wireless sensor network based on depth threshold and energy level partition" EURASIP Journal on Wireless Communications and Networking (2019) 2019: vol 228, pp 1-15 https://doi.org/10.1186/s13638-019-1533-y.

[22] S. Butt, K. Bakar, N. Javaid et al., "Exploiting layered multipath routing protocols to avoid void hole regions for reliable data delivery and efficient energy management for IoT enabled underwater WSNs," Sensors, vol.19, no.3, article no.510,2019.

[23] S. Ashraf, A. Raza, Z. Aslam, H. Naeem and T. Ahmed, "Underwater Resurrection Routing Synergy using Astucious Energy Pods", Journal of Robotics and Control, vol 1, pp 173-184 (2020)

[24] D. Dharmaraju, A. R. Chowdhury, P. Hovareshti, and J.S Baras, "INORA. A unified signaling and routing mechanism for QoS support in Mobile Adhoc Networks," proceedings of ICPPW 2002. pp 86-93.

[25] C.R. Lin and M. Gerla, "Real time support in multihop wireless networks," ACM/ Baltzer Wireless Networks Journal, vol 5, no 2, pp 125-135,1999.

[26] W. A. Xiong and Y. H. Gong, "Secure and Highly Efficient Three Level Key Management Scheme for MANET', Proc. of WSEAS Transactions on Computers, vol. 10, pp. 6-15, 2011.

[27] M. Celestin, S. Vigila, and K. Muneeswaran, "Implementation of text-based cryptosystem using Elliptic Curve Cryptography", Prof. of 1st Int. Conf. on Advanced Computing, vol. 9, pp. 82-85, 2009.

Application of Homomorphic Encryption on Neural Network in Prediction of Acute Lymphoid Leukemia

Ishfaque Qamar Khilji[1], Kamonashish Saha[2], Jushan Amin Shonon[3], Muhammad Iqbal Hossain[4]
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Abstract*—Machine learning is now becoming a widely used mechanism and applying it in certain sensitive fields like medical and financial data has only made things easier. Accurate Diagnosis of cancer is essential in treating it properly. Medical tests regarding cancer in recent times are quite expensive and not available in many parts of the world. CryptoNets, on the other hand, is an exhibit of the use of Neural-Networks over data encrypted with Homomorphic Encryption. This project demonstrates the use of Homomorphic Encryption for outsourcing neural-network predictions in case of Acute Lymphoid Leukemia (ALL). By using CryptoNets, the patients or doctors in need of the service can encrypt their data using Homomorphic Encryption and send only the encrypted message to the service provider (hospital or model owner). Since Homomorphic Encryptions allow the provider to operate on the data while it is encrypted, the provider can make predictions using a pre-trained Neural-Network while the data remains encrypted all throughout the process and finally sending the prediction to the user who can decrypt the results. During the process the service provider (hospital or the model owner) gains no knowledge about the data that was used or the result since everything is encrypted throughout the process. Our work proposes a Neural Network model which will be able to predict ALL-Acute Lymphoid Leukemia with approximate 80% accuracy using the C_NMC Challenge dataset. Prior to building our own model, we used the dataset and pre-process it using a different approach. We then ran on different machine learning and Neural Network models like VGG16, SVM, AlexNet, ResNet50 and compared the validation accuracies of these models with our own model which lastly gives better accuracy than the rest of the models used. We then use our own pre-trained Neural Network to make predictions using CryptoNets. We were able to achieve an encrypted prediction of about 78% which is close to what we achieved when validating our own CNN model that has a validation accuracy of 80% for prediction of Acute Lymphoid Leukemia (ALL).

***Keywords*—CryptoNets; neural network; Acute Lymphoid Leukemia (ALL); homomorphic**

I. Introduction

We are trying to make a system where there will be an assurance about privacy and will also give an initial prediction i.e. whether the patient has ALL (blood cancer) or not. This will also decrease the cost of the system because the initial tests are expensive and in our model the price will be less to give an initial prediction. This system can be used in case of banks, hospitals and other sectors. In our model we included Homomorphic encryption as mentioned earlier. In this system, it will allow one party to have a public key such as in hospitals where a lot of patients can send their data through the public key which will be encrypted and stored in local servers (cloud used in future works). The owner, in our case the hospital administration, lab technicians, doctors and patients can have policies to decrypt the data when necessary. This will ensure the encryption and decryption in a proper manner and will also ensure proper privacy of the user if they want to store or export their information. In the encryption process, the owner will only have the private key and will be able to decrypt the data, on the other hand, the service provider does not have any key and hence will not be able to decrypt the data and thus they won’t know about the data inside or be able to get any information about the predicted data. This will provide a better privacy and will also decrease the overall cost since there is only one private key.

Existing works of running machine learning models on encrypted data include Grapel et al. [7], where they propose confidential algorithms for binary classification based on polynomial approximations to least-squares solutions found by a small number of gradient descent steps. They show experimental validation of the confidential machine learning pipeline and discuss the give and takes involving computational complexity, prediction accuracy and cryptographic security. Zhan et al. [8] works say that their paper considers how to conduct k-nearest neighbor classification in the following scenario: multiple parties, each having a private data set, want to collaboratively build a k-nearest neighbor classifier without disclosing their private data to each other or any other parties. They intend to develop a secure protocol for multiple parties to carry out the desired calculation. All the parties take part in the encryption and in the calculation involved in learning the k-nearest neighbor classifiers. Qi &Atallah, [9] say that they use techniques to also solve the general multi-step k-NN search, and describe a particular expression of it for the case of sequence data. The protocols and correctness evidence can be extended to cope with other privacy-preserving data mining tasks, like classification and outlier detection. Aslett et al. [10][11] propose modified algorithms in application of extreme random forests, involving a new cryptographic stochastic fraction estimator, and naïve Bayes, involving a semi-parametric model for the class decision boundary, and demonstrate how they are useful in learning while predicting from encrypted data. They also exhibit that these methods perform competitively on several different classification data sets and provide detailed information about the calculative practicalities of these and other FHE methods.

In our project, the unencrypted data will be first used to train the neural network. The training data sets can be difficult to find for these types of projects because they always have a privacy issue and also, it's not easily available. After training and validation, our Neural Network model can be used to give secure predictions regarding ALL to the target user. The problem of such type is also called Privacy Preserving Data Mining (Agrawal & Srikant) [23]. To come to the internal concept of our project, we used CryptoNets where we use Homomorphic Encryption on our own Neural Network to secure ALL predictions. This is also a work on running Neural Network and machine learning algorithms on encrypted data but what we have done is a practical implementation of the method in finding the secure predictions of a life threatening disease which is the first of its kind in terms of applying Neural Networks and Machine learning algorithms on encrypted data.

We propose to make a Privacy Preserving Neural Network model which can predict Blood Cancer as well as maintain the privacy of the patient. In our research, at first, we have taken a blood cancer dataset and successfully ran it on various Neural Network and Machine learning models which would accurately predict Acute Lymphoid Leukemia. The results are then compared amongst them. Moreover, we then made our own Neural Network model which is run on the dataset that we are having which is modified at first in order to run on an encryption application. The results are again compared with that of the previous models to prove our NN model is better than the others here. The model is then encrypted and HE (homomorphic encryption wrapper) is implemented on it to do computations and give predictions in a secure format. We are in the process of having our own dataset collected from different labs which we kept for future work. We want to provide a system that will not only give the initial result of whether it is cancer or non-cancer but will also be encrypted and the result will only be known by the patient with the private key which will ensure privacy.

Our objectives include detection of Blood Cancer (Leukemia) from imagery test samples after proper modification in order to run on the custom CryptoNets application. A homomorphic encryption scheme on the whole system which would be used to homomorphically encrypt the images from the Neural Network on which computations and predictions can be done even if the images are encrypted. Comparative analysis is done among the first several models run and then between them and our own NN model. The results are then compared. Work on CryptoNets is done currently in mainly 3 datasets: MNIST, CIFAR-10 and Caltech-101. CryptoNets has not been used in practical applications before. Thus, our contribution in detecting blood Cancer using imagery in a privacy preserving model (CryptoNets) will be the first of its kind. The process that we introduce will pave a way for implementations in various fields. This will ensure secure lives and provide customer satisfaction.

II. Background

A. Literature Survey

1) CNN Features: Shafique and Tehsin [1] used pre-trained AlexNet and fine-tuning to classify ALL subtypes on ALL-IDB augmented with 50 private images. Rehman et al. [2] used a pre-trained AlexNet and fine-tuning to classify ALL subtypes on a private dataset of 330 images. On the other hand, Vogado et al [3] used different pre-trained CNNs as fixed feature extractors to classify ALL on ALL-IDB. Amongst all these, the most informative ones are selected using PCA and classification is performed with an ensemble of MLP, random forest and SVM.

2) Handcrafted Features: Mohapatra et al. [4] and Madhloom et al. [5] use private dataset and classify using an ensemble of SVM, KNN, Naïve Bayes and a KNN classifier. Putzu and Ruberto [6] classify a number of features such as, compactness area and ratio between cytoplasm and the nucleus with an SVM using ALL-IDB. In the above case, the dataset used is small compared to others and also tough to compare the results. The private datasets are unavailable and the public ALL-IDB datasets are given on their own evaluation procedures. All these factors make comparisons difficult.

Our project is divided into two parts of the programming languages Python and C#. The Neural network model building and comparisons of the ML and NN models are done in the python part of the project. The encryption part after that where the "CryptoNets" application created is done on C#. Grapel et al. [7] suggested a use of homomorphic encryption for machine learning algorithms where they focused on finding the algorithms where training can be done over encrypted data and hence were forced to use a learning algorithm where the training algorithm can be expressed in a low degree polynomial. Zhan et al. [8]; Qi &Atallah, [9] looked up for nearest neighbor divisions but they do not give the same level of accuracy as neural networks. Aslett et al. [10][11] presented both of the algorithms such as naïve Bayes classifiers and random forests but their model cannot work efficiently in recognizing objects in images.

B. Homomorphic Encryption

Homomorphic encryption algorithms that require one operation, such as addition, have been known for decades, such as for the ones based on the RSA or Elgamal cryptosystems. But a homomorphic encryption method that allows an infinite number of two operations, i.e. addition and multiplication, allows the computation of any circuit and thus a complete solution of homomorphic (FHE) is gained. FHE was first presented in Gentry [12]. In Gentry, the data encrypted in the bits and for each bit in the message, a separate ciphertext is produced. It is a sort of addition and multiplication module represented by Boolean circuits with XOR and AND gates. FHE in ciphertexts contain some inherent noise which grows during homomorphic encryption and it cannot be decrypted when it gets too large. To solve this problem, Bootstrapping is used where the ciphertexts are constantly refreshed and their noise is reduced [13][14]. The parameters for Practical Homomorphic Encryption (PHE) should be chosen which would not only increase the efficiency but also preserve privacy and ensure security. In our project, we have implemented tools such as Noise Growth Simulator and Automatic Parameter Selection Module to help the user to achieve maximum performance [15]. Somewhat homomorphic encryption approaches can only evaluate a multiple but limited number of addition and multiplication activities. SWHE schemes refer to encryption systems that present certain homomorphic characteristics but lack full homomorphic capacity. The fully homomorphic encryption supported an arbitrary number of multiplications and additions, and hence compute any form of function on encrypted information. For all forms of computations on the information warehoused in the cloud, FHE must be embraced because it allows execution of operations on encrypted records without decryption. As such, the usage of FHE is a crucial step in enhancing cloud-computing security.

C. Encoding

As described above, there is a discrepancy between the atomic structures in neural networks (real numbers) and the atomic structures in the homomorphic encryption schemes (polynomials in $\boldsymbol{R^n}_t$) [16]. An encoding scheme will map each other in a manner that preserves the operations of addition and multiplication. Such a scheme of encoding can be constructed in several ways. For example, real numbers can be converted to fixed precision numbers, and then their binary representation can be used to convert them into a polynomial with the binary expansion coefficients. This polynomial will have the property of returning the encoded value when evaluated at value 2. Another alternative is to encode as a constant polynomial the fixed number of precisions. This encoding is simple, but in the sense that only one polynomial coefficient is being used may seem inefficient. One problem with the scalar encoding is that when homomorphic operations are performed, the only coefficient of the message polynomials grows very rapidly.

D. Encoding Large Numbers

As we have already explained, in this encryption scheme, a major challenge for computation is to prevent the coefficients of the plaintext polynomials from overflowing, **t**. These forces us to pick large values for t, which allows the noise to grow faster in the cipher texts and reduces the total amount of noise tolerated (with q fixed). Therefore, for security reasons, we need to choose a larger q, and then a larger n. One way to overcome this problem partially is to use the Chinese Remainder Theorem (CRT). The concept of using multiple primes is $\boldsymbol{t_1 \ldots t_k}$; given a polynomial $\sum \boldsymbol{a_i x^i}$ we can convert it to k polynomials in such a way that the j-th polynomial is $\sum[\boldsymbol{a_i}(\boldsymbol{mod t_j})]\boldsymbol{x^i}$. Each such polynomial is encrypted and manipulated identically. The CRT guarantees that we will be able to decode back the result, as long as its coefficient does not grow beyond $\prod \boldsymbol{t_j}$. Therefore, this method allows us to encode exponentially large numbers while increasing time and space linearly in the number of primes used.

E. Plaintext Space and Homomorphic Operations

Plaintext elements (messages encrypted by homomorphic encryption schemes) can be represented as a polynomial **ring R**, with coefficients minimalized modulo the integer, **t**. Cipher text elements (encrypted plaintext elements) on the other hand can be similarly represented but instead has coefficients minimalized modulo the integer, **q** [15]. Formally, this means that the plain-text space is the ring **Rt := R/tR = Zt[X]/(Xn + 1)**, and the ciphertext space is contained in the ring **Rq := R/qR = Zq[X]/(Xn + 1)**. However, some of the elements in Rq are invalid ciphertext. A ciphertext created by the function used for encryption in the scheme that we are using encrypts one plaintext message polynomial **m in Rt**. If a homomorphic addition (resp. multiplication) is done on ciphertext that encrypts two plaintext messages for example **m1, m2 in Rt**, the output ciphertext will encrypt the summation of **m1+m2** (resp. the product **m1.m2**). Plaintext element computations are done in the ring Rt. Thus, in case of homomorphic addition, the output ciphertext will encrypt the coefficient wise summation m1+m2, where the coefficients are likewise reduced modulo the plaintext modulus**, t**. In case of homomorphic multiplication, the output ciphertext will encrypt the product **m1.m2 in Rt**, meaning the polynomial will likewise be reduced modulo **Xn+1** where –1 will substitute all powers of Xn and continued till no monomials of **n degree** or higher than that is remaining. Just like homomorphic addition, the coefficients of polynomial m1.m2 will likewise be deducted modulo integer, **t**.

F. Selecting Encryption Parameters

The particular scheme that is used in SEAL is the more practical derivation of the YASHE scheme. Encryption parameters of the scheme are: degree n, the moduli q and t, the decomposition word size w, and distributions Xkey, Xerr. Thus, parameters: = (n,q, t, w, Xkey, Xerr). These parameters are explained in more details below.

- **n**, here is used as the maximum number of terms in the polynomials used for showing the plaintext as well as ciphertext elements. SEAL shows n always as a power of 2. Xn + 1 polynomial is the polynomial modulus, shown as polymodulus in SEAL.
- **q**, the coefficient modulus, is an integer modulus operated in reduction of the coefficients of ciphertext polynomials. SEAL represents q as coeff modulus.
- **t,** the plaintext modulus, is an integer modulus taken in reduction of the coefficients of plaintext polynomials. SEAL shows t, as plain modulus.
- Integer coefficients are decomposed into smaller parts according to the integer base w. The integer calculates the number *w,q:=blogw(q)c+ 1* of parts when decomposing an integer modulo **q** to the base **w.** Practically, we take w, as a power of two, and take the decomposition bit count as *log2w*. SEAL shows *log2w* as decomposition bit count.
- **Xkey** distribution is a probability distribution on polynomials of degree at most n-1 with integer coefficients implemented to sample polynomials with small coefficients that are taken in the key generation procedure. In SEAL, coefficients are sampled uniformly from [1,0,1].

- Likewise, the distribution **Xerr** on polynomials of degree at most n-1 is used for sampling noise polynomials, essential in time of both key generation and encryption. SEAL has the distribution Xerr as a shortened discontinuous Gaussian centered at zero having standard deviation. SEAL has it called Noise Standard Deviation.

G. Algorithms used

The encryption scheme we use is a public-key, homomorphic encryption scheme, and consists of the following algorithms [15]:

- A key generation algorithm **KeyGen (parms)** that, on input the system parameters "parms", generates a public/private key pair **(pk; sk)** and a public evaluation **key, evk**, which is used during homomorphic multiplication.
- An encryption algorithm **Enc(pk;m)**, that encrypts a plaintext , m, using the public **key, pk**
- A decryption algorithm **Dec (sk; c),** that decrypts a cipher text, c, with the private **key, sk**.
- A homomorphic addition operation **Add (c1; c2)** that, given as input encryptions c1 and c2 of m1 and m2, outputs a ciphertext encrypting the **sum, m1 + m2**
- A homomorphic multiplication operation **Mult (c1; c2)** that, given encryptions c1 and c2 of m1 and m2, outputs a ciphertext encrypting the **product, m1. m2**

H. Neural Network Models used

The term Neural Network is an artificial network which is composed of circuits or neurons or artificial nodes. These are leveled circuits and in layers and are usually found in an order where the last layer is the input layer and the first being the output layer. Each layer consists of nodes and they all are incorporated with a value of the features of the project. In these layers, the above or previous nodes of the layer compute a function based on the nodes of the layers under it and the first node in the stack becomes the output layer.

On pre-trained CNN models as well as SVM (Support vector machine) models of our own. The CNN models that we used include VGG16 and VGG19, AlexNet and ResNet. After running these models with the mentioned dataset, we compared the accuracies (both train and test accuracies).

1) VGG16 and 19: In VGG16 architecture, the images are passed through a sequence of convolutional layers which are of fixed size (224x224 RGB image). Thus, we use the default image size for this model in our dataset. In one of the configurations, it also utilizes a 1×1 convolution filter. The convolution stride is fixed to 1 pixel. Spatial pooling is carried out by five max-pooling layers, which follow some of the convolutional layers (not all the convolutional layers are followed by max-pooling). Max-pooling is performed over a 2×2 pixel window, with stride of 2. There are three fully connected layers which have different depths in different architectures. Amongst them, the first two have 4096 channels, and the third performs 2-way classification of the Leukemia dataset and contains two channels for each individual class and the last layer is a soft-max layer. This configuration is the same in all the networks. We are using pre-trained VGG16 and VGG19 models of ImageNet dataset. Thus, in building our own VGG16 model we use the "Weights" of ImageNet. We then extract features of our dataset that are used through VGG16 and VGG19 convolutional base. After the feature extraction, the data then passes through the layers described above (VGG 16 and VGG 19). The models are then fitted and trained for 100 epochs.

2) SVM: Supervised Vector Machine (SVM) is a supervised machine learning algorithm which divides the dataset into two classes and is mostly used for classification and regression purposes. In order to train a linear support vector machine, the machine learning approach is used. We can use K-fold cross-validation where we can estimate error of our mode. Since this will be used, we can enlarge our training data by concatenating the train and the validation sets. After the feature extraction using the convolutional base of VGG16, the output tensor [2] is used in the model fitting of the SVM model. Thus, no separate feature extractions of the pre-processed images that are used are required. The model is run for 100 epochs. Lastly, we ensure that the SVM classifier has one hyper parameter which is a penalty parameter C of the error term.

3) AlexNet: Classifying the image is a major problem and AlexNet fixes it by taking the input image of one of 1000 different groups and generally giving output of a vector of 1000 numbers. There are two groups here instead of 1000 so an output vector of only two will be present. The sum of all output vector elements is 1. AlexNet takes an RGB image size 224x224 input picture from the preprocessed dataset. Nevertheless, unless the image is not in RGB or in grayscale, it is converted to RGB by replicating the single channel in order to get a 3 channel RGB picture. AlexNet has 3 Fully Connected Layers and 5 Convolutional Layers**.**

- Multiple Convolutional Kernels: Multiple convolutional kernels are also many times called filters that extract the necessary features out of an image where the single convolutional layers consist of multiple similar size kernels.
- The first two Convolutional layers: The third, fourth and fifth layers of convolution are joined directly. After the fifth convolutional layer comes an Overlapping Max Pooling layer, whose output goes through a sequence of two fully integrated layers. The second fully integrated layer feeds heuristic SoftMax labelling into two classes.
- Max Pooling layers: The depth is kept unaltered by sampling the sample's height and width. Overlapping Max Pool layers are compared to Max Pool layers, other than neighboring windows where the max is estimated to overlap. Makers of the model used to pool 3x3 size windows between opposite windows, with two steps. This overlapping complexity of pooling has

helped to lower the top -1 error rate by 0.4 percent, the top-5 error rate by 0.3 percent, compared to using non-overlapping 2x2 sized pooling windows with step 2, giving identical output dimensions.

AlexNet's use of the nonlinearity function within the layers is an important feature. Activation functions of sigmoid or Tanh functions used to be the traditional method of training a neural network model. AlexNet has displayed that deep CNNs can be trained much more rapidly using ReLU's nonlinearity feature rather than using saturated activation functions such as tanh or sigmoid. Until feeding the data into the layers and constructing the model, various techniques such as image mirroring, shuffling and random cropping of images in data augmentation to minimize overfitting. This is stated earlier in this section, in which the data set explanation is present.

In dropout, one neuron with a probability of 0.5 is removed from the network. If a neuron is lost, this does not lead to propagation which is either forward or backward. Thus, each input goes through different architecture of the network due to which the learned weight parameters are therefore more robust, and are not readily overfitted. There is no dropout during testing, and the entire network is utilized, but output is scaled by a factor of 0.5 to adjust for the neurons lost during training. Dropout raises the number of iterations required to converge by a factor of 2 but AlexNet will significantly overfit without it.

4) ResNet50: Above is a pre-trained model of the ResNet50 architecture. The model has "50" layers with weights. Residual Networks or ResNet creates networks through models known as residual models and also known as the degradation problem. Although increasing depth increases the accuracy of the network, the problem increases when the vanishing gradient arises. Another issue that occurs while training the deeper network is greater training error as it adds the layers when performing optimization on large parameter space. The architecture of ResNet is identical to that of VGGNet which has 3x3 filters. The ResNet50 model we will use is a pre-trained model trained on the dataset ImageNet.

III. Proposed Model

A. Dataset Pre-Processing and Feature Selection

In 2018, another dataset with in excess of 10,000 preparing pictures and a separate test set of ordinary B-lymphoid forerunners and threatening B-lymphoblasts has been discharged as an online test open to the general population. In 2019, it was made available for general use [24]. The enormous size of this new dataset permits to make improved classifiers dependent on profound neural systems and furthermore gives an increasingly dependable correlation of contending approaches. In this work we present our way to deal with the arrangement of sound and dangerous cells on the referenced dataset utilizing a convolutional neural system. The test dataset [17,18,19,20,21], from now on alluded to as C_NMC dataset, contains pictures of white platelets taken from 154 individual subjects, 84 of which show ALL. Table I gives a nitty gritty breakdown of the quantity of subjects and cells in preparing and test sets. The dataset is imbalanced with about twice the same number of ALL cells as ordinary cells. Each picture has a goal of 450 × 450 pixels and contains just a solitary cell as a result of preprocessing steps applied by the dataset creators: A mechanized division calculation has been utilized to isolate the cells from the foundation. Every pixel that was resolved not to be a piece of the cell is hued totally dark. In any case, since the division calculation isn't great, there are examples where parts of the cell are coincidentally shaded dark or pointless foundation is incorporated. Moreover, the sum total of what pictures have been preprocessed with a stain-standardization system that performs white-adjusting and fixes blunders acquainted due with varieties in the recoloring compound [17]. See Fig. 1 for instance pictures from the dataset.

Table I shows Composition of the dataset. At the time of writing the ground truth for the final test set is not yet released, so some information is missing.

Despite the fact that the dataset contains in excess of 10,000 pictures, a few information enlargement strategies can be applied to build the measure of preparing information further and improve the preparation of our convolutional neural system. Since tiny pictures are invariant to flips and turns, we perform level and vertical flips with 50% likelihood each and pivots with an edge from [0, 360] degrees picked consistently at irregular. Since convolutional neural systems with pooling tasks or walks bigger than one are not flawlessly interpretation invariant, we additionally perform arbitrary interpretations of up to 20% of each side-length in flat and vertical ways. Also, the pictures are further focus trimmed to 100 × 100 pixels to diminish the dimensionality of the information. This will for the most part make learning a classifier quicker and simpler. Despite the fact that the editing disposes of huge pieces of the picture, it has no impact on the arrangement exactness in light of the fact that without a doubt, not very many cells are really bigger than this harvest. Much of the time, pictures that are not totally dark outside of the harvest are division disappointments that incorporate pieces of the foundation. The dataset is further trimmed, labeled and pre-processed into CIFAR-10 format so that we can run our CryptoNet model with ease. This part is explained further in the coming section.

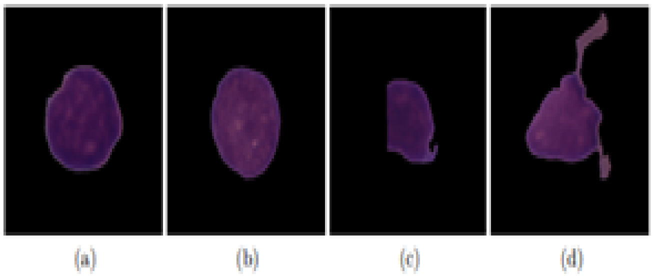


Fig 1. Images in the Training Set. (a) ALL cell (b) Normal cell (c), ALL cell with Part of the Cell Cut Off Due to an Imperfect Segmentation (d) Normal cell with Superfluous Background Due to an Imperfect Segmentation.

TABLE I. TRAIN AND TEST SUBJECTS AND THE CORRESPONDING NUMBER OF SAMPLES

Dataset part	ALL subjects	Normal Subjects	ALL cells	Normal cells
Train	47	26	7272	3389
Preliminary Test	13	15	1219	648
Final test	9	8	?	?

B. *Model Description*

According to the workflow diagram illustrated previously in Fig. 2, firstly the C_NMC Challenge 2019 dataset is modified, pre-processed to CIFAR-10 format, split into training and test and taken in numpy arrays accordingly. The conversion of the dataset to CIFAR-10 format is essential because previously CryptoNets model has been run on mainly three datasets, namely, Cifar-10,MNIST and Caltech-101 as mentioned earlier of which Cifar-10 is much more convenient in dealing with real-life image classification and has an organized "labeling" along with "classes" of images in binary format, all of which are convenient in running the CryptoNets application using the SEAL version 3.2 HE-wrapper in C and .NET framework version 4.6.2 [16].

The conversion of the dataset to numpy array and using it to train our own cancer predicting Convolutional Neural Network, generating encryption parameters and conversion of test samples to binary version of CIFAR-10 are done prior to building the CryptoNets wrapper around it is done using code of python version 3.5.

1) *Dataset Conversion and taking into Array*

- After the pre-processing has been done; our 10,000 training images are at first separated equally and placed into two different folders with names: "Cancer" and "Normal".
- From each class sub folder, we are taking 80% of the images for training and 20% of the images testing. After placing the images, the class subfolders and the images inside the folder are iterated accordingly. An array is first created with dimensions of 32x32 images and an RGB value of "3". Thus, the shape of the array would be (32,32,3). For each class subfolder, each image in the subfolder is sliced to obtain the "R", "G" and "B" values which are then into that array that are concatenated as iteration is done over each image. The array is then appended.
- For the "index" value, a separate array is declared. Each class folders in the input directory would correspond to an image label. Thus the "index" value is assigned to each class folder namely "0" for "Cancer" and "1" for "Normal". Each class folder is iterated for images inside and the assigned "index" value is appended into an array for each iterated image in the subfolder.
- The above steps are repeated for another class subfolder.
- The above steps are repeated for the rest 20% of the training images. The test and train image arrays and the corresponding test and train image labels are saved in variables "X_train,Y_train" and "X_test, Y_test". Since the label numpy array is being iterated and concatenated within the same loop as the same array, one-hot encoding is not necessary here. But we are doing it anyway just to be on the safe side. Thus, numpy arrays are then one-hot encoded where input, that is, list of a ground truth table where "0" is Cancer and "1" is Normal. Thus, the image data taken in the test and train arrays are in Cifar -10 format as with each image taken in "X" the corresponding "Y" label is inserted in the arrays accordingly.

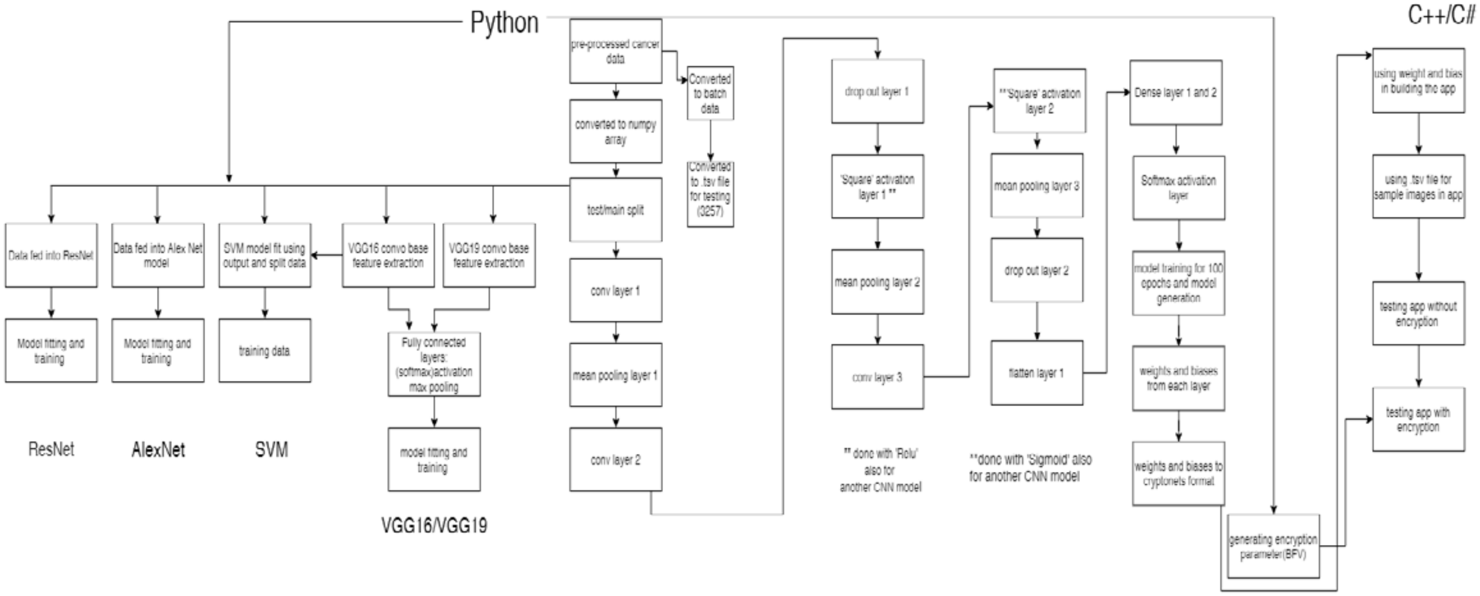

Python
C++/C#
pre-processed cancer data
Converted to batch data
converted to numpy array
Converted to .tsv file for testing (3257)
test/main split
conv layer 1
mean pooling layer 1
conv layer 2
Data fed into ResNet
Model fitting and training
ResNet
Data fed into Alex Net model
Model fitting and training
AlexNet
SVM model fit using output and split data
training data
SVM
VGG16 convo base feature extraction
VGG19 convo base feature extraction
Fully connected layers: (softmax)activation max pooling
model fitting and training
VGG16/VGG19
drop out layer 1
'Square' activation layer 1 **
mean pooling layer 2
conv layer 3
** done with 'Relu' also for another CNN model
**'Square' activation layer 2
mean pooling layer 3
drop out layer 2
flatten layer 1
**done with 'Sigmoid' also for another CNN model
Dense layer 1 and 2
Softmax activation layer
model training for 100 epochs and model generation
weights and biases from each layer
weights and biases to cryptonets format
generating encryption parameter(BFV)
using weight and bias in building the app
using .tsv file for sample images in app
testing app without encryption
testing app with encryption

Fig 2. Overview of Proposed Approach.

2) Model Details

Our Neural Network model has 14 layers in total of which 3 are “Convolutional”, 3 are “Activation”, 2 “Dense” layers, 1 “flatten” layer before the output layer and the rest are “Mean Pooling” and “Dropout” layers.

The model is put into training for 100 epochs. Our own model is set to training using a different set of “Activation” layer functions twice. At first training, the first 2 “Activation” layers are “Relu” layers and the last “Activation” layer being a “Sigmoid” layer. The model is again trained this time with “Square” function instead of “Relu” and “Softmax” instead of “Sigmoid”.

Below are the descriptions of the “Activation” functions mentioned.

Sigmoid: Take the value of one of the nodes in the feeding layer and evaluate the function

z ↦1/ (1+exp (-z))

Rectified Linear: Take the value of one of the nodes in the feeding layer and compute the function

z ↦max(0,z)

Square Activation Layer: This layer squares the value at each input node.

Softmax Layer: This activation function forces the values of output neurons to take values between zero and one, so they can represent probability scores.

“Sigmoid” and “Relu” activation functions are non-polynomials. The fix was to estimate these functions with low-degree polynomials but here we will be using a different method [15]. We tried to manipulate the trade-off between possessing a non-linear transformation required by the learning algorithm and also need to maintain the degree of the polynomials minimal to make the parameters of homomorphic encryption realistic. We opted to use the non-linear lowest degree polynomial function, which is the Square function: **sqr (z):=z^2**. It has been suggested by a theoretical study of a problem regarding neural networks with polynomial activation functions and dedicated the majority of their study to the square activation function [22]. For the training stage, the sigmoid activation function is used to get reasonable terms of error when running the gradient descent algorithm. However, in the encrypted world, we don’t have a reasonable way to deal with the sigmoid. Fortunately, once we have our weights set and would like to make predictions, we can just take it out. This is because the neural network’s prediction is given by the index of its output vector’s maximum value, and since the sigmoid function is increasing monotonously, whether we apply it or not will not affect the prediction.

The validation accuracies for both the times are recorded. For the first time the accuracy is recorded to be 78% and the second time it is recorded to be 80%.

3) Converting Weights and Biases to CryptoNets Format

Once the model is training the next step is to convert the weights and bias vectors to a format that CryptoNets recognizes. CryptoNets expects the weights to be in a CSV file where the weights for each layer are in a separate line. One challenge is to collapse the immediate previous or next linear layers into a single linear layer. For each layer with trainable weights (a dense layer or a convolution layer) a bias file and a weights file should be generated. Once done for all the relevant layers, we combine all the weights into a one file and all the biases into a second file. Below is the code snippet of how the “weights” and “biases” of the “Convolutional” and “Dense” layers are obtained as a separate file. A total of 10 files (5 for weights and 5 for biases) are generated for the 3 “Convolutional” and 2 “Dense” layers. Values in the files are now in single columns. Thus, each column in each file of all the weights and biases for each layer is transposed into single rows. All the “weights.csv” and “bias.csv” files are combined to a single “all_weights.csv” and “all_bias.csv” file

4) Building and Testing the Application without Encryption

The model is first tested without any encryption parameters. Prior to that, the “test.tsv” file is created in python. At first we a create “.bin” file similar to the binary version of the CIFAR-10 dataset for our test samples of the cancer dataset which had been trimmed, pre-processed and put into folders with labels “0” and “1” in order to work with CryptoNets like the Cifar-10 dataset. The test samples of the cancer dataset are thus arranged accordingly. The “.bin” file hence is a batch file created containing a binary version of the 3527 test samples arranged in bytes in the .bin file. The model is first tested without any encryption parameters. Prior to that, the “test.tsv” file is created in python. At first we a create “.bin” file similar to the binary version of the CIFAR-10 dataset for our test samples of the cancer dataset which had been trimmed, pre-processed and put into folders with labels “0” and “1” in order to work with CryptoNets like the Cifar-10 dataset. The test samples of the cancer dataset are thus arranged accordingly. The “.bin” file hence is a batch file created containing a binary version of the 3527 test samples arranged in bytes in the .bin file. The “.bin” is then converted to “.tsv” file where should have one line per image where each line contains 1 + 33232 tab separated columns in which the first column is the label and the other column are the RGB values of a 32*32 image. The bytes in the “.bin” file is converted to strings when converting to “tsv”. This is done using C#.

The application is coded in C# using “Visual Studio 2019” and was tested in the windows environment used .Net framework version 4.6.2. This project depends on SEAL version 3.2. Thus a “Nuget” package containing SEAL, is added as a reference which is essential. The “all weights” and “all biases” are passed in the “WeightsReader” function and the parameters are loaded. The string file is passed into the application. The project is then built in x64 architecture in release mode.

Prior to “building” the project, the line of code:

```
var Factory = new
RawFactory((ulong)batchSize);
```

is added. The use of the “RawFactory” function is explained further.

5) *Selecting Encryption Parameters*

The theoretical process and mathematical formulae to calculate the correct parameters are given in the previous section "Parameter Selection". To allow correctness the parameters should support large enough numbers to be processed. Much like in traditional programming where a program might fail if numbers are allocated with insufficient space (short integers vs. long integers or floats vs. doubles), the same thing may happen when using homomorphic encryption. Thus, the first step is to determine the amount of space needed. When running without encryption (using the RawFactory), CryptoNets keys track of the size of number processes in the line of code:

```
Console.WriteLine("Max computed value {0}
({1})", RawMatrix.Max,
Math.Log(RawMatrix.Max) / Math.Log(2));
```

We print the maximum number used (in absolute value) and the number of bits this number required to encode this number. To determine the number of bits needed, we add 1 to this number since an additional bit is required to hold the sign of the number.

To provide the required number of bits, a number of prime numbers is provided such that the product of these numbers is at least the required number of bits. For example, if 70 bits are needed, we can use 2 prime numbers with 35 bits each. Working with more prime numbers increases the running time. However, smaller primes allow more computation to be done before the noise budget exceeds.

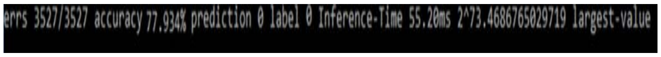
Noise budget is another important parameter of Homomorphic Encryption. In a nut-shell, a freshly encrypted number has a certain amount of noise budget. Every operation on such numbers (addition, multiplication, etc.) reduces this budget. Once this budget equals zero, the decryption will fail to provide correct results. The amount of noise budget available is determined by several parameters, the most important of them are the dimension used. (N) and the size of the prime numbers used as plaintext-modulus. The dimension N should be a power of two, the larger it is, the greater the noise budget is. However, the larger N is, the slower the program runs. Typical values for "N" range from 2^12 to 2^15. On the other hand, a greater noise budget is available when the plaintext modulus is smaller. However, working with smaller plaintext modulus requires using more plaintext modulus to achieve the required number of bits and therefore slows down the application. Selecting a good set of parameters is currently done manually.

After determining the required number of bits, select a value for N and the number of primes to be used. 3 parameters are specified to generate the encryption parameters that are to be passed in the application. The code in python 3 generates these parameters in the code, 3 parameters are set where "bits" is the minimal number of bits of each prime, "ndegree" is the number of bits in N and "count" is the number of primes to generate. The code above generates parameters of 957181001729 and 957181034497.These parameters are passed into the application and the line of code for the CryptoNets build:

```
var Factory = new EncryptedSealBfvFactory(new ulong[]
{ 957181001729, 957181034497 }, 16384);
```

where 16384 is the value of "N". Since 2 prime numbers were demanded with 39.8 bits each, these parameters can support 79.6 bits.

The following is an output for a prediction sample generated after the CryptoNets model is run is as follows:



```
errs 3527/3527 accuracy 77.934% prediction 0 label 0 Inference-Time 55.20ms 2^73.4686765029719 largest-value
```

Fig 3. Output for a Prediction Sample.

Here in Fig. 3, label "0" is correctly predicted with an accuracy of 77.934% at an inference time of 55.20 ms.

IV. Experiments and Result Analysis

Each model mentioned earlier in the paper is trained on a PC of GTX 750ti, 8gb Ram and a processor of core i5 4th generation. Each model is trained for100 epochs except for AlexNet and ResNet which are trained for approximately 20 epochs since they are better CNN models with more convolutional layers and training them for more epochs may result in "overfitting". The training and validation accuracies of the models are illustrated below:

From Fig. 4, the VGG-16 model is trained for 100 epochs. The training accuracy increases at a decreasing rate whereas the validation accuracy decreases but is very much fluctuating. At 100 epochs approaching, both the accuracies tend to become constant.

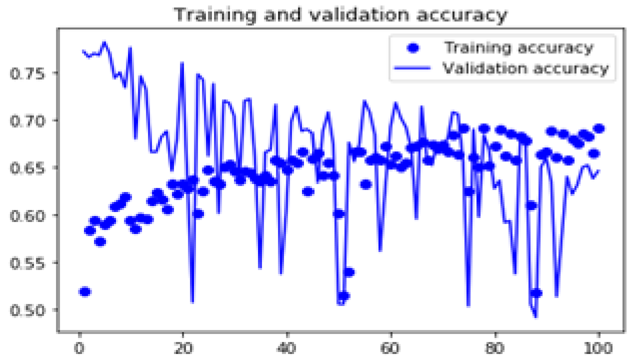


Fig 4. Training and Validation Accuracy for VGG16.

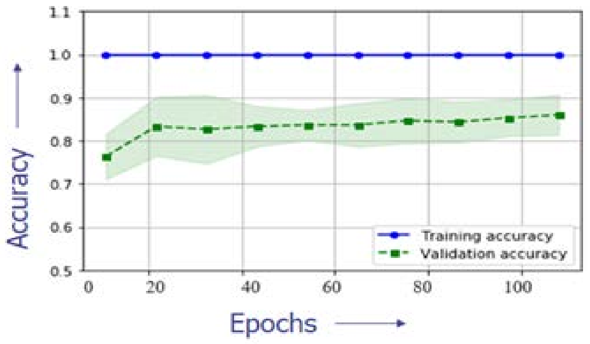


Fig 5. Training and Validation Accuracies for SVM.

It can be seen from Fig. 5, that the training accuracy is always constant at 100% which is practically unrealistic in terms of machine learning. Hence, it can be stated that this is due to overfitting of the data and we should not take this result into account.

The AlexNet model is trained for 20 epochs as depicted by the graph in Fig. 6. After 15 epochs, we see that the training accuracy is approximately 73% which is higher than the steady increasing validation accuracy of 68%. The model thus is not over-fitting. Both the model's training and testing accuracy increases at a decreasing rate.

The ResNet50 model is trained for 20 epochs. The graph depicts the Validation and Training accuracies of the model after 15 epochs Fig. 7. We see that the training accuracy is approximately 72.50% which is higher than the steady increasing validation accuracy of 67.80%. The model it seems is not overfitting. The model's training accuracy increases at a decreasing rate but the validation accuracy remains constant.

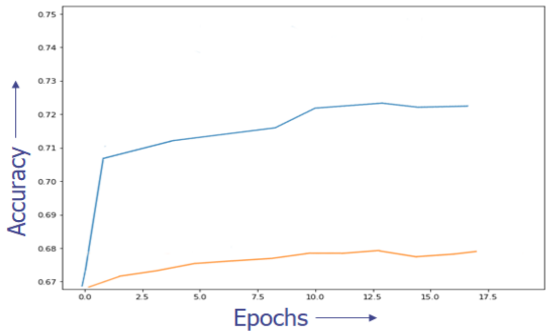


Fig 6. AlexNet Validation (Orange) and Train Accuracy (Blue).

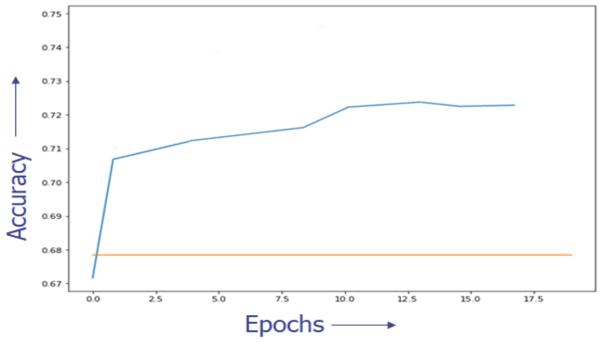


Fig 7. ResNet50 Validation (Orange) and Train Accuracy (Blue).

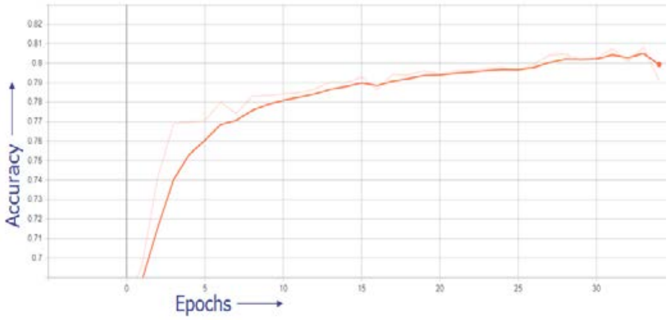


Fig 8. Validation Accuracy using Square and Softmax.

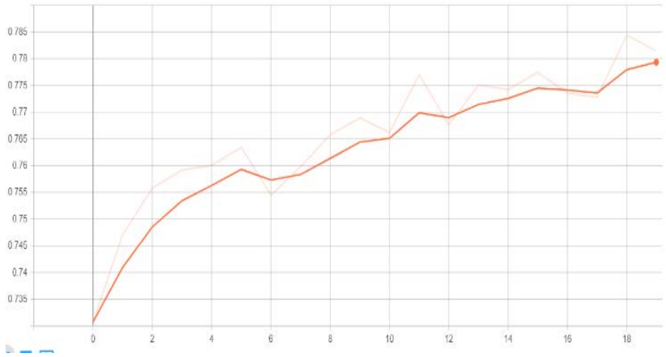


Fig 9. Validation Accuracy using Relu and Sigmoid.

Our own Neural Network model is defined as above and is trained for 100 epochs. Like mentioned earlier our model is first fitted using the "Relu" function in the first two "Activation" layers and "Sigmoid" function in the last "Activation" layer and the Neural Network is trained. The same process is repeated using the "Square" function instead of "Relu" and "Softmax" instead of "Sigmoid". The graph of the validation accuracy of our own model using different sets of functions twice is illustrated in Fig. 8 and 9. The graphs were obtained from Tensorflow. Although the models with different functions are trained for different numbers of epochs, they are trained with the same dataset. Thus, there won't be much of a difference in accuracy.

The model with the "Square" and "Softmax" activation functions have higher test or validation accuracy of 80% than the previous AlexNet and Resnet models when compared and also has more validation accuracy than that when the other two functions are used to build our own model (Fig. 9).

Table II shows the comparison between all the other models.

From Table II, we see that SVM has the most validation accuracy. It is surprising how an ML model had performed better than the rest of the Neural Network models. This may be due to "over-fitting" of the model after put into training taking the output tensor of the convolutional base of VGG16 into the model for feature extraction. The VGG19 model also works the same way except that there are differences in layers. Since we have included the work of VGG19 in our workflow diagram, our implementation on this will be for future works.

TABLE II. COMPARISON BETWEEN OUR MODELS

	AlexNet	ResNet50	VGG16	SVM	CNN (our model)
Training accuracy	72.90%	72.50%	68.20%	100%	82.6%
Validation Accuracy	68%	67.80%	64.80%	86%	80%
Encrypted Neural Network (Accuracy)					77.934

V. Conclusion

From the earlier validation accuracies of our current model using different "Activation" functions and taking the validation accuracies of the models into account, the model with the "Square" and "Softmax" activation functions have higher test or validation accuracy of 80% than the previous "AlexNet" and "ResNet" models when compared and also has more validation accuracy than that when the other two functions are used to build or own model. The prediction accuracy of our encrypted CNN model (77.934%) is slightly less than that of the un-encrypted CNN model (80%). This may be due to the noise generation which should reduce if correct encryption parameters are selected. In our future work, after creating the CryptoNet model, the model with the data will be stored in the cloud and hence the cloud can charge money for the storage and will also be financially beneficial for both the user and the supplier. The cloud system does not have any key and hence will not be able to decrypt the data and hence it won't know about the data inside or be able to get any data about the predicted data. This will provide a better privacy and will also decrease the overall cost and since there is only one private key. The secure predictions of Acute Lymphoid Leukemia (ALL) can thus be carried out through the cloud and the particular patient can access the corresponding results with ease.

According to our literature survey and our previous research, it can be seen that there are several works which used several machine learning and Neural Network algorithms in classification of Acute Lymphoid Leukemia, however our approach was different and we were able to attain a high accuracy while encrypting our dataset and using our CNN model.

Moreover, the CryptoNet model that we implemented here is currently based on The Brakerski/Fan-Vercauteren (BFV, 2012) scheme from the built in SEAL library. Our future works would also include implementing the CryptoNet model for real life applications using the faster Cheon-Kim-Kim-Song (CKKS, 2016) scheme for better accuracy in the CryptoNet model used. We are currently in the process of developing the algorithm using the CKKS scheme to precisely suit our CryptoNets model and its calculations. Also, we are collecting ALL- Acute Lymphoid Leukemia images with "patient id", "age", and "gender". For now, we have 290 images which is more than the ALL-IDB dataset which is frequently used in detection of blood cancer using ML and NN models. Previous works done on ALL detection used ALL-IDB dataset which has about 270 ALL blood cancer images. As of now, we are using the CNM-C dataset of our model which is significantly larger than the ALL-IDB dataset and has about 10000 training images of which we are using 3257 images for testing. We are hopeful to successfully collect about 2000 images, label it and run it on our own CryptoNets model for secure prediction of Cancer.

Moreover, it will provide a comparatively less expensive preliminary screening and will also ensure the proper privacy of the user.

References

[1] S. Shafique and S. Tehsin, "Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks", Technology in cancer research & treatment, vol. 17, pp. 1-533, 2018.

[2] A. Rehman, N. Abbas, T. Saba, S. I. u. Rahman, Z. Mehmood, and H. Ko-livand, "Classification of acute lymphoblastic leukemia using deep learning", Microscopy Research and Technique, vol. 81, no. 11, pp. 1310–1317, 2018.

[3] L. H. S. Vogado, R. D. M. S. Veras, A. R. Andrade, F. H. D. De Araujo,R. R. V. e Silva, and K. R. T. Aires, "Diagnosing leukemia in blood smear images using an ensemble of classifiers and pre-trained convolutional neural networks", 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE, pp. 367–373, 2017.

[4] S. Mohapatra, D. Patra, and S. Satpathy, "An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images", Neural Computing and Applications, vol. 24, no. 7-8, pp. 1887–1904, 2014.

[5] H. T. Madhloom, S. A. Kareem, and H. Ariffin, "A robust feature extraction and selection method for the recognition of lymphocytes versus acute lymphoblastic leukemia", International conference on advanced computer science applications and technologies (ACSAT) , IEEE, pp. 330–335, 2012.

[6] L.Putzu and C. Di Ruberto, "White blood cells identification and counting from microscopic blood image", in Proceedings of World Academy of Science, Engineering and Technology, World Academy of Science, Engineering and Technology (WASET), p. 363, 2013.

[7] T.Graepel, K. Lauter, and M. Naehrig, "Ml confidential: Machine learning in encrypted data", in International Conference on Information Security and Cryptology, Springer, pp. 1–21, 2012.

[8] J. Z. Zhan, L. Chang, and S. Matwin, "Privacy preserving k-nearest neighbor classification.", IJ Network Security, vol. 1, no. 1, pp. 46–51, 2005.

[9] Y. Qi and M. J. Atallah, "Efficient privacy-preserving k-nearest neighbor search", The 28th International Conference on Distributed Computing Systems , IEEE, pp. 311–319, 2008.

[10] L. J. Aslett, P. M. Esperan ̧ca, and C. C. Holmes, "Encrypted statistical machine learning: New privacy preserving methods",arXiv preprint arXiv:1508.068, 2015.

[11] L. J. Aslett, P. M. Esperan ̧ca, and C. C. Holmes, "A review of homomorphic encryption and software tools for encrypted statistical machine learning",arXiv preprint arXiv:1508.06574 , 2015.

[12] C. Gentry et al., "Fully homomorphic encryption using ideal lattices.", in Stoc, vol. 9, pp. 169–178, 2009.

[13] Z. Brakerski and V. Vaikuntanathan, "Fully homomorphic encryption from ring-lwe and security for key dependent messages", in Annual cryptology conference, Springer, pp. 505–524, 2011.

[14] Z. Brakerski, C. Gentry, and V. Vaikuntanathan, "(leveled) fully homomorphic encryption without bootstrapping",ACM Transactions on Computation Theory (TOCT), vol. 6, no. 3, p. 13, 2014.

[15] N. Dowlin, R. Gilad-Bachrach, K. Laine, K. Lauter, M. Naehrig, and J. Werns-ing, "Manual for using homomorphic encryption for bioinformatics", 2015.

[16] R. Gilad-Bachrach, N. Dowlin, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy", in International Conference on Machine Learning,pp. 201–210, 2016.

[17] R. Gupta, P. Mallick, R. Duggal, A. Gupta, and O. Sharma, "Stain color normalization and segmentation of plasma cells in microscopic images as a prelude to development of computer assisted automated disease diagnostic tool in multiple myeloma",Clinical Lymphoma, Myeloma and Leukemia, vol. 17,no. 1, e99, 2017.

[18] A. Gupta, R. Duggal, R. Gupta, L. Kumar, N. Thakkar, and D. Satpathy, "Gcti-sn: Geometry-inspired chemical and tissue invariant stain normalization of microscopic medical images", 2018.

[19] R. Duggal, A. Gupta, R. Gupta, M. Wadhwa, and C. Ahuja, "Overlapping cell nuclei segmentation in microscopic images using deep belief networks",in Proceedings of the Tenth Indian Conference

on Computer Vision, Graphics and Image Processing, ACM, p. 82.42, 2016.

[20] R. Duggal, A. Gupta, and R. Gupta, “Segmentation of overlapping/touching white blood cell nuclei using artificial neural networks”,CME Series on Hemato-Oncopathology, All India Institute of Medical Sciences (AIIMS). New Delhi,India, 2016.

[21] R. Duggal, A. Gupta, R. Gupta, and P. Mallick, “Sd-layer: Stain deconvolutional layer for cnns in medical microscopic imaging”, in International Conference on Medical Image Computing and Computer-Assisted Intervention,Springer, pp. 435–443,2017.

[22] R. Livni, S. Shalev-Shwartz, and O. Shamir, “On the computational efficiency of training neural networks”, in Advances in neural information processing systems, pp. 855–863, 2014.

[23] Agrawal, R. and Srikant, R. Privacy-preserving data mining. *ACM SIGMOD Record*, 29(2), pp.439-450, 2000.

[24] "Gupta, A., & Gupta, R. ALL Challenge dataset of ISBI 2019 [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/tcia.2019.dc64i46r",2019.

On the Digital Applications in the Thematic Literature Studies of Emily Dickinson's Poetry

Abdulfattah Omar*
Department of English, College of Science & Humanities, Prince Sattam Bin Abdulaziz University, Saudi Arabia
Department of English, Faculty of Arts, Port Said University, Egypt
Correspondence: Abdulfattah Omar, Department of English, College of Science & Humanities, Prince Sattam Bin Abdulaziz University, Al-Kharj, Riyadh, 11942, Kingdom of Saudi Arabia.

***Abstract*—Thematic studies in literature have traditionally been based on philological methods supported by personal knowledge and evaluation of the texts. A major problem with studies in this tradition is that they are not objective or replicable. With the development of digital technologies and applications, it is now possible for theme analysis in literary texts to be based at least partially on objective replicable methods. In order to address issues of objectivity and replicability in thematic classification of literary text, this study proposes a computational model to theme analysis of the poems of Emily Dickinson using cluster analysis based on a vector space model (VSM) representation of the lexical content of the selected texts. The results indicate that the proposed model yields usable results in understanding the thematic structure of Dickinson's prose fiction texts and that they do so in an objective and replicable way. Although the results of the analysis are broadly in agreement with existing, philologically-based critical opinion about the thematic structure of Dickinson's work, the contribution of this study is to give that critical opinion a scientific, objective, and replicable basis. The methodology used in this study is mathematically-based, clear, objective, and replicable. Finally, the results of the study have their positive implications to the use of computational models in literary criticism and literature studies. The success of computer-aided approaches in addressing inherent problems in the field of literary studies related to subjectivity and selectivity argues against the theoretical objections to the involvement of computer and digital applications in the study of literature.**

Keywords—Cluster analysis; digital applications; Emily Dickinson; lexical content; philological methods; thematic studies; Vector Space Model (VSM)

I. Introduction

The analysis of literary texts according to thematic criteria has long been central to literary criticism. There is an established discipline in literary criticism, here referred to as thematic literary criticism (TLC), which studies literary texts in terms of their assessed themes [1-7]. TLC has traditionally been carried out on the basis of philological criteria and/or according to predefine**d** templates or stereotypical classifications. Missing from such studies, however, is any discussion of the issues of objectivity and replicability, or indeed evidence of any awareness that these are issues at all. They are, however, fundamental to all areas of science [8-12]. This study addresses these issues in a literary context in relation to the poetry of Emily Dickinson (1830–1886). The aim is to make some progress towards developing an objective and replicable method for the thematic studies of Emily Dickinson's poetry that can be extended to more general literary classification, overcoming the subjectivity of traditional philological methods. This study builds on work undertaken on Information Retrieval (IR), Automated Text Classification (ATC), and related technologies with the ultimate aim of developing an effective framework for thematic literature studies based on empirical grounds.

In order to identify the thematic structures in the poetry of Emily Dickinson, vector space clustering (VSC) methods are used. VSC is an effective tool for identifying and forming meaningful groups of the objects. The hypothesis thus is that VSC methods can be used in generating an experimentally replicable, objective and conceptually useful analysis based on empirical evidence abstracted from Emily Dickinson's poetry. The remainder of the article is organized as follows. Section 2 is a brief survey of the thematic studies of the poetry of Emily Dickinson. Section 3 describes the methods and procedures of carrying out the computational thematic analysis of the data. Section 4 is analysis and discussions. Section 5 is the conclusion.

II. Literature Review

Different approaches have been developed in the critical study of the thematic structures of literary texts. These include New Criticism, Phenomenology, Structuralism, Deconstructionism, Post-structuralism, Psychoanalysis, Post-Colonialism, Marxism, Feminism, and Historicism [13]. Critics and researchers are usually free to adopt any of these approaches or even adopt their individual style of analysis. In identifying the thematic significance of a given text, a critic may focus on the text, or views it within its larger historical or sociocultural framework. Another critic focuses primarily on economic critique, often exploring how identity is related to social class [14]. Apart from these methodologies, numerous thematic discussions rely heavily either on the author's biographical considerations or even the critic's personal anecdotes, voice, and experience. The problem with such studies is that they are neither objective nor replicable. Regardless of the adopted critical approach, thematic studies of literary works in the philological tradition are in one way or other reflections of the critics' own judgments, which can be affected by personal feelings, emotions, impressions, or prejudices. Moreover, a critic cannot set definite criteria he

*Corresponding Author

Submission Date: May 31, 2020
Acceptance Date: June 13, 2020

used for his classification so that it can be replicated or repeated by another researcher. Even worse, it cannot guarantee that two critics following the same approach. As a result, two readings of a given text can result in completely different interpretations of the same text. It is true thus to suggest that thematic studies in literature have traditionally been based on philological methods supported by personal knowledge and evaluation of the texts [15].

Referring to the literature on the poetic production of Emily Dickinson, thematic studies have been given due attention. Emily Dickinson is one of the most important American poets of the nineteenth century and is considered by many critics as one of America's greatest and most original poets of all time [16]. For many critics, Dickinson's poetry is widely regarded as a milestone in American literature [17]. Dickinson wrote forty volumes of almost 1,800 poems [18, 19]. Many critics argue that Dickinson's poems speak of love, death and nature [20, 21]. One major problem with these studies is that critics have been generally selective in their treatment of the thematic analysis of her poems. They have directed their attention towards particular thematic aspects of her work. For instance, there is a strong body of criticism that confines the works of Dickinson to the subject of death [20, 22-24]. Evidently, many commentators stress the preoccupation morbid in her poetry. Critics generally have focused on her most celebrated works, classified as death poetry. The work on Dickinson is thus best described in terms of its 'selectivity'. Critics have been concerned with particular issues in Dickinson's work and to that end they have tended to select particular pieces of writing for criticism and investigation.

Rommel [25] argues that the problem and limitation of exclusion is accepted as an integral aspect of traditional approaches to textual analysis, and for this reason most literary critics deal with representative textual phenomena when they talk about the surface features of a text. He points out that in the majority of literary critical studies that adopt traditional methods, some kind of textual sampling takes place and critics occasionally make judgments according to the frequency of occurrence or absence of certain textual features. He makes it clear that traditional philological methods are insufficient when dealing with literary texts, since their length makes it too difficult for any traditional approach to measure the frequency of an element efficiently. Rommel concludes that empirical evidence that is truly representative of the whole text is extremely difficult to come by, and mainstream literary scholarship has come to accept this limitation as a given fact. In the face of this limitation, quantitative and computational approaches have been suggested to address the problems of selectivity and lack of objectivity in literary studies. Although these approaches have been most naturally associated with applications related to authorship and style, "they can also be used to investigate larger interpretive issues like plot, theme, genre, period, tone, and modality" [26]. This study seeks to bridge this gap in the literature by looking into computational approaches that address the problems of philological methods of thematic analysis and classification.

III. Methods and Procedures

Recent years have witnessed the development of different computational approaches in document clustering theory. This is a broad framework that includes numerous methods for grouping similar texts together [27-30]. These methods include: vector space clustering (VSC); latent semantic indexing (LSI); concept mining; explicit semantic analysis (ESA); and Netword. The approach that seems most theoretically consistent with our goal, however, is VSC. This is a clustering method whereby texts are clustered into distinct sets based on their semantic similarity [27, 29, 30]. This approach has two steps. Firstly, the relevant documents are mathematically quantified as vectors in high-dimensional space using the vector space model (VSM). Secondly, the similarity between documents is computed using exploratory multivariate analysis (EMVA) methods and hierarchical cluster analysis methods [31, 32]. The rationale is that: (1) the research question directing the present discussion is exploratory, since it is concerned with generating hypotheses about the conceptual structure of Dickinson's corpus; (2) the discussion is concerned with grouping texts of identical/similar themes into distinct sets, which suggests that the idea of analysis is a multivariate data-solving problem [33]; and finally EMVA methods have proved successful in many VSC applications [34]. EMVA encompasses numerous techniques, but for the present purposes cluster analysis is the most appropriate. This is a multivariate mathematical technique for finding relatively homogeneous clusters of cases based on proximity measurements. The rationale of using cluster analysis is that it is the most appropriate technique for organizing a collection. More importantly, cluster analysis methods are used when we do not have any prior hypotheses about the data [29, 35-38]. This serves the principle of objectivity, a primary concern of this research.

In order to support objective and reliable generalizations about Emily Dickinson's poetry, a corpus of all Dickinson's poems (recently collected in *Emily Dickinson's Poems As She Preserved Them* by Cristanne Miller) is built. These are 1775 poems. Dickinson's letters to Susan Gilbert (the woman who was her friend, her muse, mentor, primary reader and editor, fiercest lifelong attachment, and Only Woman in the World) were not included in this study. This study is only concerned with the poetic production of Emily Dickinson. One requirement, however, is that the texts must be pre-processed prior to their representation as data in the corpus. In the current case, the poems were reduced to lists of tokens with only content words retained. That is, function words, like determiners and prepositions, have been removed. 59,378 content-type words were identified in this way, forming the basis for analysis.

Documents were then represented using the vector space model (VSM). This model is both conceptually simple as well as convenient for computing semantic similarities within documents. A data matrix, D*ij*, was built where the rows D*i* represent the documents; the columns D*j* represent the lexical-type variables; and the value of the matrix D*ij* encompasses the frequency of lexical type *j* in document *i*. The data matrix D*ij* was constructed from 59,378 variables representing 1775 poems. As such, each row of the matrix represents a lexical frequency profile for the corresponding text. Because each lexical variable in the profile has a semantic set, the profile gives a representation of what the text is about; what it is not about; and gradations of meaning in-between. However, it should be noted that the matrix 59,378 has some characteristics that could adversely affect the validity of clustering results. Firstly, some poems are long while others are very short. Secondly, its data space dimensionality is so large as to be unwieldy.

To address the variation in text length, the row vectors of the matrix were normalized to compensate for variations in length among texts so that their lexical frequency profiles could be meaningfully clustered. This normalization was related to mean text length using the function:

$$Freq = Freq\ F_i(\frac{\mu}{length\ (i)})$$

The effect of this is to reduce the values in the vectors that represent long documents, while increasing the values of the vectors representing the shorter ones. For documents that are near to or at the mean value, little or no change occurs in the corresponding vectors. The overall effect is to make all the corresponding documents similar in length for the purposes of analysis. As a next step, the problem of high dimensionality was considered. To achieve this, two simple methods of dimensionality reduction were applied. These were: the elimination of relatively low-variance variables; and the retention of highest TF-IDF (term frequency-inverse document frequency) variables.

As shown in Fig. 1, relative variance can now be clearly seen with variables of high variance on the left and variables of low variance on the right. The high-variance variables have to be retained, since they are the main criteria by which the texts can be distinguished from one another. The flat area on the right represents the low-variance variables that contribute little or nothing to distinction between texts—these variables, starting at about 1001 and moving to the right, can be discarded. Variables 1001–59378 were eliminated because of their relatively low variance. The reason for retaining the first 1,000 variables is that these were thought to be the most important for the current analysis. This indicates that a certain amount of subjectivity was at play in determining the number of variables to retain. Finally, TF-IDF was used to identify the most distinctive variables within the dataset. Given that the highest TF-IDF variables are the most important, each column was calculated by means of TF-IDF using the function:

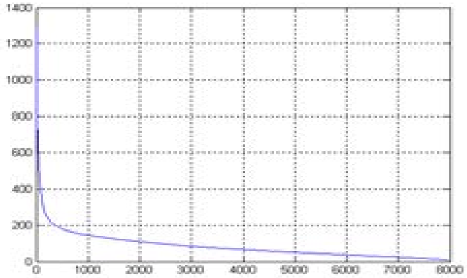


Fig 1. Term Weighting by Variance for the Matrix 59,378.

$$tfid(t_j) = tf(t_j)log_2(\frac{\text{D}}{df_j})$$

Where $tf(t_j)$ is the frequency of *term* t_j across all documents in the data matrix. Using this formulation, the TFIDF of a lexical type A that occurs once in a single document is 1 x $\log_2$ (1000/1) = 9.97; and the TFIDF of a type B that occurs 400 times across 3 documents is 400 x $\log_2$ (1000/3) = 3352. As can be seen in this example, B is far more useful for document differentiation than A, which is more intuitively satisfying than the alternative. The variables are sorted in descending order, as shown in Fig. 2.

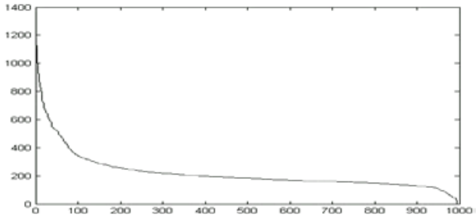


Fig 2. An Illustration of TF-IDF Term Weighting.

As can be seen in the Fig. 2, variables 1–200 were retained and variables 201–1000 were removed. The result is a transformed data matrix 200, which provided the basis for subsequent analysis.

IV. Analysis and Discussions

Agglomerative hierarchical cluster analysis methods were used to find meaningful clusters in the data, which can be used to empirically derive the structure of the thematic concepts of the poetry of Emily Dickinson. The data matrix was hierarchically analyzed first using Ward linkage clustering (or what is usually referred to as the increase in the sum of squares) with the Euclidean distance between points. This is the most suitable method for our analysis because it allows the clearest partitioning of the matrix rows. Ward's method of clustering allows us to discover useful associations and meaningful groupings in the dataset. Hierarchical cluster analyses are presented in the form of diagrams known as dendrograms. These are visual representations of cluster structures that show how clusters are related to each other, which clusters are merged or fused at each stage of the analysis, and how the distance between them is calculated at the time of their merging or fusion [39].

One advantage of this clustering method shown in Fig. 3 is that it offers a solution to a common problem in cluster analysis—how to decide on the optimal number of clusters to fit a dataset. The strong tendency towards left branching associated with other clustering methods is avoided in Ward clustering. The matrix rows are assigned to three main groups, which are assigned as groups A, B, and C. For clustering validation purposes, a cross-validation approach was used. The texts were randomly divided into two subsets, *A* and *B*, and cluster analysis was carried out separately on each. The level of similarity in the results indicates validity [40]. Comparison shows a close fit between the results of hierarchical analysis. There is no contradiction between the results of the two clustering structures.

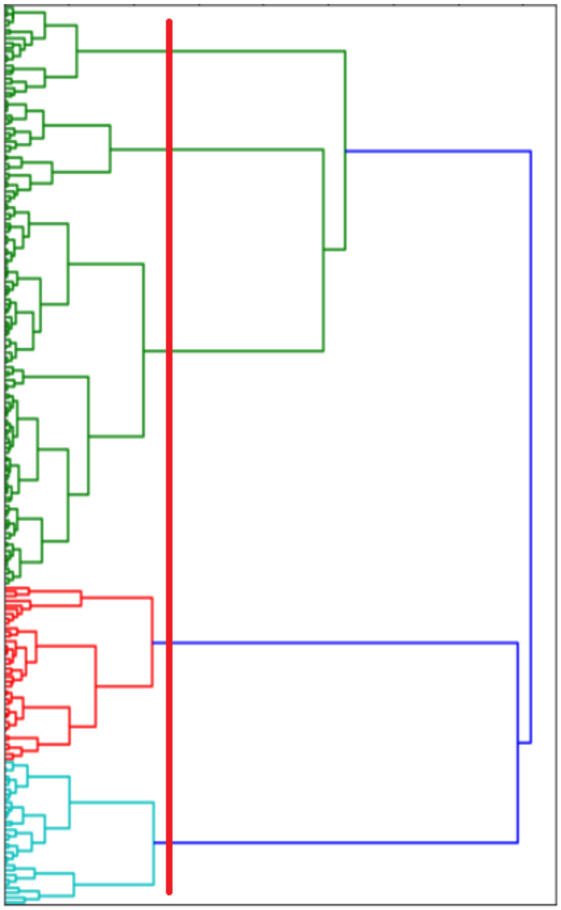


Fig 3. Hierarchical Cluster Analysis of Dickinson's Matrix using Euclidean Distance & Ward Linkage Clustering.

Given that the texts were clustered on the basis of lexical frequency vectors, each cluster has a characteristic lexical frequency profile that distinguishes it from the others. Based on this assumption, it should be possible to identify the most important variables for each group, and, on the basis of the lexical semantics of these items, to infer thematic characteristics of the respective groups. To do this, a centroid analysis was carried out. A centroid is the center of a given geometric figure. Centroid vectors were constructed by means of the vectors in the Dickinson matrix constituting the four groups A-C in accordance with the function:

$$V_i = \frac{\sum_{i=1\ m\ \mathrm{D}ij}}{m})$$

Where:

V j is the jth element of the centroid vector (for j = 1 … the number of columns in D);

D is the Dickinson data matrix, and;

m is the number of row vectors in the cluster in question

The resulting vector groups were compared to show how, on average, the three groups differed for each of the 180 lexical variables. The aim was to identify the variables in which the difference was greatest and the thematic characteristics of each group can then be inferred.

Group A comprises 894 poems including, "Because I could not Stop for Death, It was not Death, for I stood up, I Heard a Fly Buzz when I Died, and I felt a Funeral in my Brain." This group is characterized by words like *die*, *death*, *funeral*, *soul*, *Heaven*, *clergyman*, *Father*, *Christ*, *God*, and *immortality*. These are frequently used in the poems of Group A. Correlating the results of the lexical profiles above with some knowledge about these texts, it can be observed that they are concerned with idea of death.

The most distinctive lexical features of Group B, in turn, are *sea*, *feathers*, *bird*, *storm*, *wild*, *light*, *woods*, *valley*, *world*, *nature*, *dew*, *flower*, *summer*, *shower*, *bee*, *garden*, *Grass* as well as colors names such as *yellow* and *purple*. Poems of this group include, "A Dew Sufficed Itself, A Service of Song, May Flower, My Garden, Summer Shower, The Bee is not afraid, The Grass, The Purple Clover, and The Sea of Sunset." Based on the lexical-semantic features of these words, it can be suggested that they are concerned with nature.

Finally, Group C included 628 poems including, "That I did always love, Heart We Will forget him, I Cannot Live Without You, You Left Me, and I know that he exists." The most distinctive lexical features of this group are *sweet*, *love*, *heart*, *beloved*, and *charm*. It can be suggested thus that these poems are centered on the theme of love as reflected in the lexical-semantics of the words.

It can be finally concluded that the clustering structures identified in this study correspond in principle to the classification of Dickinson's poetry in the philological tradition of literary criticism outlined earlier. It can be claimed however that quantitative and computational approaches to literature provide accurate and acceptable methods of classification and analysis [41]. Furthermore, these approaches, using scientific and objective methodologies, can be used in the service of traditional literary studies to help critics cope with the huge amount of electronic text now becoming available [42, 43].

V. Conclusion

Computational analysis of Emily Dickinson's poetry has yielded a replicable, objective, and conceptually useful thematic structuring of her works. Although the results of the analysis are broadly in agreement with existing, philologically-based critical opinion about the thematic structure of Dickinson's work, the contribution of this study is to give that critical opinion a scientific, objective, and replicable basis. The methodology used in this study is mathematically-based, clear, objective, and replicable. It has been shown to be effective in the literary study of Dickinson's work and is thus potentially applicable in literary scholarship more generally. Quantitative and computational methods can be used to empirically derive taxonomies of thematic concepts of the poetry of Emily Dickinson.

Equally important, nonetheless, computers and machines cannot be replacements for humans in terms of reading and interpreting literature. I suggest that the computational element in literary criticism can develop concrete evidence to support or refute hypotheses or interpretations that have in the past been based on personal readings and the somewhat serendipitous noting of interesting features. In other words, what computational methods give us is an objective clustering giving insight into alternative interpretations based on criteria that are definitely in the text and which constrain our subjective interpretations. This is the main point of this study. It does not claim that this method is better or replaces all human interpretations of literary texts, but rather it constrains human subjective interpretation by presenting classification criteria that must be taken seriously precisely because they are objective and replicable. The clustering results of this study can serve as a base for future studies and criticisms of the thematic analysis of Emily Dickinson's poetry.

Finally, the results of the study have their positive implications to the use of computational models in literary criticism and literature studies. The success of computer-aided approaches in addressing inherent problems in the field of literary studies related to subjectivity and selectivity argues against the theoretical objections to the involvement of computer and digital applications in the study of literature.

Acknowledgments

I take this opportunity to thank Prince Sattam Bin Abdulaziz University in Saudi Arabia alongside its Deanship of Scientific Research, for all technical support it has unstintingly provided towards the fulfillment of the current research project.

References

[1] M. E. Atwood, Survival: a thematic guide to Canadian literature. Toronto: Anansi, 1972.

[2] V. H. Brombert, Novels of Flaubert: A Study of Themes and Techniques. Princeton: New Jersey: Princeton University Press, 2015.

[3] F. Hammill, Canadian literature (Edinburgh critical guides to literature). Edinburgh: Edinburgh University Press, 2007.

[4] M. L. Jockers and D. Mimno, "Significant themes in 19th-century literature," Poetics, vol. 41, no. 6, pp. 750-769, 2013/12/01/ 2013.

[5] W. R. Sanborn, The American Novel of War: A Critical Analysis and Classification System. Jefferson, North Carolina; London: McFarland Incorporated Publishers, 2012.

[6] W. Sollors, The return of thematic criticism (Harvard English studies). Cambridge, Mass.: Harvard University Press, 1993.

[7] T. Todorov, The fantastic : A Structural Approach To A Literary Genre. Ithaca, N.Y.: Cornell University Press, 1975.

[8] B. L. Berg, Qualitative research methods for the social sciences. Boston: Allyn and Bacon, 1998.

[9] I. Holloway, Basic concepts for qualitative research. Oxford: Blackwell Science, 1997.

[10] R. Gomm, Key concepts in social research methods (Palgrave key concepts). Basingstoke: Palgrave Macmillan, 2009.

[11] G. Payne and J. Payne, Key concepts in social research. London: SAGE, 2004.

[12] M. Q. Patton, Qualitative Research & Evaluation Methods, 3rd ed ed. London: Sage, 2002.

[13] L. Tyson, Critical Approaches to Literature. London; New York: Routledge, 2018.

[14] W. Sollors, The Return of Thematic Criticism. Harvard University Press, 1993.

[15] A. Omar, "Addressing Subjectivity and Replicability in Thematic Classification of Literary Texts: Using Cluster Analysis to Derive Taxonomies of Thematic Concepts in the Thomas Hardy," Journal of the Chicago Colloquium on Digital Humanities and Computer Science, vol. 1, no. 2, pp. 1-14, 2010.

[16] B. Steffens, Emily Dickinson. Lucent Books, 1998.

[17] G. Grabher, R. Hagenbüchle, and C. Miller, The Emily Dickinson Handbook. University of Massachusetts Press, 1998.

[18] R. Gray, A Brief History of American Literature. Wiley, 2010.

[19] C. Miller, "Emily Dickinson's Poems As She Preserved Them." Harvard University Press, 2016.

[20] W. Martin, The Cambridge Companion to Emily Dickinson. Cambridge Cambridge University Press, 2002.

[21] N. Tandon and A. Trevedi, Thematic Patterns of Emily Dickinson's Poetry. Atlantic Publishers & Distributors, 2008.

[22] M. Dauben, "Emily Dickinson" - The Death Motif in the Poetry of Emily Dickinson. GRIN Verlag, 2010.

[23] N. Dietrich, Emily Dickinson's Death Poetry. GRIN Verlag, 2003.

[24] B. Lindberg, "The theme of death in Emily Dickinson's poetry," Studia Neophilologica, vol. 34, no. 2, pp. 269-281, 1962.

[25] T. Rommel, "Literary Studies," in ACompanion to Digital Humanities, S. Schreibman, R. Siemens, and J. Unsworth, Eds. Oxford: Blackwell, 2004.

[26] D. L. Hoover, "Quantitative Analysis and Literary Studies," in A Companion to Digital Literary Studies, R. G. Siemens and S. Schreibman, Eds. Malden, MA: Blackwell Publishers, 2013, pp. 517-533.

[27] J. Kogan, Introduction to Clustering Large and High-Dimensional Data. Cambridge: Cambridge University Press, 2007.

[28] C. D. Manning, P. Raghavan, and H. Schütze, An Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.

[29] H. Moisl, Cluster Analysis for Corpus Linguistics. De Gruyter, 2015.

[30] W. Wu, H. Xiong, and S. Shekhar, Clustering and Information Retrieval. Springer 2013.

[31] F. Husson, S. Le, and J. Pagès, Exploratory Multivariate Analysis by Example Using R. CRC Press, 2017.

[32] A. Lüdeling and M. Kytö, Corpus Linguistics (no. v. 2). De Gruyter, 2009.

[33] R. Adams, "Perceptions of innovations: exploring and developing innovation classification," PhD, School of Management Cranfield University, 2003.

[34] M. L. Eaton, Multivariate Statistics: A Vector Space Approach (Institute of Mathematical Statistics. Lecture notes-monograph series). Beachwood, Ohio: Institute of Mathematical Statistics, 2007.

[35] M. R. Anderberg, Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Elsevier Science, 2014.

[36] E. J. Bynen, Cluster analysis: Survey and evaluation of techniques. Springer Netherlands, 2012.

[37] B. S. Duran and P. L. Odell, Cluster Analysis: A Survey. Springer Berlin Heidelberg, 2013.

[38] C. Hennig, M. Meila, F. Murtagh, and R. Rocci, Handbook of Cluster Analysis. CRC Press, 2015.

[39] A. Fielding, Cluster and Classification Techniques for the Biosciences. Cambridge, UK; New York: Cambridge University Press, 2007.

[40] A. C. Rencher, Methods of Multivariate Analysis, Second Edition ed. John Wiley & Sons, INC, 2002.

[41] R. Siemens and S. Schreibman, A Companion to Digital Literary Studies. Wiley, 2013.

[42] C. Mullings, S. Kenna, M. Deegan, and S. Ross, New Technologies for the Humanities. De Gruyter, 2019.

[43] S. Zyngier, M. Bortolussi, A. Chesnokova, and J. Auracher, Directions in Empirical Literary Studies: In honor of Willie van Peer. John Benjamins Publishing Company, 2008.

Situational Modern Code Review Framework to Support Individual Sustainability of Software Engineers

Sumaira Nazir[1], Nargis Fatima[2], Suriayati Chuprat[3]
Razak Faculty of Technology and Informatics[1, 2, 3]
University Technology Malaysia (UTM)
Kuala Lumpur, Malaysia
Faculty of Engineering and Computer Science[1, 2]
National University of Modern Languages (NUML)
Islamabad, Pakistan

***Abstract*—Modern Code Review (MCR) is a socio-technical practice to improve source code quality and ensure successful software development. It involves the interaction of software engineers from different cultures and backgrounds. As a result, a variety of unknown situational factors arise that impact the individual sustainability of MCR team members and affect their productivity by causing mental distress, fear of unknown and varying situations. Therefore, the MCR team needs to be aware of the accurate situational factors, however, they are confronted with the issue of lack of competency in the identification of situational factors. This study aims to conduct the Delphi survey to investigate the optimal and well-balanced MCR-related situational factors. The conducted survey also aimed to recognize and prioritize the most influencing situational factors for MCR activities. The study findings reported 21 situational factors, 147 sub-factors, and 5 Categories. Based on the results of the Delphi survey the identified situational factors are transformed into the situational MCR framework. This study might be helpful to support the individual sustainability of the MCR team by making them aware of the situations that can occur and vary during the execution of the MCR process. This research might also help the MCR team to improve their productivity and sustain in the industry for longer. It can also support software researchers who want to contribute to situational software engineering from varying software engineering contexts.**

***Keywords*—Situations; situational factors; Modern Code Review (MCR); sustainable software engineer; situational software engineering**

I. INTRODUCTION

Software development paradigm shift from conventional to more, complex, and sustainable software [1] demands better software engineering techniques and practices. Software methods that can effectively manage software engineering activities, operations, and maintenance are needed [2]. In the modern age of software development situation-aware software development is highly desirable. Situation-aware software engineering also known as situational software engineering [3] allows software engineers to deal with known situations instead of being unproductive with unknown situations. These situations have a strong influence on the individual sustainability of software engineers as well as on successful software development. In the literature, these situations are termed as *"situational factors"*.

Individual sustainability is one of the core dimensions of sustainable software engineering [4] that refers to psychological well-being, self-respect, and the sustainability of software engineers [5], [6] in the industry. It is argued that individual sustainability of software engineers can be ensured by situational software engineering [5] that can help the software engineers to deal with varying and unknown situations [7]. Situational software engineering by identifying the situational factors [1] boosts the confidence, competence, and productivity of the software engineers.

It is contended that effort has been contributed regarding situational factor identification for software development generally [8] and specifically software requirement engineering [9], [10], [11], however, other software engineering activities need further insight. Presently, the software engineers engaged in Modern Code Review (MCR) are facing the issue of lack of competency in the identification of situational factors [12], [13], [14], [15].

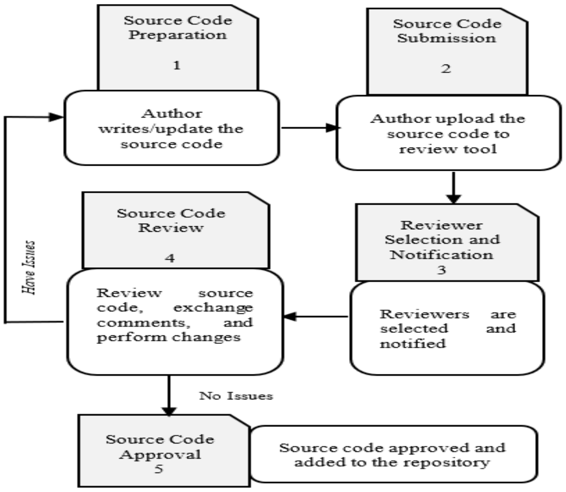

Source Code Preparation
1
Author writes/update the source code
Source Code Submission
2
Author upload the source code to review tool
Reviewer Selection and Notification
3
Reviewers are selected and notified
Source Code Review
4
Review source code, exchange comments, and perform changes
Have Issues
No Issues
Source Code Approval
5
Source code approved and added to the repository

Fig 1. MCR Overview [19].

MCR is a significant light weight [16], [17] software engineering activity. In MCR [18], [19] the reviewer reviews the source code submitted by the author, identifies the issues and provide feedback to the author for source code improvement. The author then performs changes according to the feedback of the reviewer and again submit the source code for review. The reviewed cycle continues till the source code is approved by the reviewer. After the approval, the source code is added to the repository. MCR process is executed with the support of review tools [16], [18], [20] such as Code Flow and Gerrit. The MCR process overview is given in Fig. 1.

Though the existing research [8], [9], [10], [11], [21] has provided attention to situational software engineering by identifying situational factors, however, the existing research has discussed the situational factors' identification for software development generally and specifically for requirement engineering. There are fewer considerations regarding the exploration of situational context for MCR activities to help software engineers' engaged in MCR activities to sustain in the industry for longer [12], [13], [22] lack of such research can cause mental distress of software engineers and ultimately failure of software [6], [12].

No, systematized investigations are available that comprehensively report the situational factors in the context of MCR. Hence, there is a need to recognize situational factors [14], [22] and to develop a comprehensive situational MCR framework to address the issue of lack of competency in the identification of situational factors, to support the individual sustainability of software engineers, and for successful software development.

Therefore, this study aims to conduct the Delphi survey to finalize the list of situational MCR factors, sub-factors, and categories for their practicality, naming, grouping, and sub-grouping. Delphi survey also aimed to identify and prioritize the most influential situational factors for MCR activities. The results obtained after the Delphi survey are used to develop the situational MCR framework to support the individual sustainability of software engineers engaged in MCR activities. This study is an extension of our previous research that is Systematic Literature Review (SLR), to identify the situational factors from literature and expert review to evaluate by the experts the list of situational factors obtained as a result of SLR. The results of SLR and expert review are presented in [13], [23], [24].

The paper is organized as Section II delivers the literature review. Section III highlights the details regarding the methodology employed to conduct the research. In Section IV Delphi survey results are presented. The research discussion is provided in Section V. Sections VI and VII provides the conclusion and work opportunities in the future. The research contributions are highlighted in Section VIII.

II. Literature Review

Over the past 10 years, the trend of situational software engineering has been observed to successfully develop software projects. The situational context in software engineering was explored by numerous researchers over the last decade. Researchers have highlighted the importance of situational factor identification in the context of requirement management. It is conveyed that situational factors such as varying business needs, communication channels, project stakeholders, and technologies influence the requirement management [25].

Moreover, a comprehensive situational framework for the software development process has been introduced by [8]. They followed SLR and data coding procedures of grounded theory to identify the situational factors. According to their investigations team size, culture, task complexity, organization structure, and customer satisfaction are the situational factors that affect the software development and sustainability of engineers. Likewise, situational factors for the requirement engineering process have been explored by [21]. They followed SLR, data coding based on grounded theory to identify situational factors. They argued that management awareness, project characteristics, requirement estimation, and requirement understanding are the situational factors that create an impediment for the requirement engineer to perform requirement engineering activities. It is also argued by [26] that situational factors such as organization size, management, and people impact project success.

TABLE I. Comparison of State-of-the-Art of Situational Software Engineering Research

Reference	Domain of Contribution	Contribution regarding situational Awareness
Ghosh et al., (2011)	Requirement management	Reported situational factors such as varying business needs, communication channels, project stakeholders, and technologies
Clarke and Connor., (2012)	Software development process	Introduced situational framework.
Huma et al., (2014)	Requirement engineering process	Introduced situational framework.
Bakht et al., (2015)	Requirement Engineering	Reported situational factors organization size, management, and people.
Clarke et al., (2016)	Software Development Process	Assessed the impact of the Situational factors such as task, complexity, management commitment, customer satisfaction, and organization structure.
Mark et al., (2017)	Software Development Process	Analyzed the impact of situational factors such as team size, culture, productivity, commitment, skills, turnover, experience, and developer profile
Gulzar et al., (2018)	Requirement Engineering	Reported situational factors such as organization, distance, knowledge management, trust, stakeholder, project, tools and technology, national culture, and physical factors.
Knononenko et al., (2018)	Modern code Review	Reported 'merge type' situational factor
Sadowski et al., (2018)	Modern code Review	Reported situation factors such as source code change size, number of the reviewer, number of comments on the change, distance, social interactions, and review subject
Ebert et al., (2019)	Modern code Review	Reported 'Confusion' situational factor

Additionally, it is reported that situational factors make it difficult to harmonize software development process activities, therefore the researchers should explore the situational factors for software engineering activities to avoid software failures [3]. Similarly, the impact of situational factors such as team size, culture, productivity, commitment, skills, turnover, experience, developer profile on the software development process has been analyzed by [1]. It is contended that software development is highly dependent on situational factors. Likewise, situational context concerning requirement engineering has been explored by [11]. They have identified situational factors such as organization, distance, knowledge management, trust, stakeholder, project, tools and technology, national culture, and physical factors.

Little effort has also been provided concerning situational factors identification in MCR. The recent effort provided by [27], they highlighted the situational factor 'Confusion' in code review. They argued that confusion is an integral part of human problem-solving, which normally arises from the information. They stated that this situational factor causes a decrease in confidence, abandonment of the project, and negative emotions. Likewise, [28] explored situational factors such as source code change size, number of the reviewer, number of comments on the change, distance, social interactions, and review subject can impact the review process outcome. Similarly, [29] argued that the merge type situation has a noteworthy effect on peer review merge time. Table I shows the comparison of the state-of-the-art of situational software engineering research.

From the literature, it is observed that effort has been provided concerning the situational context in software development. It is also analyzed from the literature that three frameworks on situational factors are presented by researchers. One of them focused on the software development process [8], though this framework is the foundation for the research concerning situational factors identification, however it discussed the situational factors' identification for software development generally. Two of them focused on requirement engineering [11], [21].

There is a gap in the literature regarding situational factors' identification for MCR. As a result, there is an unavailability of situational guidelines and frameworks for MCR to aid software engineers to enhance their competency in the recognition of situational factors for MCR activities. To fill this gap this study focused on situational factors' identification in the context of MCR as the MCR team members are facing the individual sustainability challenge of situational factors identification [12], [13], [14], [15]. Besides this, the research aimed to prioritize the most influencing situational factors for MCR activities. The study also aimed to provide a comprehensive Situational MCR framework to support the sustainability of software engineers engaged in MCR activities.

III. RESEARCH METHODOLOGY

This study conducted a Delphi survey to come up with a unique and validated list of situational factors for MCR, as well as the prioritized list of most influencing situational factors for MCR activities. This study is the extension of our previous research work that involves SLR, [23], [24] and expert review [13]. The research methodologies are discussed in the following sub-sections. The research procedure is summarized in Fig. 2.

A. Systematic Litearture Review (SLR)

SLR has been performed to identify the initial list of situational factors from the literature. The guideline given by [30] has been followed to conduct the SLR. The conducted SLR involved 158 studies. As a result of SLR 23 situational factors, 167 sub-factors and 5 categories have been recognized. The primary results of SLR are presented in [23], [24].

B. Expert Review

After conducting the SLR, the obtained list of situational factors was evaluated through expert review. The guideline given by [31] were followed to conduct the expert review. As a result of expert review, 23 situational factors, and 5 categories were obtained, however, the list of sub-factors was reduced to 152 according to the recommendations of the experts. The detailed results of the expert review are presented in [13].

C. Delphi Survey

To evaluate the list of situational factors from industry experts, the Delphi survey was conducted. The Delphi technique is a reliable research method with the potential for use in problem-solving, decision making, and group consensus in a wide variety of areas [32]. In the Delphi method the questionnaires are provided to the experts in such a way that the anonymity of their responses is preserved. Feedback against questionnaires continues till convergence of expert's opinion, or consensus is attained. The work product of the Delphi method is consensus among the experts and their comments, on questionnaire items.

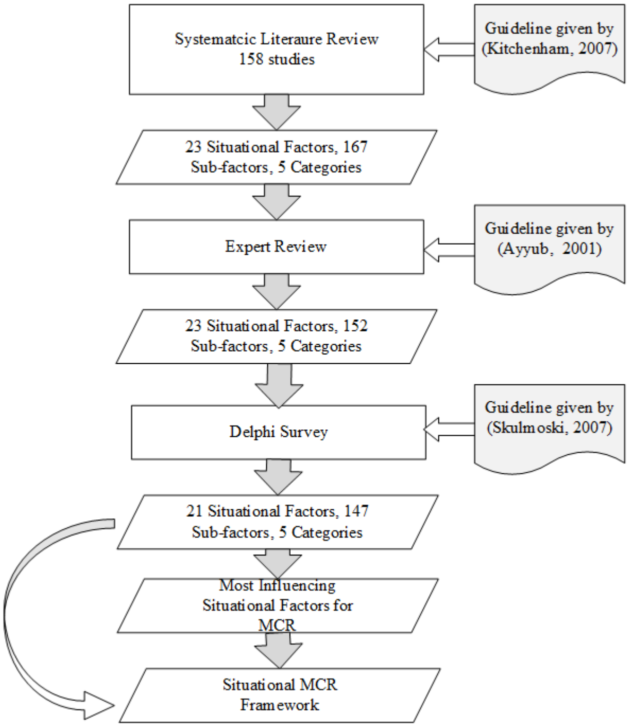


Fig 2. Research Procedure.

D. Delphi Survey Objectives

The Delphi survey has been performed 1) to assess the practicality of the identified situational factors, sub-factors, and categories regarding MCR concerning industry 2) to recognize and prioritized the most influential situational MCR factors for MCR activities 3) to get the recommendation about naming conventions, grouping, and sub-grouping of provided situational MCR factors, sub-factors, and categories 4) to distinguish new industry-based situational MCR factors, with their connected sub-factors, and categories for MCR. The guidelines given by [33] were used to conduct the Delphi Survey.

E. Delphi Experts' Selection and Panel Size

The selection of the experts for the Delphi study is a very important aspect as the Delphi results are intensely based on the views of selected experts [34]. It is claimed that there are four key criteria that must be met while experts' selection: (1) knowledge and experience with the problems under investigation; (2) willingness to participate (3) satisfactory time to give in survey (4) good communication skills [33]. Based on these criteria for this study the experts are selected having experience of 8 or more than 8 years in the industry, must have knowledge of MCR, situational software engineering, and sustainable software engineering particularly individual sustainability. Besides this their capacity and willingness to participate in the study as well as time commitment were also considered for their selection.

Expert panel size deals with the number of experts to participate in the study. It is conveyed that for Delphi study the expert panel size is flexible [33] and a similar group of people, ten to fifteen experts might be enough" [33]. Therefore, we contacted fifteen experts, however, ten experts agreed to participate.

F. No. Rounds in Delphi Study

The performed Delphi survey involved two rounds. The Delphi experts' input was collected through questionnaires. It is claimed that in Delphi study most convergence of panel responses occurs between one to two rounds [35].

G. Delphi Questionnaire Design

The questionnaire is the core aspect of the Delphi study. The Round 1 questionnaire was distributed into four segments. Segment I intended to gather the background of the experts. Segment II was composed of a list of identified situational factors, connected sub-factors, and categories produced as a result of our earlier study grounded on SLR [13], [23], [24] and expert review [13] and expert review. In Segment II the experts were requested to grade the situational factors for their practicality and their level of influence for each MCR activity. Segment III intended to collect new industry based situational factors, connected sub-factors, and categories that must be included in the list. In the same segment the experts were also asked to indicate any recommendation regarding naming conventions, grouping, and sub-grouping of the provided list of factors, sub-factors, and categories from the industry perspective. In segment IV, the experts were requested to provide current real project examples for which they performed MCR activities with the set of situational factors that they experienced. This section was particularly planned for creating the scenario for experimental evaluation of developed situational MCR framework. The Round 2 questionnaire design was similar to Round 1 excluding segment I that aimed to collect the experts' demographic information.

H. Pilot Study

Prior to giving the questionnaire to the experts, it was evaluated in a pilot study by five software engineering researchers for clarity and understanding. As it is conveyed that though a pilot study is an optional aspect, it helps in the identification of obscurities in the questionnaires that might affect the outcome [36]. The feedback received was encouraging. No changes were suggested in the pilot study.

I. Procedure for Data Analysis

This study involved descriptive statistics as it is a basic analytical approach. It gives a basic quantitative approach for examination and produces an overall view of the results [37]. To grade the practicality of situational factors a five-point Likert scale i.e. from 1 to 5 (Very High- 5, High - 4, Moderate - 3, Low- 2, Very Low – 1) was utilized. Likewise, to grade the level of influence of situational MCR factors for each MCR activity, a five-point Likert scale i.e. 1 to 5 (5-Most Influential, 4-Influential, 3-Moderate, 2-Weakly Influential, 1-Not Influential) was utilized. For calculating the practicality of situational MCR factors and to recognize and prioritize the most influential situational MCR factors, the mean values were gathered into the discrete categories as shown in Table II. Table II also shows the grouping of the mean values to measure the practicality of MCR factors and the level of influence of situational MCR factors.

The mean practicality values along with mean influential values of sub-factors were computed primarily and were further converted into a single composite mean value showing composite mean practicality and composite mean influence value for the connected situational MCR factors. To obtain the consensus on the situational MCR factors' practicality and their influence level the standard deviation has been utilized as presented in Table II. Primarily the standard deviation of the sub-factors was computed than was further converted into a single composite standard deviation for the related situational MCR factor. Grounded on the attained composite standard deviation of situational MCR factors the consensus level among the experts was obtained. Equation (1) has been formulated based on the guidelines given by [38] to compute the composite standard deviation of situational factors.

$$SD(SitF)=\sqrt{\left((SD(SitSbF_1)) +\ldots+(SD(SitSbF_k)))\right)/k} \quad (1)$$

TABLE II. GROUPING OF MEAN VALUES TO MEASURE PRACTICALITY AND INFLUENCE LEVEL OF SITUATIONAL MCR FACTORS

Mean Score =X	Level of Practicality	Level of Influence
4.0≤X≤ 5.0	Very High	Most Influential
3.0≤X< 4.0	High	Influential
2.0≤X< 3.0	Moderate	Moderate
1.0≤X< 2.0	Low	Weakly Influential
0≤X< 1.0	Very Low	Not Influential

TABLE III. DECISION PRINCIPLES FOR THE LEVEL OF CONSENSUS

Standard Deviation (SD=X)	Level of Consensus
0 ≤ X <1	High
1 ≤ X <1.5	Fair Level
1.5 ≤ X <2	Low Level
2 < X	No Consensus

Where '*SD*' refers to the standard deviation, '*SitF*' denotes to situational factor. '*SitSbF*' denotes the sub-factor of the associated situational factor and its value ranges from 1 to *k*, and '*k*' denotes to the total number of sub-factors for associated situational factors.

To achieve the consensus among the experts the standard deviation was measured based on [39]. Table III denotes the consensus levels used in this study. A standard deviation near to '0' indicated that the experts' gradings tended to be very close to each other, standard deviation far from '0' indicated that the gradings were spread out over a large range.

J. Data Collection and Analysis

This section detailed the data collected from the Delphi experts along with the analysis of collected data relying on the procedure of data analysis mentioned in sub-section 'I'. The conducted study involved two rounds. The details about data collection in Delphi rounds are discussed in sub-sections.

K. Delphi Survey Round 1

In the Delphi Round 1 the questionnaire was provided to the experts to collect their input. They were given one week to provide the feedback. The experts were contacted on phone calls to make sure about their mindfulness concerning the feedback submission date for Round 1. It takes two weeks to complete the Delphi survey Round 1. The Round I intended to collect the background information from the experts. It also intended to assess the list of situational MCR factors, connected sub-factors, and categories for their naming, grouping, and sub-grouping, which was produced as a result of our previous study based on SLR [23], [24] and expert review [2]. Round 1 includes the assessment of the situational MCR factors for their practicality and their level of influence for each MCR activity. In Round 1 the experts were also asked to state new industry-oriented situational MCR factors, connected sub-factors, and categories that need to be present in the provided list. The scale used to grade the practicality and level of influence is presented in sub-section 'I'. The particulars regarding the Round 1 questionnaire are given in sub-section 'G'. In Delphi Round 1 the expert provided some recommendations, therefore Delphi Round 2 has been conducted to get the consensus on the recommended changes among the experts.

L. Delphi Round 2

In Round 2, the experts were provided with the summary of Round 1 results along with the updated list of situational MCR factors, sub-factors, and categories based on the suggestions of the experts given in Round 1. The details about the Round 2 questionnaire are given in sub-section 'G'. Round 2 was completed in 2 weeks. In Round 2 the consensus among the experts on the provided list of all situational MCR factors was achieved, and no changes were suggested therefore we decided to stop at Delphi Round 2.

IV. RESULTS

This section highlights the Delphi survey results. The practicality level of situational MCR factors with the standard deviation for Delphi Round 1 and 2 are presented in Fig. 3 to Fig. 6. Fig. 3 and Fig. 4 shows the comparison of the Delphi survey results of mean perceived practicality values of the situational MCR factors. It also shows the changes performed based on the suggestions of experts. For instance, the factors name 'Team' was changed as 'Team Dynamics'. This comparison also indicated that the mean perceived practicality value for the situational factors have been increased in Delphi Round 2. The comparison of Fig 5 and Fig. 6 shows that the consensus level for most of the situational factors was increased in Round 2 for their practicality among the Delphi experts as standard deviation moves near to '0' in Round 2. For instance, for the factor 'Tool,' the standard deviation was '0.3855011' in Round 1 whereas in Round 2 the standard deviation was '0.3312434'.

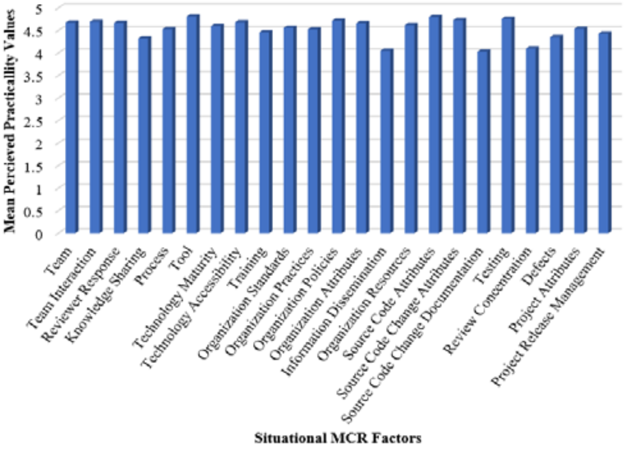


Fig 3. Composite Mean Perceived Practicality Value of Situational MCR Factors -Round 1.

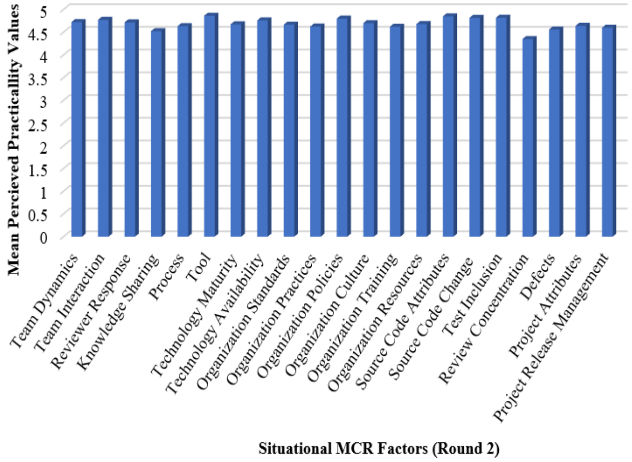


Fig 4. Composite Mean Perceived Practicality Value of Situational MCR Factors -Round 2.

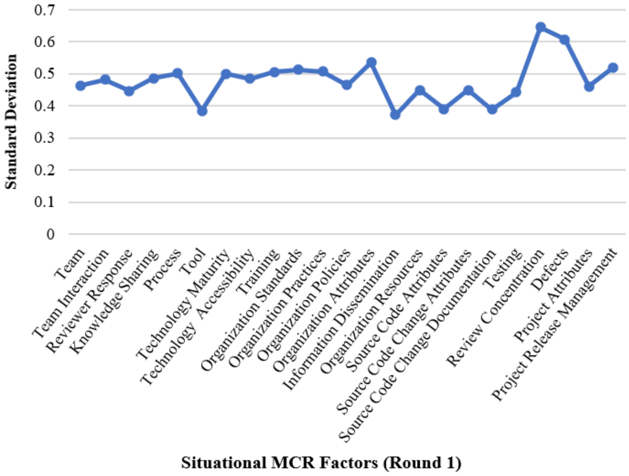


Fig 5. Consensus Level among the Delphi experts for Mean Perceived Practicality of situational MCR factors - Round 1.

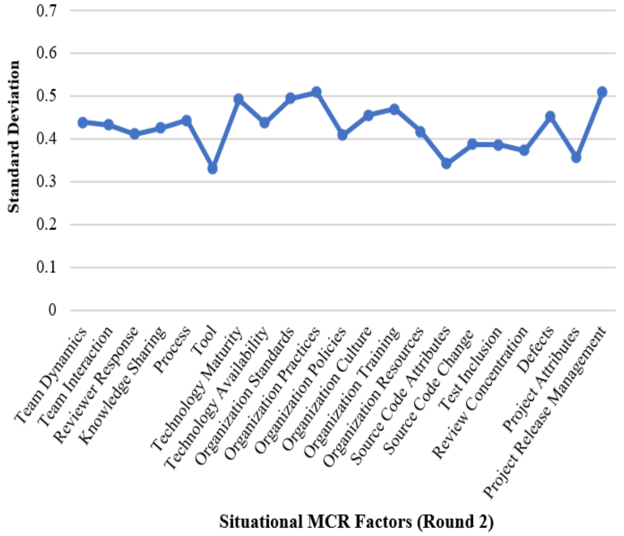


Fig 6. Consensus Level among the Delphi experts for Mean Perceived Practicality of situational MCR factors - Round 2.

Table IV displays the ranking of situational MCR factors for their level of practicality.

Concerning the most influential situational MCR factors, the mean influential values of sub-factors of each situational MCR factor in final Round for; Source Code Preparation ranges from 2.5 to 5.0, Source Code Submission ranged from 1.2 to 5.0, Reviewer Selection and Notification ranges from 1.8 to 5.0, Source Code Review ranges from 3.0 to 5.0, Source Code Approval ranges from 2.0 to 5.0. To find the most influential factors, the composite mean influential values of their related sub-factors were computed. The factors having composite mean values equal to or above 4.00 were considered as the most influential factors for that specific MCR activity. For each MCR activity, the most influential factors were identified based on their composite mean values after the final Delphi Round and are presented in Tables V to IX with the standard deviation.

TABLE IV. RANKING OF SITUATIONAL FACTORS FOR PERCEIVED LEVEL OF PRACTICALITY

Situational MCR Factors	Composite Mean Practicality Values	Standard Deviation	Rank
Tool	4.8625	0.331243449	1
Source Code Attributes	4.85	0.341565026	2
Test Inclusion	4.82	0.38586123	3
Source Code Change	4.816	0.387298335	4
Organization Policies	4.8	0.40824829	5
Team Interaction	4.772	0.43228311	6
Technology Availability	4.76	0.437162568	7
Team Dynamics	4.724	0.438684903	8
Reviewer Response	4.71667	0.410510071	9
Organization Culture	4.7	0.455420034	10
Organization Resources	4.68	0.4163332	11
Technology Maturity	4.675	0.49159604	12
Organization Standards	4.667	0.494413232	13
Project Attributes	4.641	0.356162676	14
Process	4.633	0.443053379	15
Organization Practices	4.625	0.508265023	16
Organization Training	4.62	0.469041576	17
Project Release Management	4.6	0.509175077	18
Defects	4.56	0.451335467	19
Knowledge Sharing	4.525	0.424918293	20
Review Concentration	4.35	0.372677996	21

TABLE V. RANKING AND INFLUENTIAL LEVEL OF SITUATIONAL MCR FACTORS FOR SOURCE CODE PREPARATION

Most influential Situational Factors	Composite Mean Influential Value	Standard Deviation	Rank
Source Code Attributes	4.8625	0.343592135	1
Source Code Change	4.725	0.361324723	2
Tool	4.7125	0.445658065	3
Team Dynamics	4.704	0.436144726	4
Test Inclusion	4.6	0.503322296	5
Organization Culture	4.45	0.45338235	6
Organization Policies	4.43333	0.389681731	7
Project Attributes	4.4166	0.449279258	8
Process	4.11666	0.307318149	9
Technology Availability	4.08	0.53748385	10
Project Release Management	4.066666667	0.384900179	11
Defects	4.0333	0.36004115	12
Organization Practices	4	0.307318149	13

TABLE VI. RANKING AND INFLUENTIAL LEVEL OF SITUATIONAL MCR FACTORS FOR SOURCE CODE SUBMISSION

Most influential Situational Factors	Composite Mean Influential Value	Standard Deviation	Rank
Tool	4.8375	0.170782513	1
Technology Availability	4.8	0.405517502	2
Process	4.75	0.396746024	3
Team Dynamics	4.676	0.450678501	4
Organization Practices	4.6	0.48876261	5
Project Attributes	4.3	0.344265186	6
Organization Culture	4.2	0.298142397	7
Project Release Management	4.166	0.349602949	8
Source Code Attributes	4.05	0.333333333	9
Source Code Change	4.025	0.358752984	10
Organization Policies	4.016	0.129099445	11
Test Inclusion	4	0.529150262	12

TABLE VII. RANKING AND INFLUENTIAL LEVEL OF SITUATIONAL MCR FACTORS FOR REVIEWER SELECTION AND NOTIFICATION

Most influential Situational Factors	Composite Mean Influential Value	Standard Deviation	Rank
Team Dynamics	4.888	0.27325202	1
Team Interaction	4.82727	0.325669474	2
Reviewer Response	4.71666	0.440958552	3
Tool	4.625	0.329140294	4
Source Code Attributes	4.5375	0.385501117	5
Project Attributes	4.533	0.394405319	6
Source Code Change	4.5083	0.311804782	7
Process	4.45	0.387298335	8
Technology Availability	4.32	0.333333333	9
Organization Practices	4.25	0.414996653	10
Organization Culture	4.2333	0.344265186	11
Organization Resources	4.18	0.294392029	12
Organization Policies	4.05	0.396746024	13
Project Release Management	4.03	0.182574186	14

Based upon the practicality and influence level of situational MCR factors, the situational MCR framework has been developed to support individual sustainability of software engineers engaged in MCR activities. The developed framework guides the MCR team members about the situational factors as well as most influential situational MCR factors for each MCR activity. The developed framework might also help the software engineers to be aware of upcoming situations and to improve their competence and productivity. The situational MCR framework is attached in the Appendix I (Fig. 7a and Fig. 7b).

TABLE VIII. RANKING AND INFLUENTIAL LEVEL OF SITUATIONAL MCR FACTORS FOR SOURCE CODE REVIEW

Most influential Situational Factors	Composite Mean Influential Value	Standard Deviation	Rank
Source Code Attributes	4.925	0.252762515	1
Source Code Change	4.866	0.288675135	2
Test Inclusion	4.84	0.359010987	3
Team Dynamics	4.792	0.354024481	4
Team Interaction	4.781818	0.386645767	5
Reviewer Response	4.683	0.341565026	6
Tool	4.5625	0.305050087	7
Organization Culture	4.45	0.36767538	8
Process	4.4	0.272165527	9
Organization Policies	4.383333	0.36767538	10
Technology Availability	4.38	0.380058475	11
Project Attributes	4.25	0.423827358	12
Defects	4.15	0.319142369	13
Organization Practices	4.1	0.25819889	14
Organization Training	4.02	0.391578004	15

TABLE IX. RANKING AND INFLUENTIAL LEVEL OF SITUATIONAL MCR FACTORS FOR SOURCE CODE APPROVAL

Most influential Situational Factors	Composite Mean Influential Value	Standard Deviation	Rank
Source Code Attributes	4.8125	0.35551215	1
Source Code Change	4.6583	0.267013663	2
Defects	4.6	0.370185139	3
Project Attributes	4.5	0.430331483	4
Team Dynamics	4.464	0.388730126	5
Tool	4.35	0.263523138	6
Organization Culture	4.23	0.316227766	7
Process	4.15	0.324893145	8
Project Release Management	4.1	0.278886676	9
Technology Availability	4.06	0.374165739	10
Organization Policies	4.016	0.36767538	11

V. DISCUSSION

Software engineers engaged in MCR activities belongs to various cultures and backgrounds, as a result, various unknown and varying situational factors arise that impact their sustainability and productivity due to mental distress. Therefore, this study has presented a situational MCR framework based on Delphi survey. The developed framework guides the software engineers about the situational factors as well as the most influential situational MCR factors for each MCR activity. The developed framework might also support the sustainability of software engineers, help them to be aware of upcoming situations, and to improve their productivity. The situational MCR framework is attached in Appendix I (Fig. 7a and Fig. 7b).

VI. CONCLUSION

This study presented a situational MCR framework to support the sustainability of software engineers engaged in MCR activities. The study findings reported 21 situational factors, 147 sub-factors, and 5 Categories. In this paper, the Delphi survey results along with the development of the situational MCR framework are presented. The developed framework might guide the software engineers engaged in MCR activities to consider the reported situations, identify situations according to their context, and improve their productivity while ensuring their individual sustainability.

VII. FUTURE WORK

This study can be further elaborated for other software engineering contexts and activities for instance software design, software testing, etc. The ongoing research plans are to validate the developed situational MCR framework through experiment and to develop a web-oriented situational MCR tool to have an electronic situational guideline for software engineers engaged in MCR activities.

VIII. CONTRIBUTION

The investigation contributed to the software engineering body of knowledge (SWEBOK) specifically situational software engineering and sustainable software engineering, particularly the individual sustainability of software engineers. The study reported the situational factors for MCR from literature, academic, and industry. The study also recognizes and prioritizes the most influential situational MCR factors and present the situational MCR framework to support the individual sustainability of software engineers engaged in MCR activities.

REFERENCES

[1] G. Marks, R. V. O’Connor, and P. M. Clarke, “The impact of situational context on the software development process – A case study of a highly innovative start-up organization,” Commun. Comput. Inf. Sci., vol. 770, pp. 455–466, 2017.

[2] C. K. Chang, “Situation Analytics: A Foundation for a New Software Engineering Paradigm,” Computer (Long. Beach. Calif)., vol. 49, no. 1, pp. 24–33, 2016.

[3] R. V. O. Connor, P. Elger, and P. M. Clarke, “Exploring the impact of situational context – A case study of a software development process for a microservices architecture,” pp. 6–10, 2016.

[4] B. Penzenstadler et al., “Software Engineering for Sustainability: Find the Leverage Points!,” IEEE Softw., vol. 35, no. 4, pp. 22–33, 2018.

[5] S. Nazir, N. Fatima, S. Chuprat, H. Sarkan, N.F. Nilam, N.A. Sjarif “Sustainable Software Engineering: A Perspective of Individual Sustainability Challenges,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 10, no. 2, pp. 676–683, 2020.

[6] S. Nazir, N. Fatima, and S. Chuprat, “Individual Sustainability Barriers and Mitigation Strategies: Systematic Literature Review Protocol,” in 2019 IEEE Conference on Open System, ICOS 2019, 2019, pp. 1–5.

[7] B. Penzenstadler, A. Raturi, D. Richardson, C. Calero, H. Femmer, and X. Franch, “Systematic Mapping Study on Software Engineering for Sustainability (SE4S),” in Proc. 18th International Conference on Evaluation and Assessment in Software Engineering, 2014, pp. 1–14.

[8] P. Clarke and R. V. O’Connor, “The situational factors that affect the software development process: Towards a comprehensive reference framework,” Inf. Softw. Technol., vol. 54, no. 5, pp. 433–447, 2012.

[9] M. Salam and S. U. Khan, “Challenges in the development of green and sustainable software for software multisourcing vendors: Findings from a systematic literature review and industrial survey,” J. Softw. Evol. Process, vol. 30, no. 8, pp. 1–21, 2018.

[10] D. Mishra, S. Aydin, A. Mishra, and S. Ostrovska, “Knowledge management in requirement elicitation: Situational methods view,” Comput. Stand. Interfaces, vol. 56, no. September, pp. 49–61, 2018.

[11] K. Gulzar, J. Sang, A. A. Memon, and M. Ramzan, “A Practical Approach for Evaluating and Prioritizing Situational Factors in Global Software Project Development,” vol. 9, no. 7, 2018.

[12] R. Chitchyan, I. Groher, and J. Noppen, “Uncovering sustainability concerns in software product lines,” J. Softw. Evol. Process, vol. 29, no. 2, pp. 1–20, 2017.

[13] S. Nazir, N. Fatima, and S. Chuprat, “Situational factors for modern code review to support software engineers’ sustainability,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 1, pp. 498–504, 2020.

[14] R. Chitchyan, L. Duboc, C. Becker, S. Betz, B. Penzenstadler, and C. C. Venters, “Sustainability Design in Requirements Engineering : State of Practice,” in IEEE/ACM 38th IEEE International Conference on Software Engineering, 2016, pp. 533–542.

[15] N. Fatima, S. Chuprat, and S. Nazir, “Challenges and Benefits of Modern Code Review-Systematic Literature Review Protocol,” in Proc. International Conference on Smart Computing and Electronic Enterprise, 2018, pp. 1–5.

[16] A. Bacchelli and C. Bird, “Expectations, outcomes, and challenges of modern code review,” in Proc. International Conference on Software Engineering, 2013, pp. 712–721.

[17] N. Fatima, S. Nazir, and S. Chuprat, “Understanding the Impact of Feedback on Knowledge Sharing in Modern Code Review,” in 6th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS) , 2019.

[18] L. MacLeod, M. Greiler, M. A. Storey, C. Bird, and J. Czerwonka, “Code Reviewing in the Trenches: Challenges and Best Practices,” IEEE Softw., vol. 35, no. 4, pp. 34–42, 2018.

[19] A. Bosu, J. C. Carver, C. Bird, J. Orbeck, and C. Chockley, “Process Aspects and Social Dynamics of Contemporary Code Review: Insights from Open Source Development and Industrial Practice at Microsoft,” IEEE Trans. Softw. Eng., vol. 43, no. 1, pp. 56–75, 2017.

[20] N. Fatima, S. Nazir, and S. Chuprat, “Knowledge sharing, a key sustainable practice is on risk: An insight from Modern Code Review,” in 2019 6th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS), 2019.

[21] H. H. Khan, M. N. Mahrin, and S. Chuprat, “Situational requirement engineering framework for Global Software Development,” 2014 Int. Conf. Comput. Commun. Control Technol., no. I4ct, pp. 224–229, 2014.

[22] P. Clarke, R. V. O. Connor, R. V. O. Connor, and B. Leavy, “A complexity theory viewpoint on the software development process and situational context,” no. May, 2016.

[23] S. Nazir, N. Fatima, and S. Chuprat, “Does Project Associated Situational Factors have Impact on Sustainability of Modern Code Review Workforce?,” in 6th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS) , 2019

[24] S. Nazir, N. Fatima, and S. Chuprat, “Situational factors affecting Software Engineers Sustainability: A Vision of Modern Code Review,” in 6th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS), 2019.

[25] S. Ghosh, A. Dubey, and S. Ramaswamy, “C-FaRM: A collaborative and context aware framework for requirements management,” 2011 4th Int. Work. Manag. Requir. Knowledge, MaRK’11 - Part 19th IEEE Int. Requir. Eng. Conf. RE’11, pp. 29–30, 2011.

[26] K. Bakhat, A. Sarwar, and Y. H. Motla, “A Situational Requirement Engineering Model for an Agile Process,” Bahria Univ. J. Inf. Commun. Technol. ISSN – 1999-4974, vol. 8, no. 1, pp. 21–26, 2015.

[27] F. Ebert, F. Castor, N. Novielli, and A. Serebrenik, “Confusion in Code Reviews: Reasons, Impacts, and Coping Strategies,” SANER 2019 - Proc. 2019 IEEE 26th Int. Conf. Softw. Anal. Evol. Reengineering, pp. 49–60, 2019.

[28] C. Sadowski, E. Söderberg, L. Church, M. Sipko, and A. Bacchelli, “Modern code review: : A Case Study at Google,” in Proc. ACM/IEEE 40th International Conference on Software Engineering: Software Engineering in Practice, 2018, pp. 181–190.

[29] O. Kononenko, T. Rose, O. Baysal, M. Godfrey, D. Theisen, and B. De Water, “Studying Pull Request Merges : A Case Study of Shopify ’ s Active Merchant,” in Proc. 40th International Conference on Software Engineering: Software Engineering in Practice, 2018, pp. 124–133.

[30] B. Kitchenham and S. Charters, “Source: " Guidelines for performing Systematic Literature Reviews in SE " , Kitchenham et al Guidelines for performing Systematic Literature Reviews in Software Engineering Source: " Guidelines for performing Systematic Literature Reviews i,” pp. 1–44, 2007.

[31] B. Ayyub, “A practical guide on conducting expert-opinion elicitation of probabilities and consequences for corps facilities,” Inst. Water Resour. Alexandria, VA, USA, no. January, 2001.

[32] H. A. von der Gracht, “Consensus measurement in Delphi studies. Review and implications for future quality assurance,” Technol. Forecast. Soc. Change, vol. 79, no. 8, pp. 1525–1536, 2012.

[33] G. J. Skulmoski, F. T. Hartman, and Jennifer Krahn, “The Delphi Method for Graduate Research,” J. Inf. Technol. Educ., vol. 6, 2007.

[34] M. Adler and Z. Erio, Gazing Into the Oracle: The Delphi Method and Its Application to Social Policy and Public Health. Jessica Kingsley, 1996.

[35] H. W. Lanford, Technological forecasting methodologies; a synthesis. New York: American Management Association, 1972.

[36] N. J. Shariff, “Utilizing the Delphi Survey Approach: A Review,” J. Nurs. Care, vol. 04, no. 03, 2015.

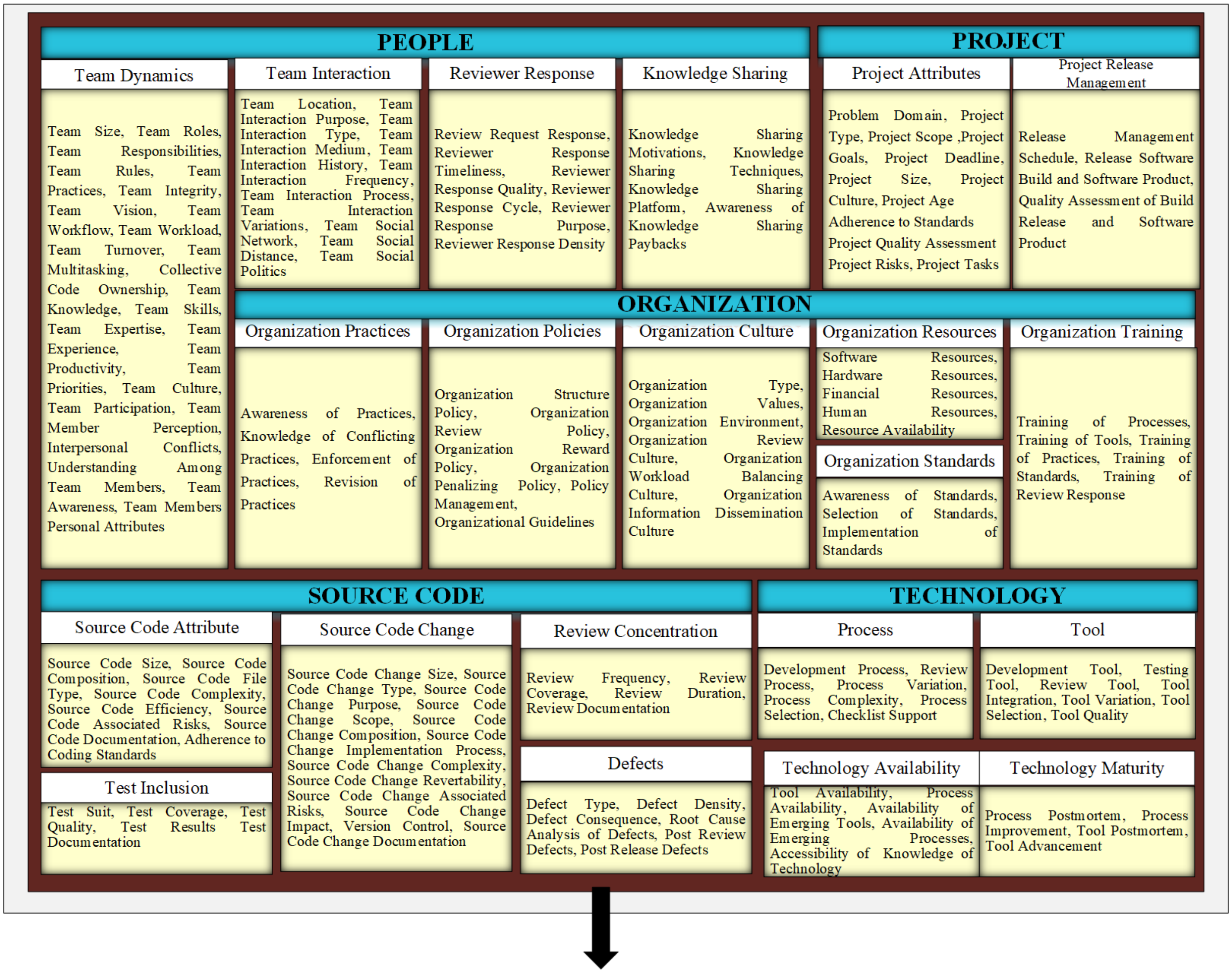
[37] S. G. Naoum, Dissertation research and writing for construction students, Second Edition, 2nd ed. Oxford : Butterworth-Heinemann, 2012.

[38] J. Cohen, Statistical power analysis for the behavioral science, 2nd ed. New: Lawrence Erlbaum Associates, 1988.

[39] M. R. Rogers and E. C. Lopez, “Identifying critical cross-cultural school psychology competencies,” J. Sch. Psychol., vol. 40, no. 2, pp. 115–141, 2002.

APPENDIX I

Situational MCR Framework



PEOPLE

Team Dynamics: Team Size, Team Roles, Team Responsibilities, Team Rules, Team Practices, Team Integrity, Team Vision, Team Workflow, Team Workload, Team Turnover, Team Multitasking, Collective Code Ownership, Team Knowledge, Team Skills, Team Expertise, Team Experience, Team Productivity, Team Priorities, Team Culture, Team Participation, Team Member Perception, Interpersonal Conflicts, Understanding Among Team Members, Team Awareness, Team Members Personal Attributes

Team Interaction: Team Location, Team Interaction Purpose, Team Interaction Type, Team Interaction Medium, Team Interaction History, Team Interaction Frequency, Team Interaction Process, Team Interaction Variations, Team Social Network, Team Social Distance, Team Social Politics

Reviewer Response: Review Request Response, Reviewer Response Timeliness, Reviewer Response Quality, Reviewer Response Cycle, Reviewer Response Purpose, Reviewer Response Density

Knowledge Sharing: Knowledge Sharing Motivations, Knowledge Sharing Techniques, Knowledge Sharing Platform, Awareness of Knowledge Sharing Paybacks

PROJECT

Project Attributes: Problem Domain, Project Type, Project Scope ,Project Goals, Project Deadline, Project Size, Project Culture, Project Age Adherence to Standards Project Quality Assessment Project Risks, Project Tasks

Project Release Management: Release Management Schedule, Release Software Build and Software Product, Quality Assessment of Build Release and Software Product

ORGANIZATION

Organization Practices: Awareness of Practices, Knowledge of Conflicting Practices, Enforcement of Practices, Revision of Practices

Organization Policies: Organization Structure Policy, Organization Review Policy, Organization Reward Policy, Organization Penalizing Policy, Policy Management, Organizational Guidelines

Organization Culture: Organization Type, Organization Values, Organization Environment, Organization Review Culture, Organization Workload Balancing Culture, Organization Information Dissemination Culture

Organization Resources: Software Resources, Hardware Resources, Financial Resources, Human Resources, Resource Availability

Organization Standards: Awareness of Standards, Selection of Standards, Implementation of Standards

Organization Training: Training of Processes, Training of Tools, Training of Practices, Training of Standards, Training of Review Response

SOURCE CODE

Source Code Attribute: Source Code Size, Source Code Composition, Source Code File Type, Source Code Complexity, Source Code Efficiency, Source Code Associated Risks, Source Code Documentation, Adherence to Coding Standards

Test Inclusion: Test Suit, Test Coverage, Test Quality, Test Results Test Documentation

Source Code Change: Source Code Change Size, Source Code Change Type, Source Code Change Purpose, Source Code Change Scope, Source Code Change Composition, Source Code Change Implementation Process, Source Code Change Complexity, Source Code Change Revertability, Source Code Change Associated Risks, Source Code Change Impact, Version Control, Source Code Change Documentation

Review Concentration: Review Frequency, Review Coverage, Review Duration, Review Documentation

Defects: Defect Type, Defect Density, Defect Consequence, Root Cause Analysis of Defects, Post Review Defects, Post Release Defects

TECHNOLOGY

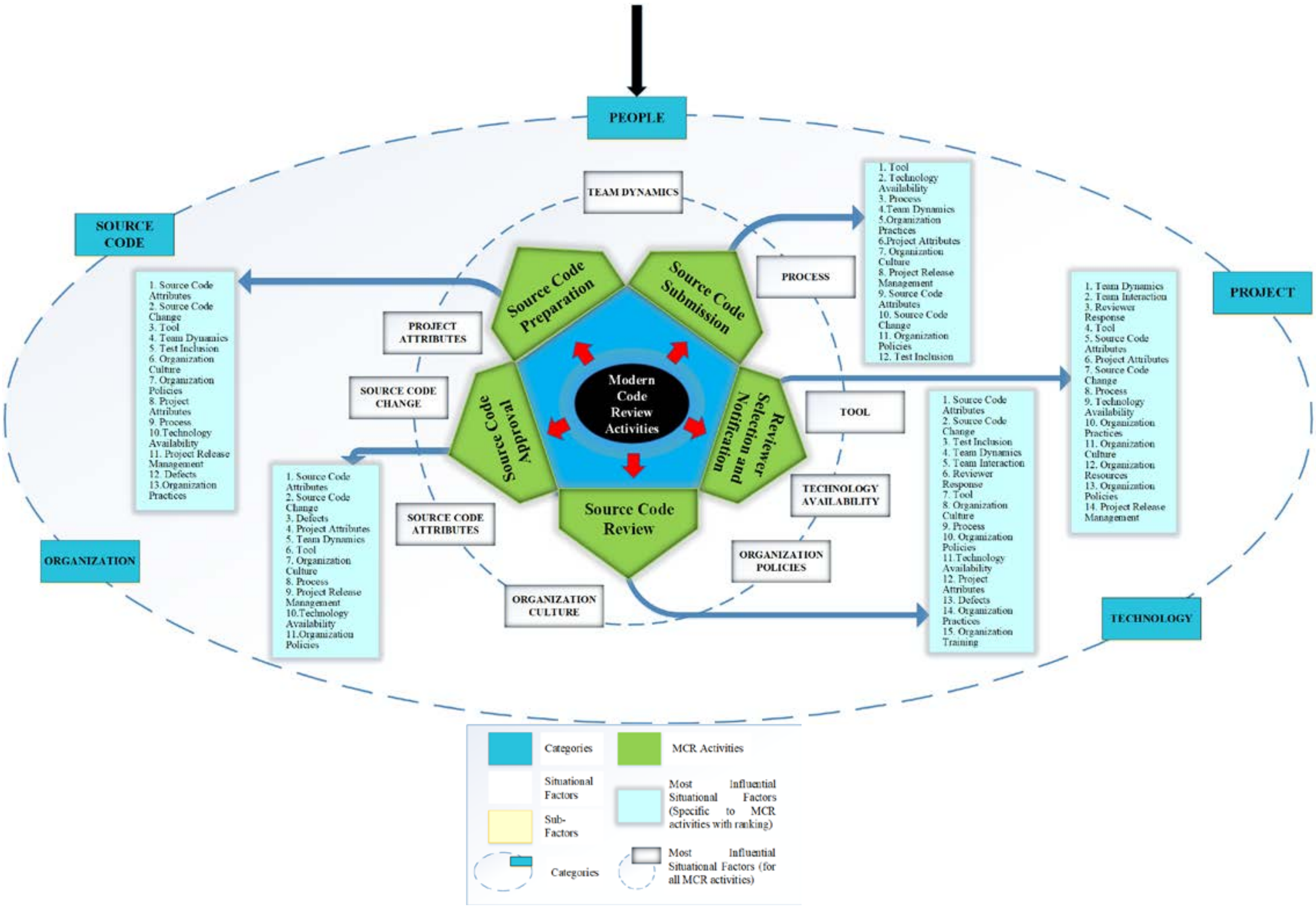
Process: Development Process, Review Process, Process Variation, Process Complexity, Process Selection, Checklist Support

Tool: Development Tool, Testing Tool, Review Tool, Tool Integration, Tool Variation, Tool Selection, Tool Quality

Technology Availability: Tool Availability, Process Availability, Availability of Emerging Tools, Availability of Emerging Processes, Accessibility of Knowledge of Technology

Technology Maturity: Process Postmortem, Process Improvement, Tool Postmortem, Tool Advancement

(a)



(b)

Fig 7. (a) Situational MCR Framewrok for Sustainable Software Engineers- Part I, (b) Situational MCR Framewrok for Sustainable Software Engineers- Part II.

Solving Travelling Salesman Problem (TSP) by Hybrid Genetic Algorithm (HGA)

Ali Mohammad Hussein Al-Ibrahim
Faculty of Information Technology
Department of Computer Science
The World Islamic Sciences and Education University (WISE)
P.O. Box 1101, Amman 11947, Jordan

***Abstract*—The Traveling Salesman Problem (TSP) is easy to qualify and describe but difficult and very hard to be solved. There is known algorithm that can solve it and find the ideal outcome in polynomial time, so it is NP-Complete problem. The Traveling Salesman Problem (TSP) is related to many others problems because the techniques used to solve it can be easily used to solve other hard Optimization problems, which allows of circulating it results on many other optimization problems. Many techniques were proposed and developed to solve such problems, including Genetic Algorithms. The aim of the paper is to improve and enhance the performance of genetic algorithms to solve the Traveling Salesman Problem (TSP) by proposing and developing a new Crossover mechanism and a local search algorithm called the Search for Neighboring Solution Algorithm, with the goal of producing a better solution in a shorter period of time and fewer generations. The results of this study for a number of different size standard benchmarks of TSP show that the proposed algorithms that use Crossover proposed mechanism can find the optimum solution for many of these TSP benchmarks by (100%) ,and within the rate (96%-99%) of the optimal solution to some for others. The comparison between the proposed Crossover mechanism and other known Crossover mechanisms show that it improves the quality of the solutions. The proposed Local Search algorithm and Crossover mechanism produce superior results compared to previously propose local search algorithms and Crossover mechanisms. They produce near optimum solutions in less time and fewer generations.**

***Keywords*—Traveling Salesman Problem (TSP); NP-Complete problem; genetic algorithms; local Search algorithm; crossover mechanism; neighboring solution algorithm**

I. The Problem of Travelling Salesman

A. Introduction

The issues of improvement to find the best solution optimization problems of the most important problems worthy of study, and characterized these issues with difficulty to reach solutions that meet all the goals, and examples of the problem of scheduling (timetabling problem) and the problem of finding the shortest route for a (travelling salesman problem) and the problem of coloring schemes (graph coloring problem) and the problem of transport (transportation problem), given the differences in the nature of each problem and the degree of difficulty cannot be to work with each algorithm with the same efficiency with all kinds of problems of finding the best solution (Newall, 1998) [18] and it has been many studies to compare the mechanism of the different algorithms and see which ones more suitable to solve the problem (Newall,1998) [18], (Burke,1995) [2].

This study is concerned the problem of travelling salesman, find the shortest route which passes a range of cities without repetition.

B. Reasons for Choosing a Sales Trip Problem

The problem of travelling salesman old, it was the first to predicate the same problem at 1759 by Lawler, who was his attention focused on solving the problem of round knight on the chessboard so that visit all the boxes, patch once without any recurrence (Michalewicz, 1999) [14] and was the first use of the term travelling salesman in 1932 by a travelling salesman call Veteran has focused on a Rand Corporation in 1948 to resolve the problem of travelling salesman and helped the company's reputation by making this problem of problems known (Michalewicz, 1999) [14]. In addition, it has helped the emergence of liner programming in that time, solving the problem of travelling salesman. In spite of oldest of this problem its simplicity, but that did not make it easy to resolve, which, the authors should be widely available to many areas of research in mathematics and linear algebra to can understand the theory behind the solution to this problem. And the researches increased in this problem by the increased development in computers in terms of processing speed and storage capacity and memory, And has this problem drew the attention of many mathematicians and computer experts (Michalewicz, 1995) [15] for several reasons including:

1) Easy description, but its solution is very difficult because it belongs to a class of optimization problems called NP-complete which mean that there are no known algorithm that can find the ideal time of the result can be represented by multiple dimensions equation (Johnson polynomial).

2) Problem of travelling salesman can be generalized to many of the problems, such as:

- The problem of scheduling the tasks set to the CPU machine.
- Network routing problem.
- Path planning in mobile robot problem.
- Airline schedule problem
- As the solution to many of the problems of artificial intelligence will be finding the order of the number of

data in certain ways as a problem of circuit boards, electrical circuit board drilling, a method of solving the problem of a sales trip

3) We have a lot of information available and the standard questions about this problem, which it considered the mother problem (Michalewicz, 1995) [15] for many of the problems, so you can make many experiments on this problem and then present the solution methods to other problems.

C. Description of the Travelling Salesman Problem

On the assumption that there are a number of cities, (c1, c2, ..., cn) represented as points scheme, and there are lines between points is called edges and each edge length is the distance between cities, d(cj, ci) and the problem of travelling salesman is to find a path going through all the points so that the edges along the shortest possible (Johnson and Mcgeoch 1998) [8]. In other words, the salesman must visit all these cities and return to the city of departure under the following conditions:

- Salesman can visit only one city at any time.
- The salesman must visit all the cities and then return to its point of departure.
- Should not visit any city more than once in one round.
- The total distance should be the least that can be.

N-1

d(Cπ(i), Cπ(j))+ d(Cπ(N), Cπ(I))] ∑ Minimize[
i=1

The equation above represents the length of the tour, which the salesman went through during a trip to visit cities in the tour route in the order and then return to starting point. The travelling salesman problem has two main types:

- Asymmetric travelling salesman problem TSP, it mean that the distance between the city(i) and the city(j), is not equal to the distance between the city(j) and the city(i)

d (Ci, Cj) ≠ d (Cj, Ci)

1 <= i,j <= N, where (N is the number round cities)

- Symmetric TSP , it mean that the distance between the city(i) and the city(j), are the same as the distance between the city(j) and the city(i).

d (Ci, Cj) = d (Cj, Ci)   For each 1 <= i,j <= N

It is clear that the symmetric travelling salesman problem is a special case of Asymmetric travelling salesman problem, but the practical experiences declare that it cannot disseminate the results Asymmetric travelling salesman problem to symmetric travelling salesman problem (Hoffman, 1985, 1991) [7], therefore this type need special research. The absence of any known algorithm to solve this problem in a time can be represented in terms of the number of NP-complete, make sure and only way to find the optimal solution is to calculate the lengths of all possible ways, and then choose the shortest route of them to be the best. But if we suppose it is possible to hold one billion a calculation every second, we will need several years to calculate the lengths of possible routes for 25 cities and numbering (2,65 × 1032) route by (Lalena, 1998-2003)[10]

So must look for alternative routes and approximate algorithms to find a solution in a reasonable time relatively (katayama and Narihisa, 1999) [21], to be any sacrifice to reach the ideal solution to dissolve and sufficiency close to which is obtained in less time.

The most important approximate algorithms is genetic algorithms, including those that use local search, genetic algorithms, and the genetic algorithms have formed the subject of numerous studies (Michalewicz, 1996) [39.40] (Mitchell, 1996) [17] (Obitko, 1998) [19], but suffer from high time for implementation to get a good result, this study aims to improve the performance of genetic algorithms to reduce the number of generations and reduce the time needed to implement them.

II. Genetic Algorithms

A. Introduction

Genetic algorithms were inspired by Darwin's theory of evolution. The first began in the description of the idea, (John Holland's, 1992) [9]owner of theory of adaptive systems in the sixties (Man, 1999) [13] at the University of Michigan. The program is adaptive to study how the program can establish that the procedures that will allow itself to be altered depending on the efficiency of the existing environment (Goldberg, 1989) [5].

The main idea of the Genetic Algorithms is to do what nature does, it simulates what happens in living nature of the selection processes and the evolution and mutations in order to produce improved generations of parents specification involved in the process of mating, the more selected good qualities from the parents led to the emergence in children (David, 1989)[3] .And genetic algorithms is search and optimization based on the mechanics and the foundations of the survival of the best as it is in the processes of natural selection, chromosomes compete with each other and the environment, in which only the best remains to be in the new gene

New algorithms start encrypts set of solutions to a problem, and the way in which the chromosome encoding or representation that would reduce or increase the efficiency of genetic algorithms, which is the key to the solution, which expands approved work (Man, 1999) [13], these encrypts called Genes, that yields chromosome which is form a solution for a problem.

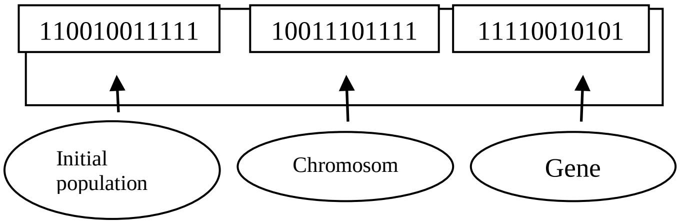


Fig 1. Genetic Algorithms Content.

Set of solutions that begins with it is called initial population, which will be subject to genetic algorithms for the formation of new solutions and to obtain or approach the final solution of the problem, as shown in Fig. 1.

One Chromosome Genes subject to simple changes in the site or the situation which leading to the emergence of recipes and new genetic. This is called mutation Change , either inherited qualities of the chromosomes of parental to sons using the mechanism of the Crossover, through the merger of parts of chromosomes have been selected a set of solutions called parents, to create new solutions are called sons (David, 1989) [3] (Michalewicz, 1996) [39, 40]. After many generations, begins the solutions proposed begins approach each other with the hope that this is closer to the ideal solution. The most important characteristic of genetic algorithms from other methods of solution, where they produce a lot of solutions to the problem of time and then to improve the Crossover and mutations in an attempt to reach the best solution to the problem under research (Michalewicz, 1996) [16].

This is done by choosing the number of solutions the primary either randomly or by using one of algorithms approximation against known, then the subject of these solutions to the operations of the mutation and the Crossover which leads to better solutions or degradation, which requires the evaluation process help to choose good solutions based on the degree of efficiency, in order to reduce the bad solutions and access to more appropriate chromosomes.

B. The Mechanism of Genetic Algorithms

Genetic algorithms depend in his work on the adaptive methods and the stochastic search method used and the possibilities in the process of local search and un local.

C. Main Components of the Genetic Algorithms

Identifying the components of the genetic algorithms of the basic things, and the main components of these algorithms are:

- ✓ development solutions tools (Operators)
- ✓ mechanism for the representation of chromosomes (Solution Representation)
- ✓ mechanism for the establishment of elementary solutions (Initialization Procedure)
- ✓ Assessment mechanism resulting solutions (Evaluation Function).

D. Importance of Genetic Algorithms

Any researcher can take advantage of genetic algorithms, as long as he can encrypt the problem solving being addressed and put that solution in the chromosome, creating a function to evaluate solutions generated, which is a key to success in genetic algorithms comes from the ease of dealing with it, the time of their ability to find good solutions and quick to the problems incurable. And genetic algorithms are effective when:

- ✓ Research in a large and complex problem, or when it is difficult to understand.
- ✓ Be difficulty in reducing the search domain to a certain extent can be dealt with by traditional methods.
- ✓ There are no ways of mathematical solving
- ✓ Traditional search methods fail.

Have used genetic algorithms to solve many problems such as:

- ✓ Improvement problems to design electrical circuits (Louis, 1997) [11] and job scheduling (Goldstein, 1991)[6] and travelling salesman problem.
- ✓ Automatic programming.
- ✓ Machine learning, genetic algorithms has been used in many of these areas such as classification, prediction and design of control circuits.
- ✓ Economic models, development models and strategies of supply demand and the integration of economic markets.

Interaction between evolution and learning. Genetic algorithms have been used to demonstrate the ability of people to learn and their ability to influence others.

III. Previous Studies

A. Nagata and Kobayashi Study

This study (Nagata and Kobayashi, 1997) [12] provided a solution to the problem of travelling salesman where the proposed Crossover mechanism for the genetic algorithms called (Edge Aassembly Crossover EAX), What distinguishes this study, it was able, using an edge assembly crossover EAX and without any algorithm local search solving the problem of travelling salesman until (3038) city and access to the ideal solution, but with implementation time over of the two and a half (Watson et al, 1998) [20] .

B. Brady Study

Bardy (Bardy, 1985) [1] adopted in his study on the local optimization algorithms, and use the algorithm (2-opt) for this purpose.

Where the following steps declare Bardy Algorithm:

1) Generate a population of K starting solutions S = {P1 ... PK}.

2) Apply 2-opt algorithm to each solution P in S., Letting the resulting locally optimal solution replace Pin S.

3) While not yet converged do the following:

3.1. Select k' distinct subsets of S of size 1 or 2 as parents (the mating strategy).

3.2. For each 1-element subset, perform a randomized mutation operation to obtain a new solution.

3.3. For each 2-element subset, perform a (possibly randomized) crossover operation to obtain a new solution that reflects aspects of bath parents.

3.4. Apply 2-opt algorithm to each of the k' solutions produced in step 3.3, and let S' be the set of resulting solutions.

3.5. Using a selection strategy, choose K survivors from PU S', and replace the contents of S by these survivors.

4) Return the best solution in S.

C. Freiseleben and Study

This study (Freiseleben and Merz, 1997) [4] proposed Genetic local search algorithm with high efficiency to solve travelling salesman Problem based on the integration of genetic algorithms with the algorithm of Lin - Kernighan also proposed a new Crossover mechanism (Distance Preserving Crossover, DPX), where the sons born between them and each of the parents is equal distances, i.e., that the distances between the Son and the first Father are the same as the distance between the Son and the second Father.

It is notes her that using of Lin – Kernighan algorithm increase the efficiency of the results, on the other hand increasing the execution time by 80% .

D. Results of Previous Studies

From previous studies we note the following: -

- ✓ Use of local search algorithms to solve the problem of travelling salesman will give an appropriate solution, but it is often located in a local ideal solution, then there is a noticeable increase in execution time.
- ✓ In most studies have been dispensed from the representation of chromosomes(cities) the way the simple binary numbers and replaced so that the round shape is the order of the actual cities, even though the operations of Crossover and the mutation is difficult.
- ✓ Been developed more than style of the Crossover but these methods did not find the ideal solution, especially when the number of cities increase, also need high space in the memory and more execution or implementation time.

IV. Proposed Algorithm for solving Travelling Salesman Problem

A. Search for Neighbor Solution Algorithm

Search for neighbor solution Algorithm Been proposed search algorithm for solving travelling salesman problem through genetic algorithms, which rely on a simple modification of the round, so that the re-arrangement of the sites of some cities for new rounds and keep optimal of them, and repeat the switch mechanism by (the number of round cities- 3) which mean that this algorithm does not need a long time to implement them. And viewed this process as a search for solutions in the vicinity of the rounds the current round, which may be one of them better than the original round.

The mechanism of this algorithm as following:

1) Cancel the path between the second city and the third city in the tour.

2) Re-arranged the tour to reverse direction of the cities on the tour starting from the end of the tour and ended by the third city (zone of separation in the tour).

3) Round length is calculated by comparing the length of the canceled tour lengths with the new tour, in order to avoid re-collection of lengths of all tracks in order to reduce the execution time of the algorithm.

4) If the efficiency of the new round is best than the original round, is retained in the provider population otherwise they be ignored.

5) Repeat the previous mechanism to separate the path between the third city and the fourth city, and then between the fourth city and the fifth city, and so on until the separation between the city and the city before the recent.

B. Proposed Crossover Mechanism

Each chromosome in the travelling salesman represents a solution to the problem. In this mechanism is to keep as much as possible of the tracks between cities and create a new paths do not exist in the rounds of parents, or even all the chromosomes in the provider population. This is opposed to most of the mechanics of the previous Crossover, which is based on inheritance as much as possible paths of parents to children, and without relying on previous information.

To apply the Proposed Crossover mechanism, two chromosomes are selected from the provider population, such that the first chromosome is the owner of the best efficiency until know, through the implementation of genetic algorithm (the best local solution), and selection of second chromosome randomly, thus the efficiency of first chromosome always the owner of the efficiency of the best local solution. On the assumption that the efficiency of the first chromosome with the best local solution has some qualities in common with the general chromosome with the best efficiency, we can also assume that the second chromosome may exhibit some qualities in common with the general chromosome with the best efficiency. So both chromosomes can similar in many paths, and tracks with similar characteristics are dominant in the total of the best parts of the general solution.

Based on previous assumptions, in the proposed Crossover mechanism to keep as much as possible from the common paths between the cities in the chromosomes of parents and the transfer of these tracks as they are to each child chromosome, in order to benefit from as much in common information in this issue. The uncommon tracks are being introduced in new common paths to choose the city closest to the city starting from the current remaining cities, hoping to get a new round with the best qualities of efficient and higher efficiency of the parents to continue the process to reach the best general solution.

Proposed Crossover mechanism actions are as follows:

a) Choose two chromosomes from the provider population to configure the parents such that the first chromosome to be the owner of the local best efficiency, and the second chromosome are selected randomly.

b) Choose a city tour of cities at random and placed in the output son chromosome and considered it the starting point.

c) Find the location of this city in both rounds.

d) Identify the neighboring towns of the city beginning in the first chromosome and determine if these cities had been previously visited or not?

e) If you were to visit the neighboring towns of the city is already beginning to go a step (g), but neighboring cities are selected for the city beginning in the second chromosome and compared these cities with cities adjacent to the city beginning in the first chromosome to determine the existence of any common paths.

f) If you find common paths between the cities of the first chromosome and the cities of the second chromosome, is the joint selection of the city and add to child chromosome and the mind as a new beginning.

g) In the absence of any tracks shared between the city of the beginning and nearby towns in the chromosome I and chromosome II or the neighboring towns of the city first had been previously visited, is selection of the city closest to the Town of the Beginning of the cities that have not been had yet to be added to the Son chromosome, and be the point a new beginning.

h) Repeat steps (c - g) to be completed the round.

Simply the Proposed Crossover mechanism as follows:

Choose two chromosomes, P1 is the local optima, P2 is chosen randomly Choose a starting city i randomly from the cities in the tour.

- Do the following until the stopping criteria matches. Find the position of the starting city in both p1 and p2.
- Make sure that the neighboring cities are not visited yet.
- If this is an unvisited city in common between the neighboring cities in p1 and p2, connect the starting city to it.
- Otherwise, find the nearest unvisited city to the starting city, and make it the new starting city.

C. Proposed Genetic Algorithm

The mechanism of the proposed genetic algorithm is as follows:

1) Create a primary provider population by the max limit.

2) Compute the efficiency of all chromosomes in the provider population.

3) Apply of the proposed local search algorithm on all chromosomes in the provider population.

4) Choose a chromosome with a better efficiency to put it in the provider population to cancel the worse the efficiency of the chromosome and replaced with a new chromosome with better efficiency, provided they do not repeat any chromosome in population provider.

5) Make sure the condition of the stay, the check expires.

6) Choose a chromosome with the best efficiency p1, and choose another chromosome randomly p2.

7) Apply the new Crossover mechanism to get a new chromosome.

8) Apply the local improvement algorithm on output chromosome.

9) Apply the mutation on the Son chromosome as the ratio of installed.

*10)*Compute the efficiency of the resulting chromosomes.

*11)*Put chromosomes that carry a better efficiency in the population provider.

*12)*Return the population provider to its original position.

*13)*Repeat the process from step 5 to step No. 12

Also see flowchart of the proposed Genetic Algorithm in Fig. 2 below.

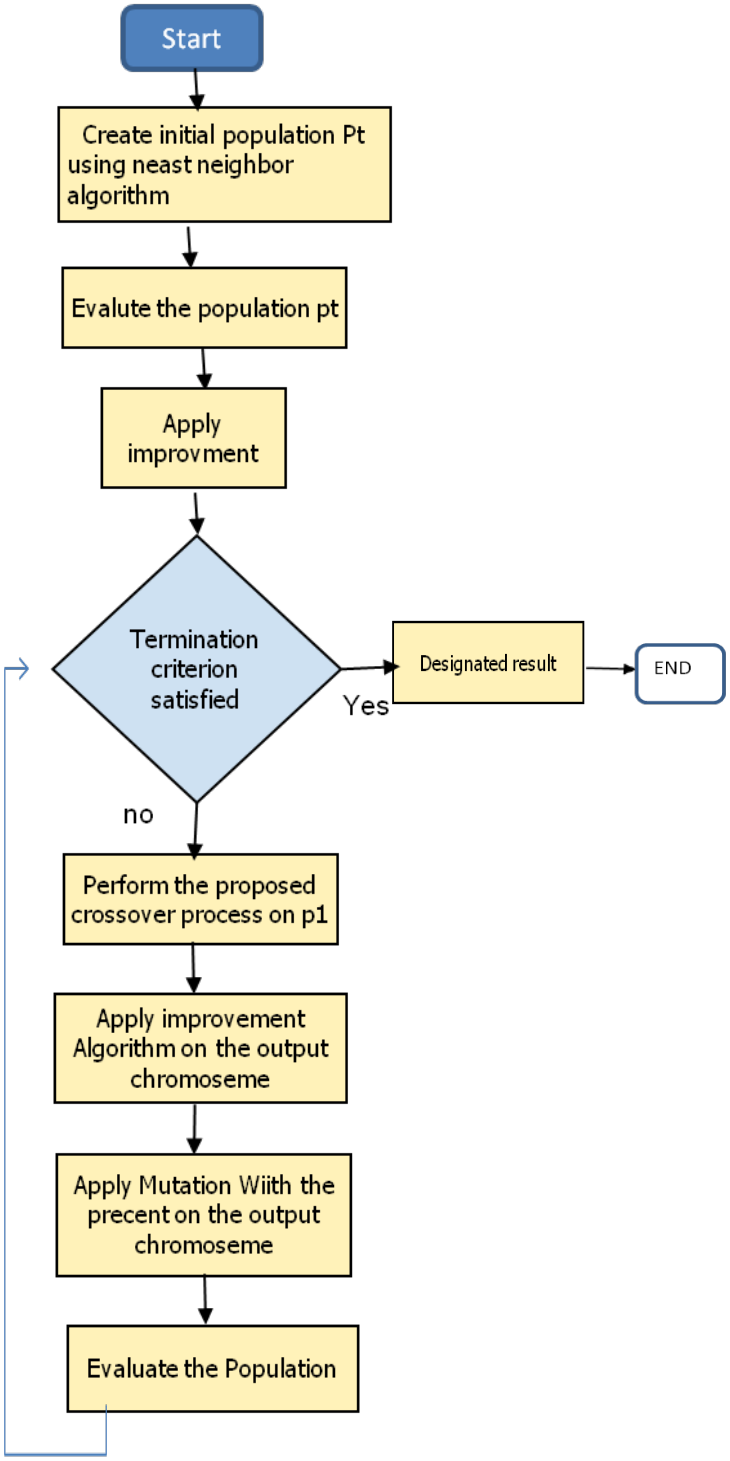

Start
Create initial population Pt using neast neighbor algorithm
Evalute the population pt
Apply improvment
Termination criterion satisfied
Yes
Designated result
END
no
Perform the proposed crossover process on p1
Apply improvement Algorithm on the output chromoseme
Apply Mutation Wiith the precent on the output chromoseme
Evaluate the Population

Fig 2. Proposed Genetic Algorithm.

V. Results and Comparisons

A. Results of Experiments

Genetic algorithm and local search algorithm to solve the travelling salesman problem implement in two forms, namely:

1) Individual form: Genetic algorithm works in single form without any local improvement local mechanism, and for comparing some of Crossover mechanism from previous studies are implemented.

2) Hybrid form: in which the local search algorithm works as a tool in the development of genetic algorithm to solve the research problem.

Individual and hybrid forms are been selected of the following objectives:

1) Define the ability of the proposed Crossover mechanism to solve the research problem.

2) show the importance of the introduce of the local improvement algorithm on the genetic algorithm

3) Note the difference between the results when the genetic algorithm works as Individual format and the hybrid

4) Determine the effect of the proposed Crossing on the performance of individual genetic algorithms by comparing the number of generations and time spent in twice.

5) Compare the different results obtained from the proposed Crossover mechanism with the results of previous studies.

these algorithms Implemented using a programming language (C + +) and carried out all experiments on a Pentium 4 (Pentium IV), which operates 1000 MHz frequency and main memory (RAM) is equal to 128 kilobytes. The program was applied in each case ten times and taking the arithmetic average.

Genetic algorithm was implemented on some of the issues the proposed standard (Benchmark) as shown in Table I, obtained from the library of a sales trip ((http:// ,,, TSPLIB) www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB/tsp)

TABLE I. Benchmark Problems

Benchmark	Cities Number	Optimal solution
A280	280	2579
Berlin52	52	7542
Ch130	130	6110
Ch150	150	6528
D1291	1291	50801
Gil262	262	2378
Rat195	195	2323
Rat575	575	6773
Tsp225	225	3916

B. Comparisons

Table II to Table VII show the practical results obtained for a sales trip in the following standard questions, Berlin52, Ch130, Ch150, Rat195, Tsp225, Gil262, A280, Rat575, D1291, using closest neighbor algorithm to the solution-building school. Known percentage of the solution resulting from the proposed genetic algorithm optimization of the solution according to the following equation:

The proportion of approaching the solution of the resulting solution =

Solution resulting from the genetic algorithm / Standard solution (Benchmark)

Table II and Fig. 3 show the result of proposed genetic algorithm using nearest neighbor algorithm to formulate primary population provider and apply proposed crossover mechanism with research for neighbor solutions for local development algorithm.

Table III and Fig. 4 shows the result of proposed genetic algorithm using nearest neighbor algorithm to formulate primary population provider and apply proposed crossover mechanism without implementation of research for neighbor solutions for local development algorithm ,which shows form the result the increasing of number of generations and faraway from optimal solution comparing with late table .

Table IV and Fig. 5 show the result of proposed genetic algorithm using nearest neighbor algorithm and using Operation Crossover (OX), which shows form the result the increasing of number of generations and faraway from optimal solution comparing with late Table II.

TABLE II. Proposed GA output use Nearest GA with New Crossover + NLSA

		Output solution	Percentage to	Execution time
Benchmarks	Number of iteration	New Cross + NLSA	optimal solution	hh:mm:ss
A280	6500	2604	99%	00:00:48
Berlin52	50	7544	100%	< 1
Ch130	4800	6181	99%	00:00:11
Ch150	4920	6557	99%	00:00:12
D1291	25000	53055	96%	00:09:35
Gi1262	9702	2416	98%	00:00:45
Rat195	2460	2346	99%	00:00:09
Rat575	18021	7013	97%	00:04:56
Tsp225	4750	3916	100%	00:00:23

TABLE III. Proposed GA output use Nearest GA with New Crossover + without NLSA

		Output solution +	Percentage to	Execution time
Benchmarks	Number of iteration	New Cross + without NLSA	optimal solution	hh:mm:ss
A280	86240	2781	96%	00:02:02
Berlin52	1900	7544	100%	00:00:02
Ch130	53940	6357	96%	00:00:24
Ch150	31970	6577	99%	00:00:21
D1291	90316	55427	92%	00:07:23
Gi1262	90934	2472	96%	00:01:52
Rat195	23290	2371	98%	00:00:20
Rat575	90096	7481	90%	00:03:53
Tsp225	31250	3973	98%	00:00:30

TABLE IV. GA OUTPUT USE OPERATION CROSSOVER (OX) + NEAR LOCAL SEARCH ALGORITHM (NLSA)

		Output solution +	Percentage to	Execution time
Benchmark	Number of iteration	OX+(NLSA)	optimal solution	hh:mm:ss
A280	21390	2890	90%	00:02:12
Berlin52	960	7549	99%	00:00:11
Ch130	8955	6402	97%	00:00:36
Ch150	6616	6872	95%	00:00:30
D1291	15000	57277	89%	00:06:36
Gi1262	19790	2591	92%	00:01:62
Rat195	6035	2388	98%	00:00:29
Rat575	36891	7483	91%	00:03:05
Tsp225	9393	4143	95%	00:00:35

TABLE V. GA OUTPUT USE OX WITH + WITHOUT NEAR LOCAL SEARCH ALGORITHM (NLSA)

Benchmark	Number of Iteration	Output solution +	Percentage to	Execution time
		OX +Without(NLSA)	optimal solution	hh:mm:ss
A280	89159	3066	84%	00:00:49
Berlin52	88375	7825	96%	00:00:13
Ch130	93083	6781	90%	00:00:16
Ch150	87747	6985	93%	00:00:16
D1291	100000	59108	86%	00:02:21
Gi1262	98734	2726	87%	00:00:35
Rat195	88195	2551	91%	00:00:17
Rat575	91257	8216	83%	00:00:58
Tsp225	90327	4673	83%	00:00:14

TABLE VI. GA OUTPUT USE MAPPED PARTIALLY CROSSOVER (MPX) WITH + NLSA

		Output solution	Percentage to	Execution time
Benchmark	Number of iteration	MPX + NLSA	optimal solution	hh:mm:ss
A280	23715	2881	90%	00:00:48
Berlin52	630	7544	100%	00:00:08
Ch130	9070	6217	98%	00:00:11
Ch150	8310	6565	99%	00:00:12
D1291	15000	56159	90%	00:04:41
Gi1262	21702	2531	94%	00:00:36
Rat195	12625	2367	98%	00:00:14
Rat575	28150	7403	91%	00:01:45
Tsp225	8447	4024	97%	00:00:18

Table V and Fig. 6 show the result of proposed genetic algorithm using nearest neighbor algorithm and using Operation Crossover (OX), without NLSA, for comparing results.

Table VI and Fig. 7 show the result of proposed genetic algorithm using Mapped Partially Crossover (MPX), with NLSA, which show Convergence of results.

TABLE VII. GA OUTPUT USE MPX WITH + WITHOUT NLSA

		Output solution	Percentage to	Execution time
Benchmark	Number of iteration	MPX+NLSA	optimal solution	hh:mm:ss
A280	88897	2949	87%	00:00:47
Berlin52	3640	7648	99%	00:00:11
Ch130	77474	6725	91%	00:00:12
Ch150	75583	6996	93%	00:00:13
D1291	100000	60024	84%	00:02:17
Gi1262	99858	2710	88%	00:00:26
Rat195	94661	2494	93%	00:00:17
Rat575	83152	8136	83%	00:01:47
Tsp225	89657	4504	87%	00:00:18

Table VII and Fig. 8 shows the result of proposed genetic algorithm using Mapped Partially Crossover (MPX), without NLSA, which show Convergence of results at Table V, but it is worse than result in Table II.

From the previous tables and when fixing crossover percentage and mutation for all problems, we noting the following:

1) The quality of result solutions from implementation of genetic algorithm on the benchmarks for travelling sales man approaching form the general optimal solution when using the proposed crossover mechanism without using any mechanisms solutions for local development algorithm.

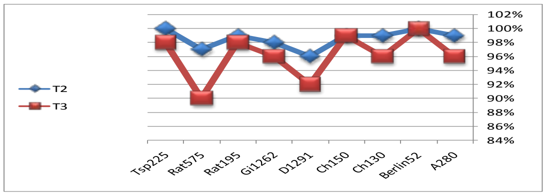


Fig 3. Proposed GA Output use Nearest GA with New Crossover + NLSA.

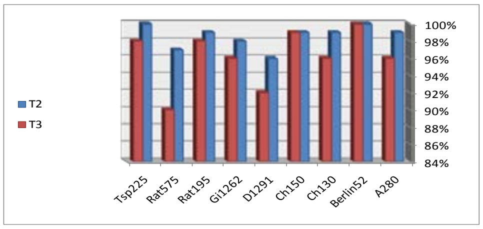


Fig 4. Proposed GA Output use Nearest GA with New Crossover + without NLSA.

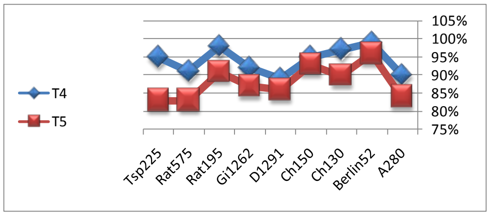


Fig 5. GA Output use Operation Crossover (OX) + NearLocal Search Algorithm (NLSA).

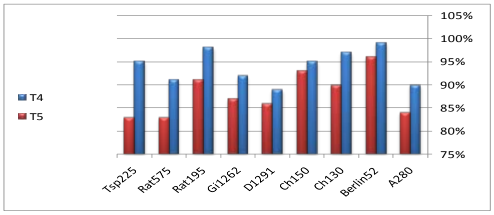


Fig 6. GA Output use OX with + without Near Local Search Algorithm (NLSA).

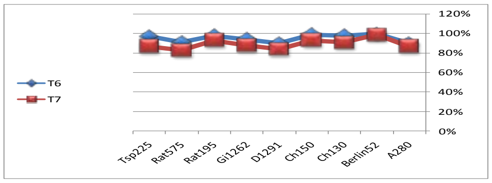


Fig 7. GA Output use Mapped Partially Crossover (MPX) with + NLSA.

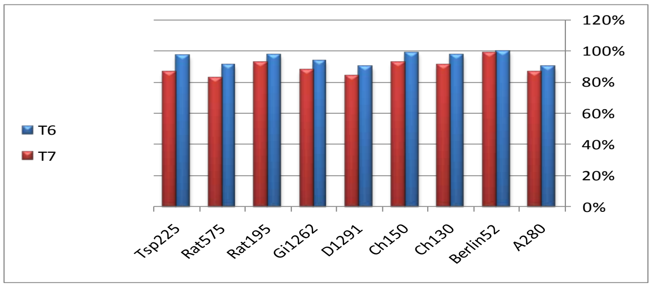


Fig 8. GA Output use MPX with + without NLSA.

2) By use local improvement algorithm on genetic algorithm by uses MPX: the percentage of performance solution is increase as shown below in Fig. 9 and Fig. 10:

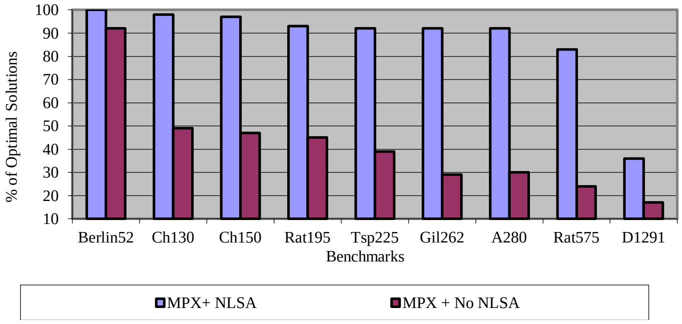


Fig 9. Compare (MPX+ NLSA) & (MPX+ No NLSA) Optimal Solution.

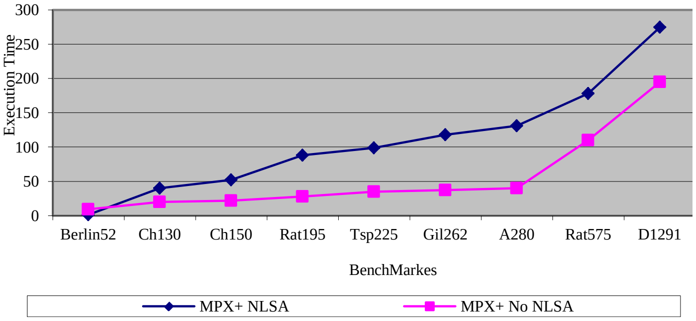


Fig 10. Compare (MPX+ NLSA) & (MPX+ No NLSA) Execution Time.

We applied genetic algorithm using new crossover technique and OX and MPX consider the follow-:

- Not use any of NLSA algorithms
- Order initiate tour randomly.
- Use same percentage (0.01, 0.8,10000) ,

as shown the following in Fig. 11.

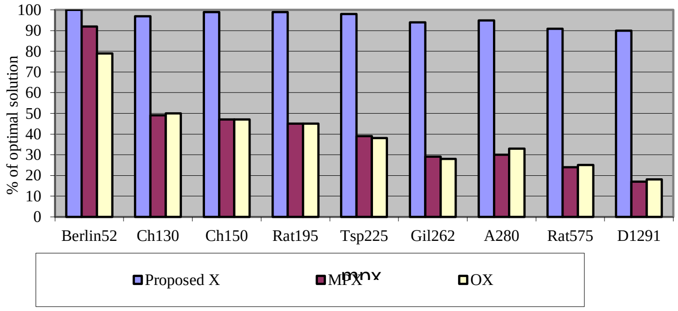


Fig 11. Compare (Proposed X) & (MPX) & (OX) Optimal Solution.

*New crossover(p) is the shortest execution time for genetic algorithm .The number of iteration for optimal solution is lowest for new crossover as shown in Fig. 12 below:-

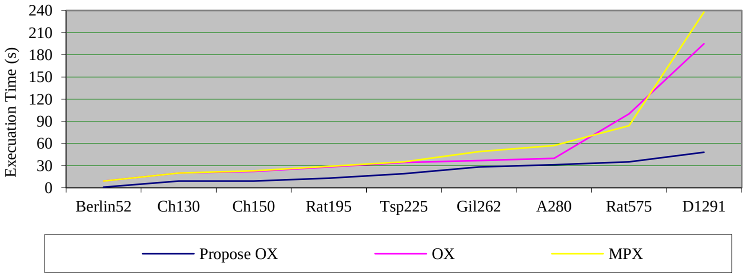


Fig 12. Compare (Proposed X) & (MPX) & (OX) Execution Time.

When we inject genetic algorithm by optimization algorithms we not the following-:

1) Increase the chance to have optimal solution.

2) Increase the result solution performance.

3) Increase in the execution time comparison by improvement quality of solutions.

As shown in Fig. 13 and Fig. 14 below:

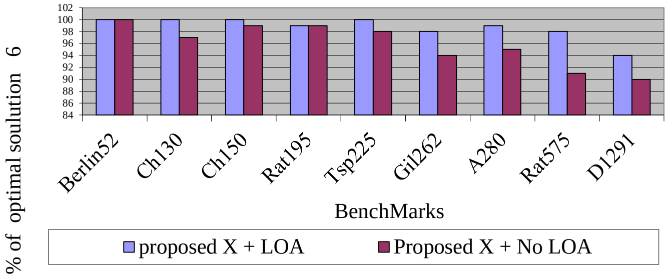


Fig 13. Compare (Proposed X+LOA) & (proposed X+ No LOA) Optimal Solution.

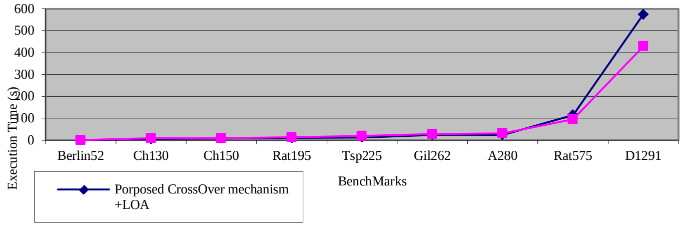


Fig 14. Compare (Proposed X+LOA) & (proposed X+ No LOA) Execution Time.

VI. Conclusions

This paper Suggested new Crossover mechanism to solve travelling salesman problem using genetic algorithms, has also proposed anew local improve algorithm to naming search for neighbor solutions algorithm entered on genetic algorithms to improve its performance and get better results and less and an increase of up to (10%) of the time implementation required to obtain the final result compared with other algorithms that have been comparison.

The following points can be drawn from this paper:

1) The proposed genetic algorithm addressed the travelling salesman problem and has a solution up to the percentage (100%) of standard solutions in some of the issues and to the proportion of (96% -100%) of the ideal solution in others matters.

2) Solve the travelling salesman problem using a proposed Crossover mechanism in genetic algorithms gave better results than in previous studies and reasonable time relative to the time required in the crossover mechanisms of the other private and tests were performed on the benchmark, number of cities where large up to (1291) city.

3) using proposed Local optimization algorithm as a tool to enhance the performance of genetic algorithms using different Crossover mechanisms raise the result efficiency and reduced the number of generations needed to get to the final solution.

4) Despite the increase in execution time when you enter local optimization algorithm on genetic algorithm, but this increase raised the efficiency of the final solution and significantly.

Acknowledgments

The researcher gratefully acknowledges and highly appreciates the financial support and the remarkable resources provided by WISE University, Jordan.

References

[1] Bardy, 1985, Optimization Strategies Gleaned from Biologic Evolution. Nature, 317: 804-806.

[2] Burke, 1995, A Hybrid Genetic Algorithm for Highly Constrained Timetabling problem, Proc. Of the 6th Int. Genetic Algorithms, San Francisco, Morgan Kaufmann, 1995.

[3] David, 1989, Proceedings of the 3rd International Conference on Genetic Algorithms, George Mason University, Fairfax, Virginia, USA, June 1989 ICGA 1989.

[4] Freiseleben and Merz, 1997, Genetic Local Search for the TSP: New Results. In T. Bäck, Z. Michalewicz, and X. Yao, eds., Proceedings of the 1997 IEEE International Conference on Evolutionary Computation, 159-164, Piscataway,NJ, IEEE

[5] Goldberg D. 1989, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, 1989.

[6] Goldstein, 1991, Genetic Algorithm Simulation of the SHOP Scheduling Problem, by Jonathan M.Goldstein, in September 1991 published by An ICMS/Shell Oil Business Consultancy.

[7] Hoffman, 1991, M., Improving LP-representations of zero-one linear programs for branch-and cut. ORSA Journal of Computing, 1991, 3(2), 121—134.

[8] Johnson David and L.A McGeoch, 1998, “The Travel Salesman Problem: A Case Study in Local Optimization”, chapter in the book Local Search in Combinatorial Optimization edited by E.H.L. Arts and J.K. Lenstra. Vol. E81A, no.5, 1998, pp. 738-750.

[9] John Holland's 1992, Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence, University of Michigan Press, 1975 (second edition: MIT Press, 1992).

[10] Lalena, 2003, Travelling Salesman Problem Using Genetic Algorithms, http;// www.lalena.com/ai/tsp.

[11] Louis S. 1997, Genetic Algorithms with Memory for Travelling Salesman Problem, 1997.

[12] Martin O., S.W. Otto, and W. Feten, “Large-Step Marcov Chains for the TSP incorporating Local Search Heuristics”, Operations Research letters, vol. 11, 1992, pp. 219-224.

[13] Man, K. F. 1999, Tang, K. S. and Kwong, S., Genetic Algorithms concepts and Designs, Springer-Verlag London, Great Britain, 1999.

[14] Michalewicz Z. 1999, Genetic Algorithms + Data Structures = Evolution Programs, 3rd edition, Springer-Verlag Berlin Heidelberg, New York, 1999.

[15] Michal Genetic Algorithms, Numerical Optimization and Constraints, Proceeding Michalewicz, 1995, s of the 6th International Conference on Genetic Algorithms, Pittsburgh, July 15-19, 1995, pp. 151-158.

[16] Michalewicz, 1996, M., Evolutionary Algorithms for Constrained Parameter Optimization Problems, Evolutionary Computation, Vol.4, No.1, 1996, pp.1-32

[17] Mitchell, 1996, an Introduction to Genetic Algorithms, MIT Press, 1996 New York

[18] Newall, 1998, Hybrid Methods for Automated Timetable, Ph.D. Thesis, University of Nottingham, 1998.

[19] Obitko, 1998 Obitko, M. (1998). V. Operators of GA. Accessed from

[20] Schaerf A., “Tabu Search Techniques for Large High-School Times tabling”, Proceedings of the 13th American Conference on Artificial Intelligence, 1996.

[21] Katayama and Narihisa, 1999, Iterated Local Search Approach Using Gene Transformation to the Travelling Salesman Problem. In W. Banzhaf, ed., Proceedings of the Genetic and Evolutionary Computation Conference, 321-328, Morgan Kaufmann, 1999.

Efficiency and Performance of Optimized Robust Controllers in Hydraulic System

Chong Chee Soon[1], Rozaimi Ghazali*[2], Shin Horng Chong[3], Chai Mau Shern[4], Yahaya Md. Sam[5], Zulfatman Has[6]

Centre for Robotics and Industrial Automation, Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia[1, 2, 3, 4]

Department of Control and Mechatronics Engineering, School of Electrical Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia[5]

Electrical Engineering Department, University of Muhammadiyah Malang, 65144 Malang, Indonesia[6]

***Abstract*—Common applications involved hydraulic system in real-time including heavy machinery, air-craft system, and transportation. Real-time applications, however, are notorious to suffered with nonlinearities due to unavoidable mechanical structures. Thus, control system engaged to manipulate and minimize the effect resulting from the nonlinearities. In this paper, few control approaches are executed in the Electro-Hydraulic Actuator (EHA) system. The control approaches including the generally used Proportional-Integral-Derivative (PID) controller, the reinforced Fractional-Order (FO) PID controller, the Sliding Mode Controller (SMC), and the enhanced hybrid SMC-PID controller. In order to obtain proper parameter of each controller, the Particle Swarm Optimization (PSO) technique is applied. The output data are then analysed based on the performance indices in terms of the consumption of the energy and the error produced. The performance indices including Root Mean Square Error/Voltage (RMSE/V), Integral Square Error/Voltage (ISE/V), Integral Time Square Error/Voltage (ITSE/V), Integral Absolute Error/Voltage (IAE/V), and Integral Time Absolute Error/Voltage (ITAE/V). It is observed in the results, based on the performance indices in terms of error and voltage, the hybrid SMC-PID capable of generating better outcomes with reference to tracking capabilities and energy usage.**

Keywords*—*Robust Control Design; optimization; tracking efficiency analysis; controller effort analysis; electro-hydraulic actuator system

I. Introduction

The behavior of a system in industrial sector is highly connected with the control system that manage and manipulated the end function of particular machineries. Practical mechanical machineries such as Electro-Hydraulic Actuator (EHA) system, however, commonly undergoes the well-known uncertainties and nonlinearities characteristics [1]. These characteristics generating direct impact to the machine over time and increase the complexities in the controller design. Improper design of the controller system may lead to the common problem such as imprecise and energy efficiency.

In hydraulic, a mathematical model that represent the physical behaviour of this system is difficult to be modelled due to the uncertainties and nonlinearities characteristics [2]. Usually, the compressibility of the oil due to the working temperature, the internal and external friction of the cylinder, the fluid flow characteristic in valve and cylinder are the trait of the nonlinearities. While two primary form of uncertainties are discovered in the past including parametric uncertainties and uncertain nonlinearities [3]. Therefore, control system raised to dealing with these issues.

Over decades, it is famed in the control and industry fields about the Proportional-Integral-Derivative (PID) controller implementation in the EHA system [4-5]. This controller became a topic of interest by researchers and academia on the basis of practical and user-friendly advantage. For years, varieties of approaches have been inserted to this controller. One of the typical approaches that often seen is the modification in terms of the controller's structure. For example, the fractional order control and the gain scheduling control that have been generally integrated with the PID controller [6-7]. These methods are proven to have more efficient performance compared with the traditional PID controller.

Adaptability and robustness are the major concern by control engineers. When addressing these features, the outstanding Sliding Mode Control (SMC) approach performed significant achievement applied in different machineries [8-9]. Unlike the PID controller, SMC is identified to commonly gain its parameter by means of try and error process [10-12]. Therefore, computer based tuning approaches including Particle Swarm Optimization (PSO), Gravitational Search Algorithm (GSA), and Genetic Algorithm (GA) have been gradually attract the attention of the control field due to their prominent performance in searching for the controller's optimal parameter [13-15].

It is noticed in the literature, EHA system is mostly executed in positioning control. The accurate positioning control are usually required in the applications such as vehicle [16], robotic [17], construction machinery [18], and aerodynamic [19]. Owing to this matter, this paper aims to develop several control approaches for the purpose of comparison in the manner of positioning performance. Additionally, further analyses in terms of precision and energy usage based on the performance indices, including Root Mean Square Error/Voltage (RMSE/V), Integral Square Error/Voltage (ISE/V), Integral Time Square Error/Voltage (ITSE/V), Integral Absolute Error/Voltage (IAE/V), and

The support of Universiti Teknikal Malaysia Melaka (UTeM) is greatly acknowledged. The research was funded by UTeM Zamalah Scheme.

Integral Time Absolute Error/Voltage (ITAE/V) are carried out.

This paper is organized as, modelling and control methods are briefly discussed in Section 2. The performance of each controller is presented in Section 3. Lastly, the summary of the outcome is drawn in Section 4.

II. MATHEMATICAL MODELLING

This study utilizes the valve operated transmission system instead of pump operated transmission system due to the efficiency [20]. Refer to Fig. 1, the general structure of the EHA system composed of valve and actuator units, sensing unit, and computer control unit.

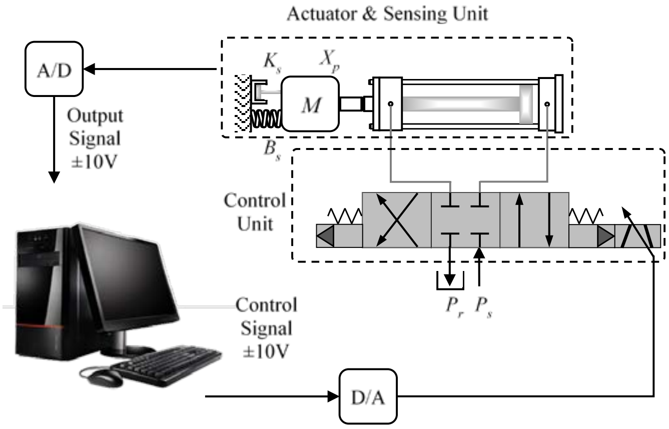


Fig 1. General Component in the EHA System.

The motor responsible to drive the spool in servo-valve. The voltage, V_v that drive the current, I_v flow to the coil that connected to the servo-valve generating torque of the motor as expressed in (1).

$$V_v = \frac{dI_v}{dt}L_c + R_c I_v \tag{1}$$

where the coil consists of inductance and resistance as denoted in L_c and R_c respectively.

The voltage and current that generating torque to the motor concurrently produces dynamic motion to the servo-valve that represented in a second order differential equation as expressed in (2).

$$\frac{d^2x_v}{dt^2} + 2\xi_v\omega_v\frac{dx_v}{dt} + \omega_v^2 x_v = I_v\omega_v^{\ 2} \tag{2}$$

The spool-valve that controlling the flow rate, Q in chambers consist of different pressure, P_v, the position of the spool valve, x_v, and the gain of the servo-valve, K_v as written in (3).

$$Q = K_v x_v \sqrt{\Delta P_v} \tag{3}$$

The fluid flow characteristic in each chamber by neglecting the effect of internal leakage occurs in servo-valve can be expressed in (4) and (5).

$$Q_1 = \begin{cases} K_{v1}x_v\sqrt{P_s - P_1} & ; x_v \geq 0, \\ K_{v1}x_v\sqrt{P_1 - P_r} & ; x_v < 0, \end{cases} \tag{4}$$

$$Q_2 = \begin{cases} -K_{v2}x_v\sqrt{P_2 - P_r} & ; x_v \geq 0, \\ -K_{v2}x_v\sqrt{P_s - P_2} & ; x_v < 0, \end{cases} \tag{5}$$

where the servo-valve gain coefficient is assumed for a symmetrical valve as K_v=K_{v1}=K_{v2}.

The pump will generate fluid flow that simultaneously produce supply pressure, P_s that driving the servo-valve. At present, the EHA system generally equipped along with pressure regulator, which can adjust the maximum operating pressure supported by particular applications. The generated pressure will form the dynamics between the pump and the servo-valve can be expressed in (6).

$$P_s = \frac{\beta_e}{V_t}(Q_{pump} - Q_L)dt \tag{6}$$

A. PID and FOPID

It is widely known that PID controller composed with three parameters which playing vital role in transient response and steady-state response analyses. Commonly, PID controller contains the transfer function as expressed in (7).

$$\begin{aligned} G_s(s) &= K_p + \frac{K_i}{s} + K_d s \\ &= K_p(1 + \frac{1}{T_i s} + T_d s) \end{aligned} \tag{7}$$

where proportional gain represented by K_p, integral gain denoted by K_i and derivative gain defined by K_d. The performance of the transient response that proportional to the steady-state error response is handled by K_p. The responsibility in reduction or elimination of error in the steady-state through the compensation of low frequency executed by K_i. While the transient response performance is improved by K_d through the compensation of high frequency [21].

For the FOPID, in 20th century, a researcher named Igor Podlubny has introduced the Fractional Order calculus [22]. This calculus is introduced in control and dynamic system by extend general differential equations into fractional order differential equation [23]. The fractional order calculus has integrated to the PID controller due to its flexibilities that emerged Fractional Order (FO-PID) controller.

Alternatively, two additional gains have been merged in the fractional order controller that yielding five parameters which are K_p, K_i, K_d, λ, and μ denoted as proportional-integral-derivative-integral order, and derivative order respectively [23-25]. Generally, PID controller composed with the element of the transfer function as expressed in (7). Whereas the integration of the additional parameters from the fractional order calculus to the PID controller leads to the transfer function in (8).

$$G(s)=\frac{U(s)}{E(s)}=K_p\left(1+\frac{1}{T_i s^{\lambda}}+T_d s^{\mu}\right) \tag{8}$$

It is clearly perceived that if both λ and μ are presumed to be one, then conventional PID can be obtained. Closed-loop control system commonly composed with the interchangeable of integer or fractional order system and controller.

FO-PID or know as $PI^{\lambda}D$ controller emerged decades and proven to be capable to elevate the performance of the conventional PID controller. However, in some practical stand point, the additional parameters or the numbers of parameters that required to be achieved somehow affecting the computational burden.

B. Conventional and PSO Tuning Methods

Several conventional tuning approaches for the PID controller including Ziegler-Nichols, Tyreus-Luyben, Damped Oscillation, Chien, Hrones and Reswich, Cohen and Coon, Fertik, and Ciancone-Marline tuning methods [26]. This article, however, only covers Ziegler-Nichols tuning method due to it prominent performance. Generally, the gains of Ki and Kd are reduced to zero in the procedure of the Ziegler-Nichols tuning method.

In the output result where K_i and K_d are in zero condition, the K_p will be increased to the ultimate gain, K_u. In this state, the sustain oscillation occurred. Then, the period of the sustain oscillation, T_u is achieved through the calculation of a full wave cycle as depicted in Fig. 2. The PID controller gains K_p, K_i and K_d are then obtained through the formula as tabulated in Table I [27].

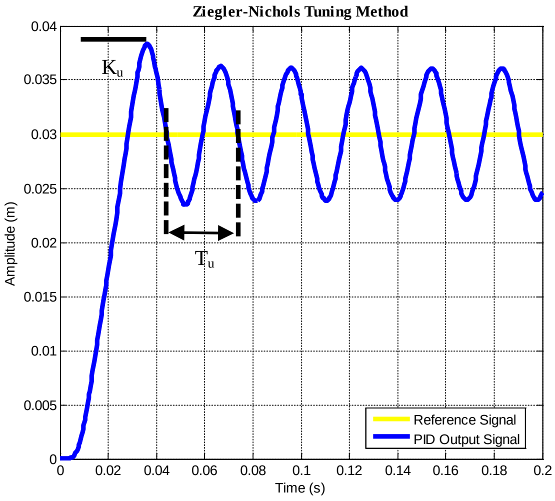


Fig 2. Sustain Oscillation to Obtain Ultimate Gain and Period.

TABLE I. GAINS FORMULATION IN ZIEGLER-NICHOLS TUNING METHOD

Gains Type	K_p	K_i	K_d
P	0.5 K_u	Inf	0
PI	0.45 K_u	T_u/1.2	0
PID	0.6 K_u	T_u/2	T_u/8

For the Particle Swarm Optimization (PSO), the establishment of this algorithm is summarized as depicted in Fig. 3. Generally speaking, the establishment operation begins with random parameter allocation of particle's velocity and position. These particles are then allocated to the region that existed with problem space, local and global borders, and started to engaging the problem space for the execution. The best solution or defined as fitness for a particle is came up after the execution in the local border and classified as local best value, *lbest*. The process repeated for each particle in seeking for their best fitness that consist of position and velocity value. These fitness value will then preserve in arrays and classified as personal best value, *pbest*. The operation repeated to the maximum criterion. When the maximum criteria are met, the excellent fitness value among these particles is finally named as global best value, *gbest*.

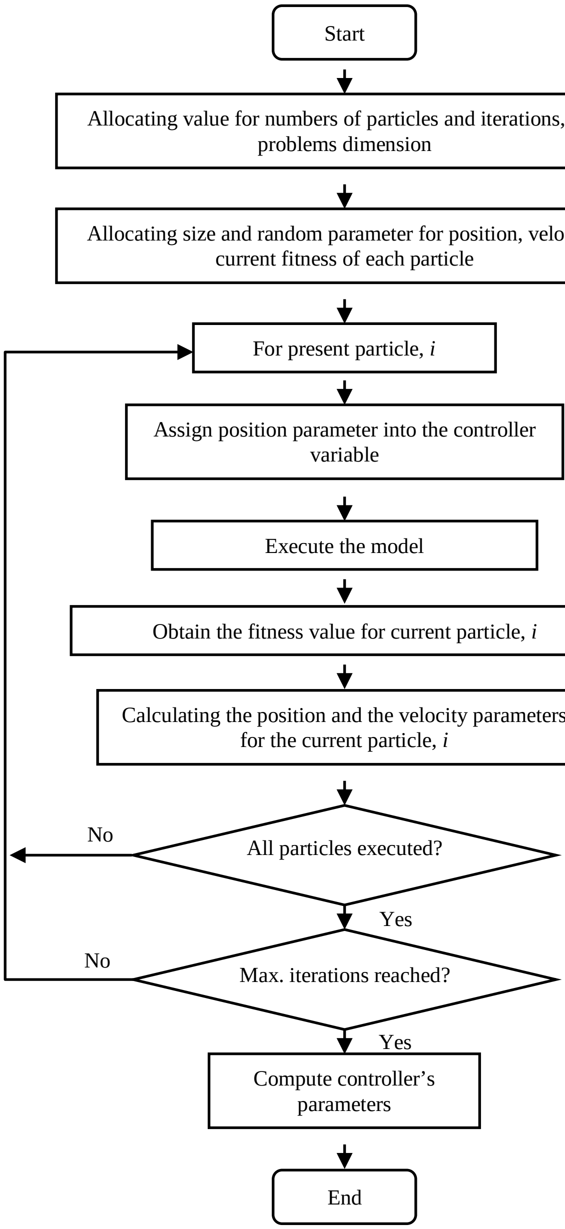

Start
Allocating value for numbers of particles and iterations, and problems dimension
Allocating size and random parameter for position, velocity, current fitness of each particle
For present particle, *i*
Assign position parameter into the controller variable
Execute the model
Obtain the fitness value for current particle, *i*
Calculating the position and the velocity parameters for the current particle, *i*
All particles executed?
No
Yes
Max. iterations reached?
No
Yes
Compute controller's parameters
End

Fig 3. General Procedures in the Establishment of the PSO Algorithm.

C. Conventional SMC and SMC with PID Sliding Surface

The outstanding work by Russia in early 1960's had introduced Sliding Mode Control (SMC) in a form of continuous time. Basically, the concept of this controller is to manipulate the control state heading the designed sliding surface. The control state will remain on the surface until the desired condition as depicted in Fig. 4.

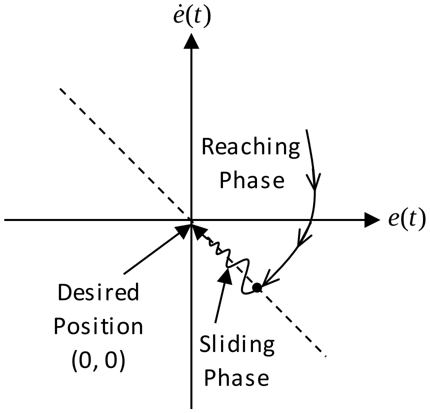


Fig 4. The Common Route to Obtain Desired Condition in SMC.

The system order, *n* is the important criterion in the design of the general sliding surface, s(t) in SMC as expressed in (9).

$$s(t) = \left(\lambda + \frac{d}{dt}\right)^{n-1} e(t) \tag{9}$$

In the SMC design, third order usually acquired for the EHA system. In the conventional SMC design, the sliding surface, s(t) is proportional to the error, *e* and the control gain, *λ* as expressed in (10).

$$s(t) = \ddot{e}(t) + 2\lambda\dot{e}(t) + \lambda^2 e(t) \tag{10}$$

By merging the PID controller in the conventional sliding surface design, following expression is obtained with the gains of PID controller K_p, K_i and K_d.

$$s(t) = K_p e(t) + K_i \int_0^t e(\tau)d\tau + K_d\dot{e}(t) \tag{11}$$

In the conditions where sliding surface is in reaching phase or s(t) ≠ 0, the switching control, u_{sw} undertake the role in carrying the control state to the sliding phase. When the control state reached the sliding phase or s(t) = 0, take over in guiding the control state to the desired point. Commonly, the design of the SMC composed the elements as expressed in (12).

$$u_{smc}(t) = u_{eq}(t) + u_{sw}(t) \tag{12}$$

In the ordinary SMC and the SMC-PID sliding surface design, first and second derivatives of the sliding surface can be obtained as expressed in (13) and (14) respectively.

$$\dot{s}(t) = \dddot{e}(t) + 2\lambda\ddot{e}(t) + \lambda^2\dot{e}(t) \tag{13}$$

$$\ddot{s}(t) = K_p\ddot{e}(t) + K_i\dot{e}(t) + K_d\dddot{e}(t) \tag{14}$$

In the simulation environment, the lumped uncertainties, L can be usually neglected in the design of the equivalent control, u_{eq}. Thus, the u_{eq} for the convention SMC and SMC-PID controllers can be defined in (15) and (16) accordingly.

$$u_{eq}(t) = \frac{1}{C}\left(\dddot{x}_r + A_n\ddot{x}_p + B_n\dot{x}_p + 2\lambda\ddot{e}(t) + \lambda^2 e(t)\right) \tag{15}$$

$$u_{eq}(t) = \left(K_d C_n\right)^{-1}\left(K_p\ddot{e}(t) + K_i\dot{e}(t) + K_d(\dddot{x}_r + A_p\ddot{x}_p + B_p\dot{x}_p)\right) \tag{16}$$

In the u_{sw} design, signum function, *sign*(*s*) that has a boundary border will be merged to the sliding surface. The u_{sw} and the boundary function is derived in (17) and (18) accordingly.

$$u_{sw}(t) = k_s sign(s) \tag{17}$$

$$sign\left(s(t)\right) = \begin{cases} 1 & ; s(t) > 0 \\ 0 & ; s(t) = 0 \\ -1 & ; s(t) < 0 \end{cases} \tag{18}$$

In SMC-PID switching control, u_{sw} design, the *sign*(*s*) that has a boundary border will also be merged to the sliding surface. The u_{sw} and the boundary function is derived in (19) and (20) accordingly.

$$u_{sw}(t) = \lambda s(t) + k_s sign\left(\dot{s}(t)\right) \tag{19}$$

$$sign\left(\dot{s}(t)\right) = \begin{cases} 1 & ; \dot{s}(t) > 0 \\ 0 & ; \dot{s}(t) = 0 \\ -1 & ; \dot{s}(t) < 0 \end{cases} \tag{20}$$

However, signum function is known to has discontinues state that lead to a chattering effect. Thus, authors in [28-30] have introduced a hyperbolic tangent function that can minimize the chattering effect. Then, the u_{sw} for both conventional SMC and SMC-PID controller can be obtained in (21) and (22) respectively.

$$u_{sw}(t) = k_s \tanh\left(\frac{\dot{s}}{\phi}\right) \tag{21}$$

$$u_{sw}(t) = \lambda s(t) + k_s \tanh\left(\frac{\dot{s}}{\phi}\right) \tag{22}$$

D. RMSE, ISE, ITSE, IAE and ITAE Analyses

In control system, few types of analyses are commonly used. These analyses including transient response analysis, steady-state error analysis, Root Mean Square Error (RMSE), Integral Square Error (ISE), Integral Time Square Error (ITSE), Integral Absolute Error (IAE), and Integral Time Absolute Error (ITAE).

Depending on the purposes of the article. In this paper, energy usage or efficiency and precision are the major concern which lead to the usage of RMSE, ISE/V, ITSE/V, IAE/V, and ITAE/V. Transient response and steady-state errors analyses are still can be used if necessary. Generally, the RMSE, ISE/V, ITSE/V, IAE/V, and ITAE/V can be expressed as below.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(r_i - y_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(e_i)^2}{n}} \tag{23}$$

$$ISE = \int_0^\infty e^2(t)dt \tag{24}$$

$$ITSE = \int_0^\infty e^2(t)tdt \tag{25}$$

$$IAE = \int_0^\infty |e(t)|dt \tag{26}$$

$$ITAE = \int_0^\infty |e(t)|tdt \tag{27}$$

III. Tracking Efficiency and Controller Effort Analyses

In this article, step input response has been applied in the performance examination of the designed controllers. Simulation works are implemented using MATLAB/Simulink 2018. By mimicking the real time environment that addressing fast response, all controller parameters are tuned only one time. It is inferred that, a well-designed control system not simply compensates the existing deficiency, but also able to mitigating an actual endeavour in operating a machine.

This article examines the tracking capability and actual effort in actuating the Electro-Hydraulic Actuator (EHA) system as aforementioned. The feedback of the designed controllers including PID, FOPID, SMC, and SMC-PID controller is portrayed in Fig. 5. In the figure, the SMC-PID is apparently outstanding with fastest response or rise time. This effect might be due to the increment of the numbers of gains in the SMC-PID controller. Referring to this, however, the FOPID that contained numbers of gains that not differ much than the SMC-PID controller showing poorest outcome even compared with the PID controller. This situation might be due to the defect of the tuning algorithm. Therefore, further analysis will be carried out on the tuning method and will implement in experiment environment to be more practical.

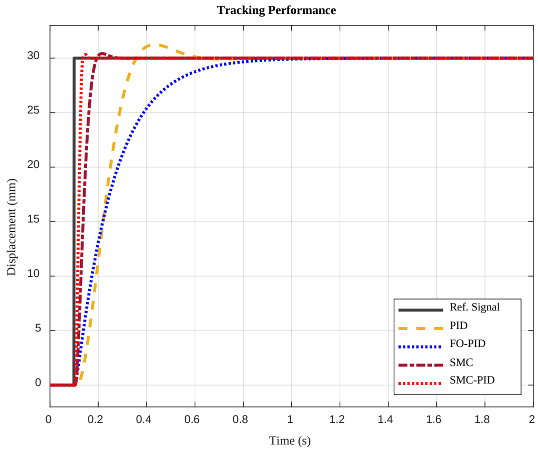


Fig 5. The Tracking Performance of the Designed PID, FO-PID, SMC and SMC-PID Controllers.

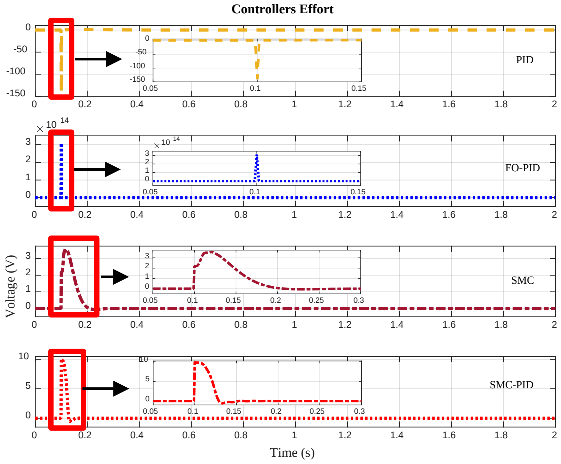


Fig 6. The Effort Required by Each Controller in Actuating the EHA System.

By taking the sample between the controller and the system, or particularly the output of the controller, and the input of the system, the control effort can be analyzed. Fig. 6 depicts the effort of the designed controllers in actuating the EHA system.

Referring to the numerical data shown in the *y* axis, the FOPID controller produced worst performance in terms of control effort, followed by PID, SMC-PID, and SMC controllers. All the feedbacks are executed according to the parameters obtained using PSO algorithm listed in Table II. To comprehensively analyse the outcome, numerical data with respect to Root Mean Square Error (RMSE) implemented to the outcome in Fig. 6 are filed in Table III. The data clearly indicate the FOPID obtained worst performance which is approaching unstable condition. Even the performance indices of ISV, ITSV, IAV, and ITAV tabulated in Table IV clearly indicate the achievement of the FOPID controller.

TABLE II. PARAMETERS GENERATED BY PSO ALGORITHM

Controller	**Parameter**				
	K_p	**K_i**	**K_d**	**λ**	**δ**
PID	10.0910	0.0013	-4.6985	1	1
FO-PID	34.8991	0.7052	8.5401	2.0296	8.1205
SMC	-	-	-	87.6240	395.7009
SMC-PID	1118.2151	0.0000073	4.0390734	10	15

TABLE III. ROOT MEAN SQUARE ERROR WITH RESPECT TO THE CONTROLLER EFFORT

Controllers	**Root Mean Square Error (Voltage)**
PID	3.1558
FO-PID	6.9289×10^{12}
SMC	0.4893
SMC-PID	5

TABLE IV. PERFORMANCE INDEX WITH RESPECT TO THE CONTROLLER EFFORT

Analysis **Controller**	**IAV**	**ITAV**	**ISV**	**ITSV**
PID	0.22310	0.04781	4.53932	0.46953
FOPID	6.88390×10^{10}	6.88390×10^{10}	2.13366×10^{25}	2.13366×10^{24}
SMC	0.18473	0.02460	0.48685	0.06117
SMC-PID	0.20662	0.02311	1.65946	0.18140

TABLE V. PERFORMANCE INDEX WITH RESPECT TO THE TRAJECTORY TRACKING EFFICIENCY

Analysis **Controller**	**IAE**	**ITAE**	**ISE**	**ITSE**
PID	0.00394	0.00075	8.23513×10^{-05}	1.29177×10^{-05}
FOPID	0.00494	0.00126	7.87748×10^{-05}	1.40427×10^{-05}
SMC	0.00122	0.00015	2.58884×10^{-05}	3.04924×10^{-06}
SMC-PID	0.00057	0.00006	1.33854×10^{-05}	1.45076×10^{-06}

To investigate the tracking capability of the designed controllers, the performance indices with reference to the error generated by each controller is listed in Table V. It is roughly seen that the SMC-PID controller is able to providing greatest tracking efficiency compared to others even the conventional SMC controller. However, with reference to the controller effort as tabulated in Table IV, conventional SMC outperform the SMC-PID controller. Logically thinking, due to the numbers of gains, and the performance demonstrated by SMC-PID controller, highest precision required highest effort. Simply to said that, the energy usage and the precision can be achieved only either one. Therefore, further enhancement will be focused on this area, which is to designing the best energy usage and precise controller.

IV. CONCLUSIONS

In this article, the energy usage that interconnected with the controller effort, and the positioning tracking efficiency that interconnected with precision have been assessed. The modelling of the EHA system also addressed in the study. Then, the discussion in terms of the controller design and the tuning algorithm are carried out. It is notorious the existence of the nonlinear and uncertain characteristics in the EHA system that concurrently increase the complexities in the controller design. In the results, compared to the PID and the FO-PID controllers, the SMC-PID is able to achieving smallest error, while the smallest effort is required by the conventional SMC. It is however, in a practical point of view, only one of the criteria between precision and the controller effort can be achieved which has been proven in the results. Further examination in the experiment environment is necessary to support the aforementioned statement.

ACKNOWLEDGMENT

The support of Universiti Teknikal Malaysia Melaka (UTeM) is greatly acknowledged. The research was funded by UTeM Zamalah Scheme.

REFERENCES

[1] C. C. Soon, R. Ghazali, C. S. Horng, C. M. Shern, Y. Sam, and A. A. Yusof, “Controllers Capabilities with Computational Tuning Algorithm in Nonlinear Electro-Hydraulic Actuator System,” *J. Adv. Res. Fluid Mech. Therm. Sci.*, vol. 52, no. 2, pp. 148–160, 2018.

[2] S. M. Rozali, M. F. Rahmat, A. R. Husain, and M. N. Kamarudin, “Robust controller design for position tracking of nonlinear system using back stepping-GSA approach,” *ARPN J. Eng. Appl. Sci.*, vol. 11, no. 6, pp. 3783–3788, 2016.

[3] Q. Guo, J. Yin, T. Yu, and D. Jiang, “Saturated Adaptive Control of an Electrohydraulic Actuator with Parametric Uncertainty and Load Disturbance,” *IEEE Trans. Ind. Electron.*, vol. 64, no. 10, pp. 7930–7941, 2017.

[4] C. M. Shern, R. Ghazali, C. S. Horng, H. I. Jaafar, C. C. Soon, and Y. M. Sam, “Performance analysis of position tracking control with PID controller using an improved optimization technique,” *Int. J. Mech. Eng. Robot. Res.*, vol. 8, no. 3, pp. 401–405, 2019.

[5] M. M. A. Alqadasi, S. M. Othman, M. F. Rahmat, and F. Abdullah, “Optimization of PID for industrial electro-hydraulic actuator using PSOGSA,” *Telkomnika*, vol. 17, no. 5, pp. 2625–2635, 2019.

[6] M. P. Aghababa, “Optimal design of fractional-order PID controller for five bar linkage robot using a new particle swarm optimization algorithm,” *Soft Comput.*, vol. 20, no. 10, pp. 4055–4067, 2016.

[7] C.-A. Bojan-Dragos, R.-E. Precup, M. L. Tomescu, S. Preitl, O.-M. Tanasoiu, and S. Hergane, “Proportional-Integral-Derivative Gain-Scheduling Control of a Magnetic Levitation System,” *Int. J. Comput. Commun. Control*, vol. 12, no. 5, pp. 599–611, 2017.

[8] F. M. Zaihidee, S. Mekhilef, and M. Mubin, “Robust speed control of pmsm using sliding mode control (smc)-a review,” *Energies*, vol. 12, no. 9, pp. 1–27, 2019.

[9] Y. Wang, Y. Xia, H. Shen, and P. Zhou, “SMC design for robust stabilization of nonlinear markovian jump singular systems,” *IEEE Trans. Automat. Contr.*, vol. 63, no. 1, pp. 219–224, 2018.

[10] A. Mohammadi, H. Asadi, S. Mohamed, K. Nelson, and S. Nahavandi, “Multiobjective and Interactive Genetic Algorithms for Weight Tuning of a Model Predictive Control-Based Motion Cueing Algorithm,” *IEEE Trans. Cybern.*, vol. 49, no. 9, pp. 3471–3481, 2019.

[11] L. Ding and G. Gao, “Adaptive robust SMC of hybrid robot for automobile electro-coating conveying,” *J. Eng.*, vol. 2019, no. 15, pp. 587–592, 2019.

[12] N. Kapoor and J. Ohri, “Improved PSO tuned Classical Controllers (PID and SMC) for Robotic Manipulator,” *Int. J. Mod. Educ. Comput. Sci.*, vol. 7, no. 1, pp. 47–54, 2015.

[13] S. M. Othman, M. Rahmat, S. Rozali, and Z. Has, “Optimization of Modified Sliding Mode Controller for an Electro-hydraulic Actuator system with Mismatched Disturbance,” *Int. J. Electr. Comput. Eng.*, vol. 8, no. 4, 2018.

[14] S. M. Rozali *et al.*, “Robust Control Design of Nonlinear System via Backstepping-PSO with Sliding Mode Techniques,” in *Asian Simulation Conference*, 2017, pp. 27–37.

[15] M. Mahmoodabadi, M. Taherkhorsandi, M. Talebipour, and K. Castillo-Villar, “Adaptive Robust PID Control Subject to Supervisory Decoupled Sliding Mode Control Based Upon Genetic Algorithm Optimization,” *Trans. Inst. Meas. Control*, vol. 37, no. 4, pp. 505–514, 2015.

[16] W. Zhao, X. Zhou, C. Wang, and Z. Luan, “Energy analysis and optimization design of vehicle electro-hydraulic compound steering system,” *Appl. Energy*, vol. 255, p. 113713, 2019.

[17] I. Davliakos, I. Roditis, K. Lika, C. M. Breki, and E. Papadopoulos, “Design, development, and control of a tough electrohydraulic hexapod robot for subsea operations,” *Adv. Robot.*, vol. 32, no. 9, pp. 477–499, 2018.

[18] J. Shi, L. Quan, X. Zhang, and X. Xiong, “Electro-hydraulic velocity and position control based on independent metering valve control in mobile construction equipment,” *Autom. Constr.*, vol. 94, pp. 73–84, 2018.

[19] J. Zhao, G. Shen, C. Yang, W. Zhu, and J. Yao, “A robust force feed-forward observer for an electro-hydraulic control loading system in flight simulators,” *ISA Trans.*, vol. 89, pp. 198–217, 2019.

[20] R. H. Wong and W. H. Wong, “Comparisons of position control of valve-controlled and pump-controlled folding machines,” *J. Mar. Sci. Technol.*, vol. 26, no. 1, pp. 64–72, 2018.

[21] Y. Li, K. H. Ang, and G. C. Y. Chong, “PID control system analysis and design,” *IEEE Control Syst.*, vol. 26, no. 1, pp. 32–41, 2006.

[22] I. Podlubny, “Fractional-Order Systems and PIλDμ-controllers,” *IEEE Trans. Automat. Contr.*, vol. 44, no. 1, pp. 208–214, 1999.

[23] M. Zamani, M. Karimi-Ghartemani, N. Sadati, and M. Parniani, “Design of a Fractional Order PID Controller for an AVR using Particle Swarm Optimization,” *Control Eng. Pract.*, vol. 17, no. 12, pp. 1380–1387, 2009.

[24] I. Podlubny, “Fractional-Order Systems and Fractional-Order Controllers,” *Inst. Exp. Physics, Slovak Acad. Sci. Kosice*, vol. 12, no. 3, pp. 1–18, 1994.

[25] M. Dulau, A. Gligor, and T.-M. Dulau, “Fractional Order Controllers Versus Integer Order Controllers,” *Procedia Eng.*, vol. 181, pp. 538–545, 2017.

[26] G. Krishnan, Karthik and Karpagam, “Comparison of PID Controller Tuning Techniques for a FOPDT System,” *Int. J. Curr. Eng. Technol.*, vol. 4, no. 4, pp. 2667–2670, 2014.

[27] J. G. Ziegler and N. B. Nichols, “Optimum Sttings for Automatic Controllers,” *Transacction of the A.S.M.E*, vol. 64, no. 11, pp. 759–768, 1942.

[28] I. Eker, “Second-order Sliding Mode Control with Experimental Application,” *ISA Trans.*, vol. 49, no. 3, pp. 394–405, 2010.

[29] I. Eker, “Sliding Mode Control with PID Sliding Surface and Experimental Application to an Electromechanical Plant,” *ISA Trans.*, vol. 45, no. 1, pp. 109–118, 2006.

[30] I. Eker and S. A. Akinal, “Sliding mode control with integral augmented sliding surface: design and experimental application to an electromechanical system,” *Electr. Eng.*, vol. 90, no. 3, pp. 189–197, 2008.

Automation of Traditional Exam Invigilation using CCTV and Bio-Metric

MD Jiabul Hoque[1], Md. Razu Ahmed[2]
Dept. of Computer & Communication Engineering
International Islamic University Chittagong
Chittagong, Bangladesh

Md. Jashim Uddin[3], Muhammad Mostafa Amir Faisal[4]
Dept. of Electronic & Telecommunication Engineering
International Islamic University Chittagong
Chittagong, Bangladesh

***Abstract*—In education, whilst e-learning has been playing a significant role over the last few years due to its flexibility and remote-based education system, majority of courses are still relying upon traditional approaches of learning due to the lack of integrity and security of online based examinations and assessments in e-learning. As such, traditional approach of examination system is considered superior method than e-examination but it has few limitations in its tag such as excessive number of physical resources (invigilators) is required and high occurrences of malpractices by the students during exam. The objective of this paper is to develop a framework for traditional pen and paper based examination system where number of invigilators will substantially be reduced and malpractices by the students during exam will be abolished. In order to implement the proposed examination system, educational institutions are required to preserve a database using Parallax Data Acquisition tool (PLX-DAQ) that incorporates bio-metric information of all students. Before entering in the examination hall, examinees go through authentication process via bio-metric reader that is attached in front of each exam hall. During the examination, examinees are monitored and controlled by an invigilator from distance through the use of 360-degree Closed-Circuit Television (CCTV) cameras as well as ultra-high sensitive microphones and speakers. Here, CCTV cameras are used to monitor examinees physical malpractices and microphones are used to control examinees vocal malpractices. Only one invigilator is required for n-number of exam halls in this process. The communication between students and invigilator can be done with microphones and speakers attached in both exam halls and invigilator room. This model will wipe out malpractices during examination. It will be a cost effective, simple and secure solution of complex traditional exam invigilation process.**

Keywords—Automation in invigilation; bio-metric authentication; CCTV monitoring; e-assessment; Parallax Data Acquisition Tool (PLX-DAQ); traditional invigilation

I. Introduction

There are two types of examination systems that are followed in Bangladesh such as online examination system and offline or paper based traditional examination system. Online examination system is encountering numerous issues related to integrity, security and ethics. As such, traditional examination system (room based) is still so popular and holding number one choice for assessment [1]. Online examination system is out of the scope of this paper.

Traditional room based examination system has been increasingly popular in any form of educational institutions such as Universities, colleges and Schools, etc. for their student's regular assessment from early nineteenth century. Besides, this type of examination system is the default assessment method for selection during recruitment of numerous organizations [2]. However, requirement of excessive number of resources and hence, the cost incurred, as well as alarming number of malpractices by examinees throughout the process of examination in recent days have made question marks on traditional room based assessment technique. Inaccurate authentication of examinees before sitting for the exam as well as massive malpractices during exam is mainly two areas of traditional exam invigilation system where examinees try to breach the system [1]. Currently, students require providing their ID card and/or admit card to enter the exam hall which is highly insecure in this technological world as anyone can forge those documents and enter the exam hall as a legitimate examinee. Moreover, it is impractical to ask few invigilators to monitor and control an exam hall filled with full of examinees successfully and hence, massive malpractices are exercised by the examinees. Furthermore, the cost associated with traditional exam invigilation system is very high as it requires many invigilators and other form of resources which makes the system unfeasible [3]. In order to abolish shocking malpractices by examinees as well as to reduce cost of existing complex traditional approach of exam invigilation system, bio-metric system of authentication need to be introduced and automated monitoring and controlling of exam hall is needed to be established.

Before making any question on existing exam invigilation system a number of literatures have been reviewed. An automated exam invigilation system using bio-metric and CCTV has been proposed in order to resolve shortcoming of traditional room based exam invigilation system. The proposed model has been implemented using fingerprint sensor module, microcontroller, DC motor, PLX-DAQ, CCTV, speakers and ultrasensitive microphones etc. Bio-metric module of the proposed system will ensure that only registered and legitimate examinees are allowed to enter in the exam hall as well as 360 degree CCTV and ultra-sensitive microphone and speaker system will guarantee that there will be zero malpractices in the exam hall and only one invigilator will ensure it from distance [3]. The proposed model will nullify malpractices currently prevalent due to flaws in exam invigilation system and will reduce substantial amount of cost incurred by the complex traditional exam invigilation system. It has been identified that proposed system will be much cheaper and highly efficient alternative of current suggested system proposed by different researchers.

The rest of the paper is prepared after introduction as follows: Section II reviews the related works and finds out the potential gaps about traditional exam invigilation systems. In Section III the proposed exam invigilation model is developed and in Section IV the proposed model is implemented. And finally, Section V focuses on some discussions, conclusions and future improvements.

II. Related Works

Even though online examination system has been emerging, traditional room based examination system is still considered as principal examination method by any types of organization for any forms of written exam such as recruitment, academic examination, and public service examination etc. [4]. Online examination system has severe issues of authentication as well as hacking of exam materials before or during exam [5].

Currently, it is observed that student ID card and/or registration form (which is easily alterable in this digital world) is used for the purpose of authenticating legitimate examinee during examination. Besides, invigilators (two or more) are used to control and monitor the exam hall which is proven costly but is failed to eradicate malpractice by the examinees during exam. Some researchers proposed some innovative exam invigilation model, prototypes or system for traditional exam invigilation [6]. However, majority of them are either impractical to deploy or insecure, or are poorly accepted by users, and overall, failed to abolish malpractices during examination.

E-assessment relies on computers and controlled exam environment where examinees fail to acquire illicit help from environment surrounding the examinees [7]. However, some authors suggested that students can violate the rules and regulations imposed by e-assessment authority [8]. Numerous authors have proposed incorporating a human invigilator during online assessment to hinder suspicious actions of examinees and to promote academic integrity [9]. Conversely, countless difficulties in involving human invigilators during e-assessment have been notified by some recent studies. For instance, invigilator allows examinees for malpractices during e-assessment and in return he/she receives some form of benefits from examinees [10]. Besides, invigilator could face difficulties in identifying lookalikes examinees. Those are the few of many reasons unsecure e-assessment has not been deployed by many institutions around the globe [11].

In case of traditional exam based studies, an embedded exam invigilation system was proposed where an improved version of regular traditional exam invigilation was presented. Each student is provided with an electronic ID card as well as an electronic card reader which is serially interfaced with invigilator’s computer to authenticate students before entering the exam hall [12]. The system had severe flaws in authentication model as anyone can sit for the exam as long as he holds the electronic ID card. Moreover, the research did not mention about how the exam hall can be monitored and controlled apart from studying authentication part. Similar trend is observed with minor improvement in the following two studies: exam hall invigilation using GSM-GPRS and RFID based exam hall proctoring. In the former one use the location of exam hall for checking authentication as such the whole system would not work if there is a change in the location of exam hall [13]. Later one has some issues as if RFID card stolen or lost then RFID detectors needs to be set up again and this system cannot eliminate impersonation let alone eradicate malpractices during exam [14]. Some researchers suggested an exam hall invigilation system that authenticates examinees based on password [15]. As we all know password can be hacked, tempered or forgotten as such even legitimate examinees are prevented to enter the exam hall and create chaos. As a result the system cannot eliminate impersonation during exam as well.

The authors has proposed a new automated exam invigilation model using CCTV, fingerprint sensor module, ultra-sensitive microphones and speakers in order to overcome the gaps found in the current and previous literatures in the context of exam invigilation. The proposed model offers simple, secure, efficient and cost effective solution of complex traditional exam invigilation process.

III. Proposed Exam Invigilation Model

In order to simplify the complex and expensive traditional examination process, an automated exam invigilation model incorporating bio-metric attendance system, 360 degree CCTV, ultra-sensitive microphones and speakers has been proposed. The proposed model of exam invigilation system is depicted in Fig. 1, where following list of acronyms has been used:

AS – Answer Scripts
BR – Bio-metric Reader
QP – Question Papers
S&M – Speakers and Microphones
St – Stapler with pins
WR – Washroom

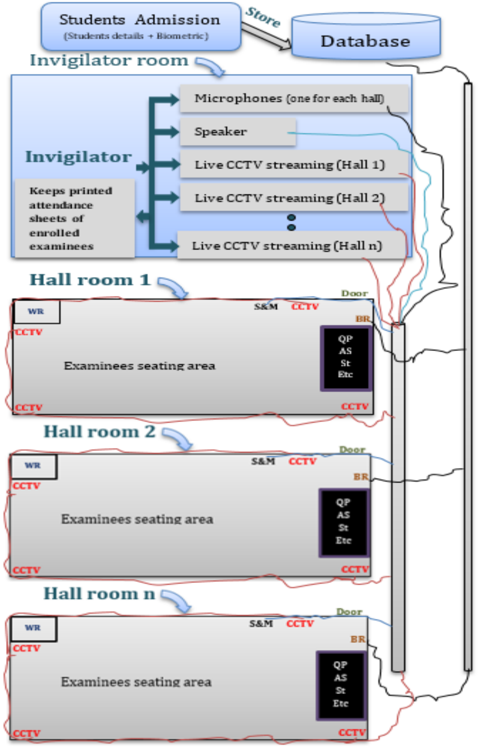

Students Admission
(Students details + Biometric)
Store
Database
Invigilator room
Invigilator
Keeps printed attendance sheets of enrolled examinees
Microphones (one for each hall)
Speaker
Live CCTV streaming (Hall 1)
Live CCTV streaming (Hall 2)
Live CCTV streaming (Hall n)
Hall room 1
Hall room 2
Hall room n
Door
WR
S&M
CCTV
BR
Examinees seating area
QP
AS
St
Etc

Fig 1. Proposed Exam Invigilation Model.

A. *Working Principle*

Following steps illustrate how the proposed exam invigilation model works:

Phase 1

Step 1: Students provide their details along with bio-metric (fingerprint [thumb]) during admission to the Institution (University/College/School/institute/coaching centre/etc).

Step 2: Students information together with bio-metric is stored in a database.

Phase 2

Step 3: Before the examination period, Institutions publish exam routine incorporating courses name, course code, exam date and time, building name, address, floor number, room number, seat number etc. based on students' enrolment of the particular subjects. Students aware about the exam details well before the exam through the notice board as well as respective web sites.

Phase 3

Step 4: Each exam hall is equipped with fingerprint bio-metric reader, CCTV, microphones, speakers, wall clock and washroom. Each exam hall is arranged in such a way that the exam hall will have only one door and the invigilator from the invigilation room can control the door remotely.

Step 5: Invigilator together with a helper set up exam halls at least 30 minutes before starting the exam.

- Arrange seat plan
- Provide answer sheets, question papers within sealed envelope, loose papers, stapler with pins, empty envelopes (to keep answer sheets after the exam) etc.

Step 6: Examinees approach to the exam centre 20 minutes before commencing the exam. Examinees make queue just outside the door of respective exam hall assigned to them. Bio-metric reader is set up next to the door of each exam hall. Each examinee provides bio-metric [for the purpose of authentication] using bio-metric reader in order to enter the exam hall.

- If the bio-metric information provided by the examinee matches with the information stored in the database then door of the exam hall will be automatically open and respective student is allowed to sit for the exam. In addition, a record of each examinee for a particular subject is sent to invigilator.

Step 8: Examinees collect answer script from a designated place within the exam hall as well as start filling up the necessary details in the answer script.

Step 9: Every examinee is instructed by the invigilator from invigilator room through microphone/speaker to collect question paper from the desk two minutes before commencing the exam.

Step 10: If everything goes well and predicted then invigilator provides signal to start writing to examinees of all exam halls.

Step 11: Invigilator room is equipped with a networked computer, a printer, CCTV live streaming display unit, microphone, and speaker. Invigilator keeps a copy of attendance of all examinees that enrolled for the examination on a particular day/time. Invigilator compares total enrolled examinees for a particular course with the actual attendance in the exam (got the list through bio-metric before commencing the exam).

Step 12: Doors of the examination halls are locked after 30 minutes of examination elapsed. If any examine comes after 30 minutes of examination then he/she will not be allowed to sit for that particular exam.

Step 13: During the examination, examinees (one at a time) are allowed to communicate with invigilator using microphone/speaker.

Step 14: Each examinee stop writing when they are instructed to do so by invigilator. After finishing the exam, examinees arrange the answer sheet (staples loose paper, if any) and submit it in a designated place within the exam hall. When all examines submit their paper then they are allowed to leave the exam hall one by one.

Step 15: No examinee is allowed to leave the exam hall within the duration of the exam. In case of unusual circumstances, he/she must discuss the matter with invigilator and get the permission to leave.

Step 16: Invigilator locks the door of exam hall after leaving all examinees. Invigilator along with a helper goes to each exam hall to collect the exam materials.

Step 17: After returning to the invigilation room, invigilator counts all the answer scripts and matches with the attendance sheet.

B. *Responsibilities of Invigilator*

Invigilator performs the following tasks before (Invigilator along with an assistant physically visits all exam halls) commencing the exam invigilation:

- Check any unusual stuff (such as writing exam material on wall or on table by students, keeping exam materials inside the washroom, etc.)
- Check to see CCTVs, bio-metric reader, Microphones and Speakers are in right place and in working condition.
- Arrange examinees seating area that identify students (attaching student name and ID on respective table)
- Placing exam materials (answer scripts, loose papers, question papers (sealed), and stapler with sufficient pin, etc.) on the desk.

Invigilator performs the following tasks during (Invigilator is in the invigilation room) examination:

- Watch CCTV Live streaming of each exam hall and communicate with examinees through microphone/speaker if required.

Invigilator performs the following tasks after elapsing particular examination:

- Perform step 16 and 17 mentioned above.
- Sealed answer scripts along with attendance sheet is sent to administration office that is responsible to send answer scripts to respective examiner for marking.

C. Block Diagram of Proposed Model

Fig. 2 depicts the block diagram of proposed exam invigilation model:

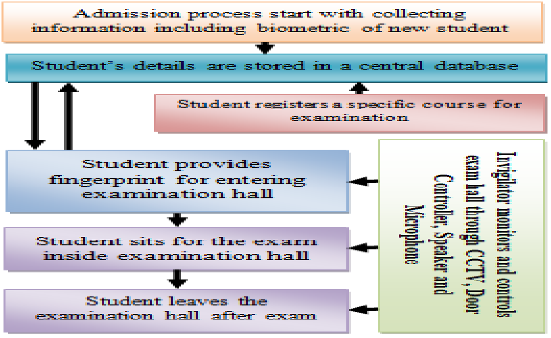

Admission process start with collecting information including biometric of new student
Student's details are stored in a central database
Student registers a specific course for examination
Student provides fingerprint for entering examination hall
Student sits for the exam inside examination hall
Student leaves the examination hall after exam
Invigilator monitors and controls exam hall through CCTV, Door Controller, Speaker and Microphone

Fig 2. Block Diagram of Proposed Exam Invigilation Model.

D. Flowchart of Proposed Model

Fig. 3 illustrates the flowchart of proposed exam invigilation model:

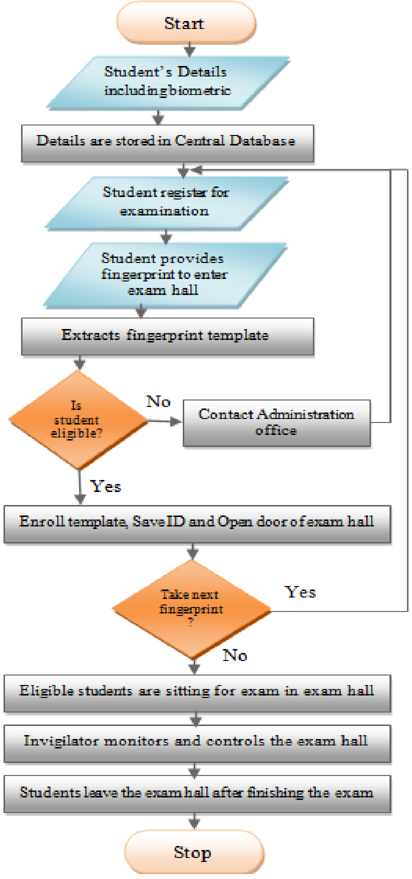

Start
Student's Details including biometric
Details are stored in Central Database
Student register for examination
Student provides fingerprint to enter exam hall
Extracts fingerprint template
Is student eligible?
No
Contact Administration office
Yes
Enroll template, Save ID and Open door of exam hall
Take next fingerprint ?
Yes
No
Eligible students are sitting for exam in exam hall
Invigilator monitors and controls the exam hall
Students leave the exam hall after finishing the exam
Stop

Fig 3. Flowchart of Proposed Exam Invigilation Model.

IV. System Implementation

In this section, hardware and software that are required for implementing proposed system have been described as well as how these materials incorporated to make a prototype of the proposed system has been illustrated.

A. Hardware Description

In this paper, the hardware comprise of fingerprint module, microcontroller, CCTV, microphone, Speaker, LED, motor driver, DC motor, door, diodes, capacitors and resistors. Among them following notable devices have been described:

a) Fingerprint Sensor Module: There are numerous fingerprint modules available in the market such as capacitive, ultrasonic, piezoresistive, piezoelectric, thermal, RF and optical etc. In this paper, an optical fingerprint sensor module (FPM10A) has been used because of its tremendous performance with low cost and low power consumption. The module reads the fingerprint pattern of the examinee and scans the images optically. Besides, it converts the scanned image to digital template and saves the output into the memory [16]. Fig. 4 shows optical fingerprint sensor module (FPM10A):

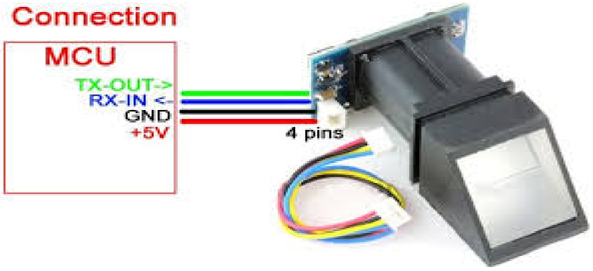


Fig 4. Optical Fingerprint Module.

b) Arduino Uno with ATmega328P Microcontroller: In this paper, Arduino Uno microcontroller board with ATmega328P microcontroller has been used:

- To connect with fingerprint module to get fingerprint template of examinee and compare it with stored fingerprint of corresponding examinee. If the fingerprint matches then the information of the examinee is stored in excel database.
- To give signals to the DC motor (attached with the door) via motor driver.

Fig. 5 illustrates Arduino Uno board with ATmega328P microcontroller [16]:

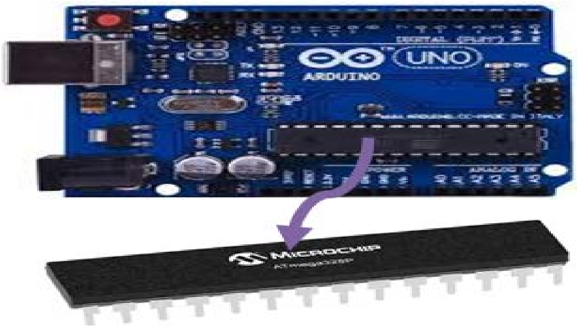


Fig 5. Arduino Uno with ATmega328P Microcontroller.

c) Driver for Motor: This is a driver for motor which can be used as a bridge between microcontroller and motor. It receives input signals from microcontroller and produces output signals for corresponding motor. Here, L293D IC has been used as a motor driver IC. It is a dual H bridge IC where one bridge is capable of driving a DC motor in bidirectional way [16]. Fig. 6 depicts the IC L293D of a motor driver:

Fig 6. Motor Driver IC.

d) Motor (DC): Fig. 7 demonstrates a small DC motor that runs with very low voltage starting from 0.5 volts. The motor is attached with the door in order to control it. The control signal comes from the microcontroller via motor driver [16].

Fig 7. Small DC Motor.

e) Door: In real scenario, a door will be used to enter the examinees in the exam hall after authenticating by the bio-metric system. However, in this implementation a small board has been used to replicate the real scenario.

*f) CCTV:*In this paper, a 360 degree CCTV camera has been used in order for invigilator to view entire exam hall at once [16]. Fig. 8 shows the CCTV camera:

Fig 8. CCTV Camera with 360 Degree View.

a) Microphone & Speaker: In order to ensure examinees in the exam hall are not taking each other a MS-MMM-1 ultra sensitive microphone has been used in this paper [17]. Fig. 9 depicts a MS-MMM-1 ultra sensitive microphone.

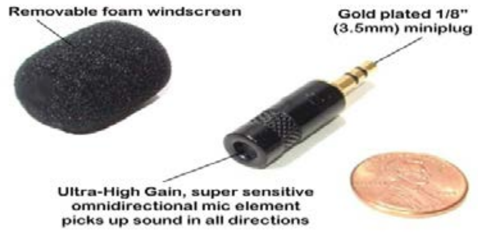

Removable foam windscreen
Gold plated 1/8" (3.5mm) miniplug
Ultra-High Gain, super sensitive omnidirectional mic element picks up sound in all directions

Fig 9. MS-MMM-1 Ultra-Sensitive Microphone.

b) LCD: A 16x4 LCD display with 12C interfaces has been used in this paper in order to display necessary information required for examinees to provide their fingerprint [17]. Fig. 10 shows a 16x4 LCD display unit.

Fig 10. A 16x4 LCD Display.

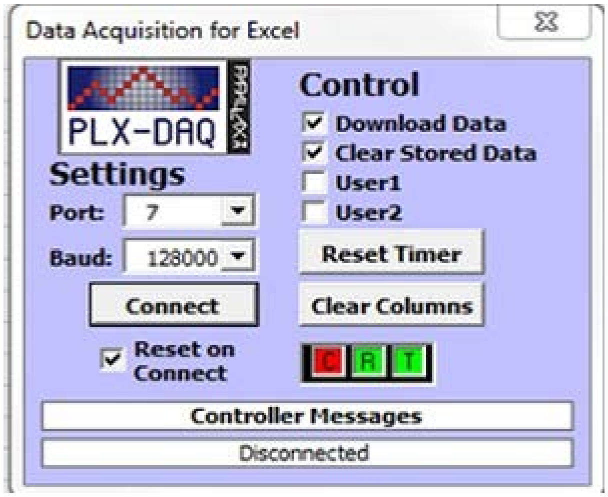
B. Software Description

The following software materials have been used to implement the proposed exam hall invigilation model:

a) PLX-DAQ & Microsoft Excel

PLX-DAQ stands for parallax microcontroller Data Acquisition which is used as add-in for Microsoft Excel. It is a software tool that is capable of recording up to 26 channels of data from sensors via microcontroller and plots those data in real time using Microsoft Excel. Moreover, rigorous analysis of collected data through sensors and monitors the equipments in real time easily possible by the use of PLX-DAQ [18].

As PLX-DAQ is an add-in of Microsoft Excel, PLX-DAQ run together with Microsoft Excel. PLX-DAQ needs to setup in order to acquire data from microcontroller. It supports any communication port from 1 to 15 and baud rate up to 128k. Here, com port 7 and baud rate 128000bps have been configured. Fig. 11 illustrates PLX-DAQ data acquisition for Excel:



Data Acquisition for Excel

Settings
Port: 7
Baud: 128000

Control
☑ Download Data
☑ Clear Stored Data
□ User1
□ User2

☑ Reset on Connect

Controller Messages
Disconnected

Fig 11. PLX-DAQ Setup for Data Acquisition.

C. Implementation

Fig. 12 illustrates how information is passing from one device to another in order to accomplish automated exam invigilation process.

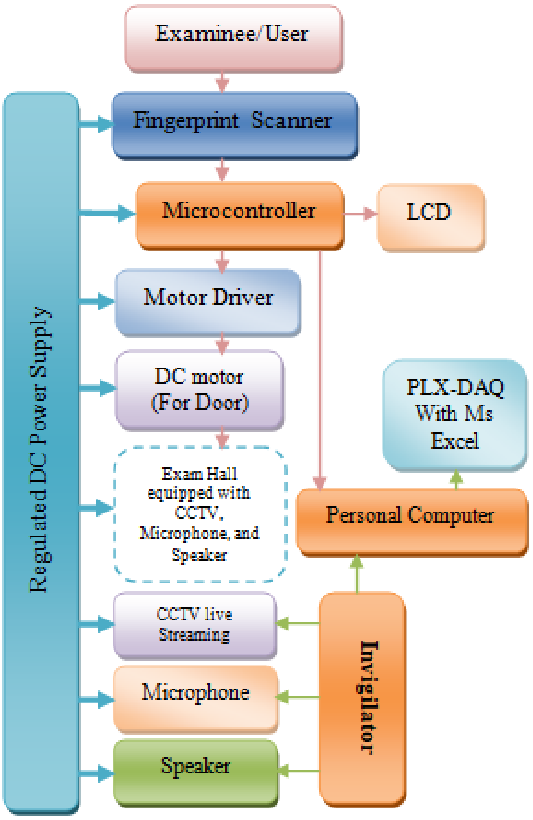


Fig 12. Information Flow Diagram of Exam Invigilation System.

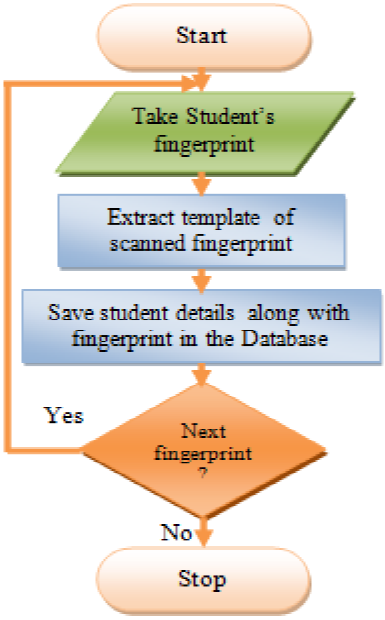

Start
Take Student's fingerprint
Extract template of scanned fingerprint
Save student details along with fingerprint in the Database
Next fingerprint ?
Yes
No
Stop

Fig 13. Flow Chart of Bio-Metric System during Admission.

It can be seen from the above figure that majority of the devices require uninterrupted and regulated DC power supply in order to run the examination process smoothly. Different devices require dissimilar level of voltage level as such regulated DC power supply provides power according to the device's power requirements [18].

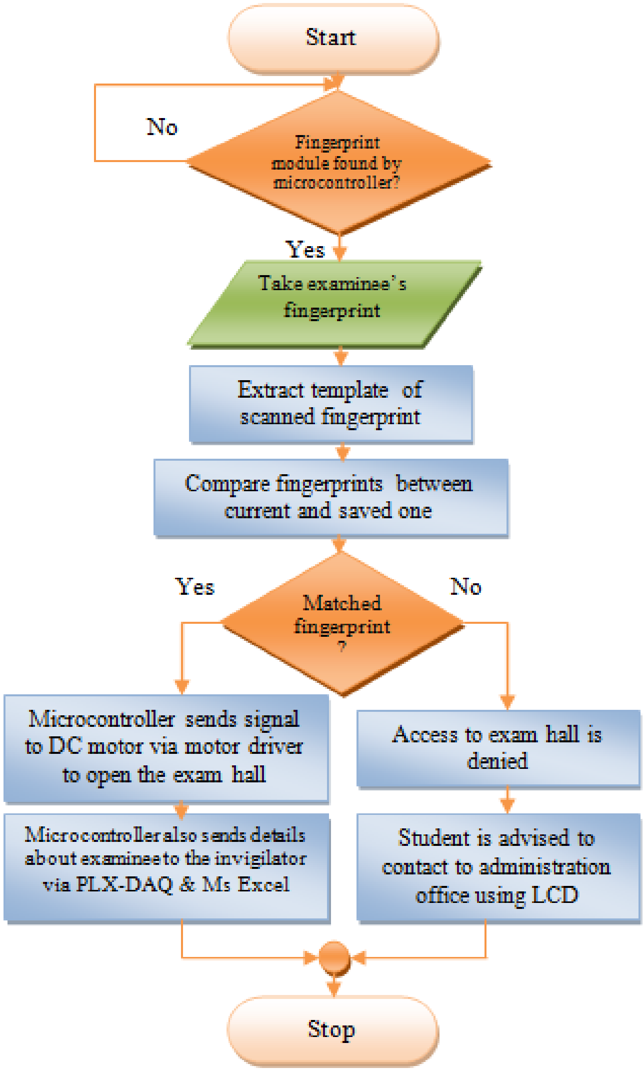

Start
No
Fingerprint module found by microcontroller?
Yes
Take examinee's fingerprint
Extract template of scanned fingerprint
Compare fingerprints between current and saved one
Yes
Matched fingerprint ?
No
Microcontroller sends signal to DC motor via motor driver to open the exam hall
Access to exam hall is denied
Microcontroller also sends details about examinee to the invigilator via PLX-DAQ & Ms Excel
Student is advised to contact to administration office using LCD
Stop

Fig 14. Flow Chart of Bio-Metric System during Exam Hall Entrance.

The exam invigilation system has fingerprint sensor scanner connected with microprocessor. The LCD display unit which is also connected with microprocessor assist user by displaying the activities of enrolment and authentication. Fig. 13 and Fig. 14 illustrate activities of bio-metric system during student's admission and during examination hall entrance for sitting examination respectively.

As long as the system is powered on, the fingerprint sensor offers user to place their finger to be scanned. Fig. 15 depicts the information which is offered to user initially.

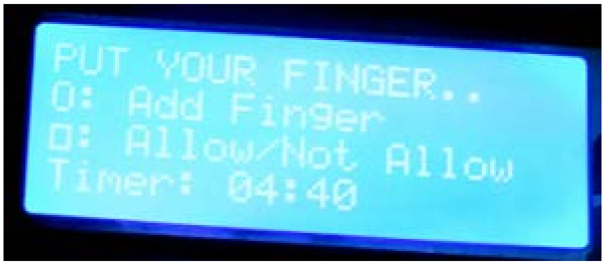

PUT YOUR FINGER..
0: Add Finger
0: Allow/Not Allow
Timer: 04:40

Fig 15. Welcome Message to the User.

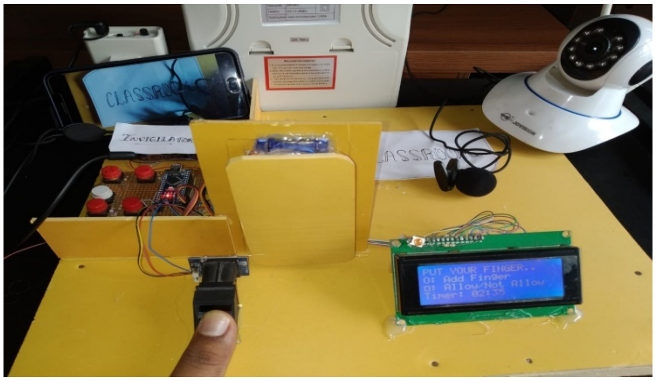


Fig 16. Examinee Puts their Fingerprint for Authentication.

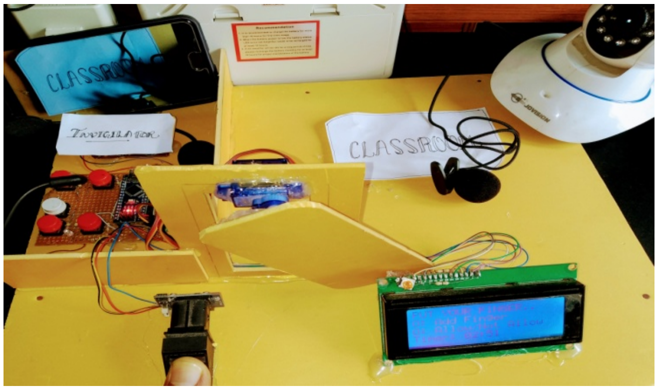


Fig 17. Opening Door Shows Successful Authentication.

User puts the finger on fingerprint sensor scanner in order to scan their fingerprint for authentication. Fingerprint module converts the scanned image to corresponding digital template. Now, microcontroller compares the current template of the scanned image with the stored one in the database scanned during exam registration. If there is a match between two then microcontroller send a signal to the DC motor via motor driver to open the door of exam hall. However, if there is no match between the fingerprint templates then the system inform user that "Access is denied" and contact to administration office as soon as possible to resolve the situation. Therefore, the system hinders unauthorized access of exam of from counterfeit users.

TABLE I. BIO-METRIC ATTENDANCE OF EXAMINEES

SL No	Bio-Metric Attendance of Examinees				
	Student ID	***Due Entry***	***Late Entry***	***Time***	***Date***
1. 1	ET-1504	YES	NO	9.35AM	7/26/2019
2.	ET-1504	YES	NO	9.37AM	7/26/2019
3.	ET-1504	YES	NO	9.48AM	7/26/2019
4.	ET-1504	YES	NO	9.55AM	7/26/2019
5.	ET-1504	NO	YES	10.03AM	7/26/2019
6.	ET-1504	NO	YES	10.05AM	7/26/2019
7.	ET-1504	NO	YES	10.10AM	7/26/2019

The following figures show the actual implementation of the proposed automated exam invigilation model where door of the exam hall is remained shut. Fig. 16 depicts authentication of examinee for exam hall entrance whereas Fig. 17 shows opening door after eligible examinee's successful authentication.

Microcontroller sends the information of entire examinees who are sitting in the examination hall after successful authentication to the Invigilator's personal computer through PLX-DAQ as Microsoft Excel document. An invigilator carefully invigilates using 360 degree CCTV as well as controls and monitors the exam hall using ultra-sensitive microphones and speakers.

Table I portrays bio-metric attendance of legitimate examinees before entering the examination hall. Invigilator collects a list of examinees who registered for examination from controller of examination (CoE) division well before commencing the exam. When examinees enter the examination hall after authenticating through fingerprint sensor module, a real time report as an Ms Excel document is sent to Invigilator's computer. Subsequently, Invigilator compares the real time list of examinees with the one already printed from CoE and takes the action about the examinees that are late or absent after the examination.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, automation of traditional exam invigilation model has been implemented using fingerprint sensor module, 360 degree camera, ultra-sensitive microphone and speaker. The model employs a fingerprint sensor module which permits only registered students who enrolled for the examination for a particular course. In this way, proxy (unauthorized student sit for the exam on behalf of registered examinee) in examination which is very familiar in Bangladesh as there are no systems in place to trace them, will be abolished. Hence, students will be motivated to study hard as they know there will be no one to sit for the exam instead of them. Besides, invigilator monitors the exam halls through CCTV live streaming and controls by using ultra-sensitive microphones and speakers that ensures zero malpractice occurrences during exam. Examinees know and fear that there will be someone watching them from distance all the time and they will get expelled for any wrong doing. Moreover, it saves time, money and resources as only one invigilator is required for total examination process. Furthermore, it simplifies the overall exam management process by automating and digitalizing the complex traditional exam invigilation process. The proposed exam invigilation model can be set up in any form of educational institutions (i.e. Universities, Colleges, Schools, and private or govt. institutions etc.) for any type of assessment (i.e. regular exam, admission test, and recruitment test, etc.) not only in Bangladesh but also anywhere in the globe.

This paper is a preliminary part of the full scale research on automation of exam invigilation system incorporating Internet of Things (IoT), Cloud computing and Image processing which will fully automates the traditional exam invigilation system that will simplify the current traditional exam invigilation system with negligible human intervention. Over the coming years, every work relating to this domain will be made public in future research papers.

REFERENCES

[1] M. A. Amin and J. Greenwood, “The examination system in Bangladesh and its impact: on curriculum, students, teachers and society,” Springer: Language Testing in Asia, vol.8, issue. 4, 2018.

[2] S. G. Anuradha and B. Kavya, “Automated face detection & recognition for detecting impersonation of candidate in examination system,” International Journal of Scientific & Engineering Research, vol. 7, pp. 159 – 160, 2016.

[3] S. Soma and S. A. Vidyashree, “An Automated Fraud Detection of Hall Ticket in an Offline Examination System using ANN Classifier,” International Journal of Computer Applications, vol. 126, pp. 7 – 12, 2015.

[4] M. P. Nerkar, “Online exam proctoring system,” International Journal of Advance Engineering and Research Development, vol. 3, pp. 110 – 118, 2017.

[5] C. J. Case and D. L. King, “E-cheating: incidence and treands among college students,” Issues in Information Systems, vol. 15, 2014.

[6] J. Roth, X. Liu, and A. Ross, “Investigating the discriminative power of keystroke sound,” IEEE Transection of Information Forensic Security, vol. 10, pp. 333 – 345, 2015.

[7] M. Hoque, M. Ahmed, and S. Hannan, “An Automated Greenhouse Monitoring and Controlling System using Sensors and Solar Power”, EJERS, vol. 5, no. 4, pp. 510-515, Apr. 2020.

[8] A. Wahid, Y. Sengoku, and M. Mambo, “Toward constructing a secure online examination system,” International Conference on Ubiquitous Information Management and Communication, ACM, pp. 95 – 102, 2015.

[9] M. Ahmed, M. Rahman, and M. Hoque, “Smart Home: An Empirical Analysis of Communication Technological Challenges”, EJERS, vol. 5, no. 5, pp. 571-575, May 2020.

[10] J. Chen and X. Liu, “Transfer learning with one class data,” Pattern Recognition Letters, vol. 37, pp. 32 – 40, 2014.

[11] Z. U. Ahmed, M. G. Mortuza, M. J. Uddin, M. H. Kabir, M. Mahiuddin and M. J. Hoque, "Internet of Things Based Patient Health Monitoring System Using Wearable Biomedical Device," International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh, 2018, pp. 1-5, doi: 10.1109/CIET.2018.8660846.

[12] P. Shenbagam and R. Kumar, “Exam hall invigilation using CCTV,” International Journal of Computer Science and Engineering, special issue, pp. 64 – 68, March 2017.

[13] Hoque, M., Kabir, S. and Hossain, K. (2018); Electricity Crisis of Bangladesh and A New Low Cost Electricity Production System to Overcome this Crisis; International Journal of Scientific and Research Publications (IJSRP) 8(7) (ISSN: 2250-3153), DOI: http://dx.doi.org/10.29322/IJSRP.8.7.2018.p7933.

[14] C. Saraswat and C. K. Amit, “An efficient automatic attendance system using fingerprint,” International Journal on Computer Science and Engineering, vol. 2, pp. 264 – 269, 2017.

[15] O.O Shoewu, M. Olaniyi, and A. Lawson, “Embedded computer based attendance management system,” Journal of IEEE Nigeria Computer Section, vol. 4, pp. 27 – 36, 2018.

[16] M. A. Mazidi, J. G. Mazidi and R. D. McKinlay, The 8051 Microcontroller: A systemetic approach, Pearson, 2012.

[17] S. Pankanti, and S. Prabhakar, “On the Individuality of Fingerprints,” IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 24, 2012.

[18] M. Habaebi and Q. Ashraf, "Autonomic schemes for threat mitigation in Internet of Things," Elsevier Journal of Network and Computer Applications (EJNCA), vol. 49, pp. 112 – 127, 2015.

Usability Evaluation of Open Source Learning Management Systems

Seren Başaran0000-0001-9983-1442 1
Near East University, Department of Computer Information Systems
Lefkoşa 98010 via: Mersin 10 Turkey, Cyprus

Rafia Khalleefah Hamad Mohammed[2]
Bright Star University, Department of Computer Engineering
El Brega City, Libya

***Abstract*—Advancements in Information and Communications Technology has enabled learning to be conducted online frequently through Learning Management Systems (LMS). The use of Learning Management Systems (LMS) as tools for learning in the present Internet age is seen as an important solution to remedy major problems particularly faced by higher education instructors, students and universities. However, any quality and usability related information regarding such widely used learning management systems are rarely encountered in the literature. The main objective of this study is to evaluate the system quality of the top five widely used open source learning management systems through the external characteristics of ISO/IEC 9126 quality standards evaluation model for Moodle, ATutor, Eliademy, Forma LMS and Dokeos with two experts. ISO/IEC 9126 quality model is adequate for evaluating important system quality metrics. Results highlighted in detail a set of usability and quality issues that are associated to external characteristics for each open LMS which require further attention of developers, educators and researchers to improve the quality of learning.**

***Keywords*—E-learning; ISO/IEC 9126; learning management systems; quality model; usability evaluation**

I. Introduction

Advancements in Information and Communications Technology (ICT) has enabled learning to be conducted online usually through Learning Management Systems (LMS). The integration of ICT in learning and its processes has resulted in increased improvements in the quality of learning generally by using several learning techniques all implemented when developing Learning Management Systems [1]. Learning Management Systems refer to different software packages designed to assist in the delivery and management of learning resources, materials and contents to students, usually via online/web platforms [2]. Basically, Learning Management Systems provide educators to create, deliver, supervise and monitor the participation of students as well as assess their performances [3]. They provide a platform for learning and obtaining knowledge at any time regardless of the geographic location of the users [4]. There are many Learning Management Systems that provide education instructors with variety of different options to select from [5]. Other Learning Management Systems provide features for assessing the learning progress of students, student registration, tracking and delivering of educational resources, materials and contents [2]. The use of Learning Management Systems as tools for learning in the present Internet age is seen as an important solution to remedy some of the problems faced by instructors, students and educational institutions in general [1]. Presently, different Learning Management Systems are implemented using different Internet technologies. The most common Internet technologies used are; (i) Open source, (ii) Cloud computing and (iii) Mobile based.

However, the most widely used technology when it comes to Learning Management Systems is the Open Source [6]. This is because open source software are tools whose source codes are available, can be modified to suit the requirements of the user and can easily be made available to the general public at [7]. This enables every institution that wishes to use any open source Learning Management System to simply download the source codes and modify them, which will enable the LMS to function according to the institute's requirements or preferences. With focus on open source Learning Management Systems, this study aims to conduct an evaluation study to compare the system quality of the top five most widely used open source Learning Management Systems available using the ISO/IEC 9126 quality evaluation model. The top five open source Learning Management Systems subject to this study are: (i) Moodle, (ii) ATutor (iii) Eliademy, (iv) Forma LMS, and (v) Dokeos [8]. The details of the aforementioned Learning Management Systems are introduced below:

Moodle: is a learning management system which is open source and free online learning platform for K12, higher education and for workplace that enables collaboration and engagement

ATutor: is another open source based LMS which has variability and extended functionality in module features. The content could be easily managed, packed and modified in standard web environment.

Eliademy: is web based learning platform where content could easily and flexibly be created, communicated and maintained. It is free as compared to its commercial competitive rivals such as Moodle and Blackboard.

Forma LMS: is open source LMS used for sharing online courses instantly.

DOKEOS: is an e-learning management system for managing content rich, flexible and effective learning and teaching experiences. DOKEOS works not only on desktop computers but also on mobile environments where no installation is required with always up to date.

The main aim of this study is to conduct an evaluation study to compare the usability of the top five open source Learning Management Systems available using the ISO/IEC 9126 quality evaluation model. To achieve this aim, a detailed study needs to carried out, which will take into consideration all the system quality and usability characteristics for each of the selected open source Learning Management Systems and then compared them with the aim of determining which amongst them has the finest system quality characteristics.

There are many open source Learning Management Systems that are available for use presently in the world. They offer very similar functionalities and have similar features as well, which are often difficult to distinguish. Some of the features enable instructors to interact and provide educational materials and resources to students during their learning process. There is vast increase in the number of Learning Management Systems especially open source LMS because they are easy to modify and customize to suit different preferences. However, the users usually do not know which open source Learning Management System is best in terms of system quality amongst the top rated open source LMSs that are available. Therefore, this paper aims to solve that problem by carrying out an evaluation study to help determining which open source LMS has the best system quality among the Top 5 open source Learning Management Systems.

The outcomes of this study will enlighten particularly instructors to determine which open source Learning Management Systems has a higher system quality in terms of external characteristics. In information technology as a whole, the system quality determines the level of acceptance and the usability quality and user satisfaction of any IT system that is used. Therefore, in order to determine which open source Learning Management System is the best, this paper carries out an evaluation of the top 5 rated open source Learning Management Systems. In addition, to authors knowledge such assessment of open source LMS has not been identified in the literature so far.

The rest of the paper will investigate the system quality of the top five most widely used open source Learning Management Systems using the ISO/IEC 9126 external quality metrics performed by two experts.

II. Literature Review

A. Learning Management Systems

Learning Management Systems refer to a platform which supports faculty, administration, learning experiences, instructor and learner/student services over the Internet [9]. There are two aspects to which Learning Management Systems can effectively be reviewed; as a social entity and as a technical entity [10]. With Learning Management Systems, quality of instructor functionality is a very important feature in LMS [11]. Due to how important instructors are to Learning Management Systems, instructors should have enough features and time to enable them interact and provide educational materials to students during the learning process [12].Another important feature of Learning Management Systems is its effectiveness to the learner/students [13], as the satisfaction of the learner/student while using the Learning Management System is an important factor when determining its success and overall usability [14], [15]. As the learning management systems provides as triadic bridge among instructors, students and administration, it is inevitable to gain information regarding the quality and usability of such tools for learning in order to improve quality.

B. ISO/IEC 9126 Quality Model

In order to evaluate the quality of a product/software, there are a set of quality characteristics that describes the system and they form the basis and the foundation for the evaluation [16]. The set of characteristics that form the basis for evaluation are called the quality model [17]. The standard of the ISO/IEC 9126 quality model is widely accepted in different countries around the world, with 1,129,446 certified companies worldwide and 485,554 companies certified in Europe alone as at 2013[11]. This study focused on only the external characteristics of the ISO/IEC 9126 quality model, which are functionality, reliability, usability and efficiency as proposed in[22] and was adopted from the model in Fig. 1 and was summarized in Table I.

1) Functionality: Learning Management Systems should have the capacity to carried out the functionalities that meets the needs of functional LMS under specified conditions [18]. As part of the functionality, Learning Management System should provide feature that are required to improve the learning experience of the learner. These include interoperability, accuracy, compliance and security [1]. Interoperability means the LMS's ability to function and interact with other applications. Compliance means developing the Learning Management Systems according to certain established specifications and guidelines. Privacy refers to the ability to protect information of the users, that is, both instructors and learners/students [18]. Functionality can easily be improved by implementing messaging services, that is, where both instructors and learners/students can send messages to each other [19].

2) Reliability: With Learning Management Systems, reliability refers to the consistency of the system and how it performs its intended functions without crashing/failure [20].

Learning Management Systems should be highly reliable, highly robust and should perform accurate well without being affected by high number of users, time of use, place of use or data access and connectivity issues [4]. One factor that determines reliability is the ability of learners/students to access the learning management systems at any given time, even under difficult situations of network connectivity [21].

[1] ISO Survey [online] Available at: https://www.iso.org/the-iso-survey.html [Accessed March 16, 2018]

[2] ISO/IEC 9126:2001. Software product evaluation—quality characteristics and guidelines for the user. Geneva: International Organization for Standardization. [online] Available at: https://www.iso.org/standard/22749.html [Accessed March 16, 2018]

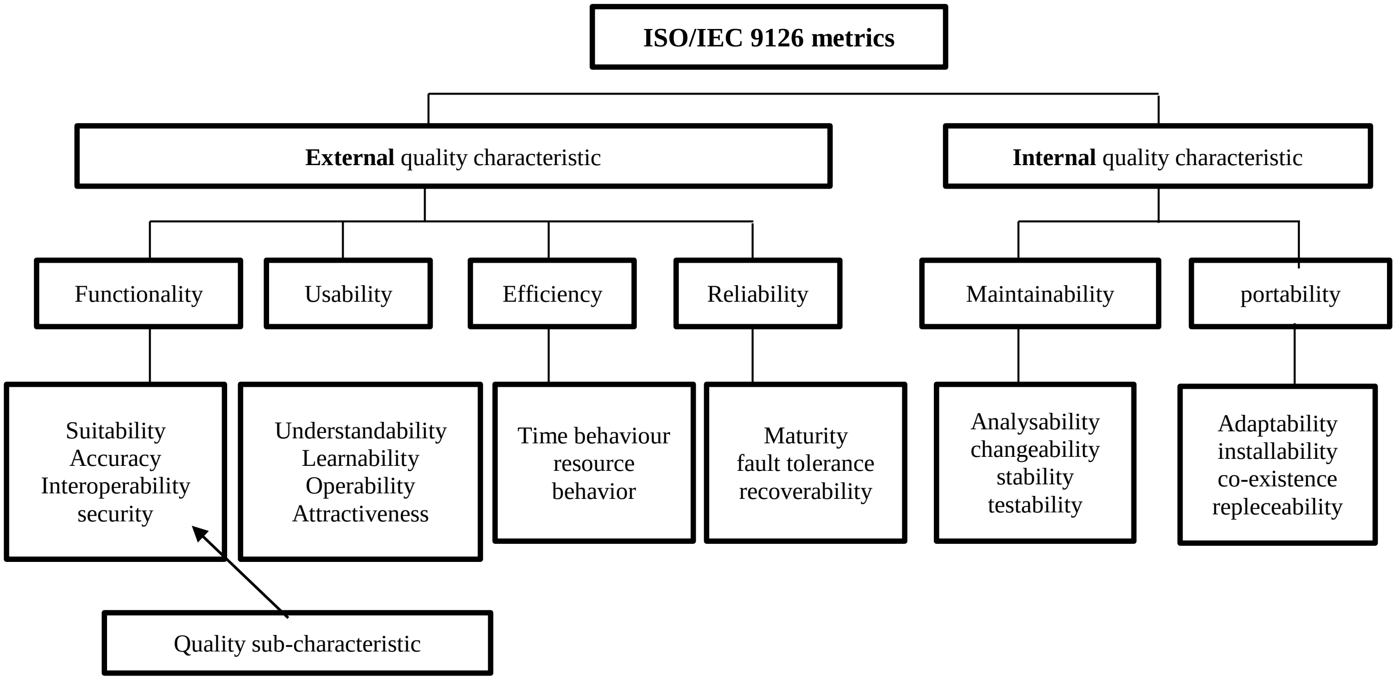

ISO/IEC 9126 metrics
External quality characteristic
Internal quality characteristic
Functionality
Usability
Efficiency
Reliability
Maintainability
portability
Suitability Accuracy Interoperability security
Understandability Learnability Operability Attractiveness
Time behaviour resource behavior
Maturity fault tolerance recoverability
Analysability changeability stability testability
Adaptability installability co-existence repleceability
Quality sub-characteristic

Fig 1. ISO/IEC 9126 Quality Model [35].

TABLE I. EXTERNAL CHARACTERISTICS AND SUB-CHARACTERISTICS [27]

ISO Characteristics	Criteria	Description
Functionality	Suitability	Can the software perform required tasks?
	Accuracy	Is the expected result achieved?
	Interoperability	Does the system interact with other systems?
	Security	Does the system stop unauthorized access?
Reliability	Maturity	Have the faults of the software been eliminated?
	Fault Tolerance	Does the system handle errors?
	Recoverability	Does the system still work after data loss?
Usability	Understandability	Is the system easy to use?
	Learnability	How easy can the user learn to use the system?
	Operability	Can the system be used with less effort?
	Attractiveness	Does the user interface look good?
Efficiency	Time Behavior	How quickly does the system respond?
	Resource Utilization	Does the system utilize resources?

3) Usability: Usability in terms of Learning Management Systems refers to the characteristics the define the quality, that is, it deals with how Learning Management Systems can be used by both the instructors and learners/students to achieve certain goals in the most efficient and effective manner possible in any given situation [22]. The guidelines that govern usability deals with mechanisms used to measure, monitor and improve the system processes of the Learning Management Systems [23]. It is very important for a Learning Management Systems to have a very clear and user-friendly user interface [24]. Therefore, in order to have a high usability level for Learning Management Systems, there is a need for the software developers to ensure that the LMS systems are adaptive and sensitive to different environments [25].

4) Efficiency: Efficiency refers to the performance level, the response time and how the over performance satisfies the needs of the user. In order to be efficient, the performance and response time must be fast so as to fully satisfy the needs of the users. Applications and systems should be able to grant the user fast access to vital information with good network speed available. This is because the main issue that affects the efficiency of Learning Management Systems are low bandwidth, lower security and interference among others [26].

C. System Quality

There are different guidelines for measuring the quality of a system and this outlines the characteristics of a software application, one of such guidelines is the ISO/IEC 9126 quality model[3] [3]. Due to the many factors involved in the process of software development, there is a need to focus strictly on the characteristics of Learning Management Systems. The LMS characteristics are selected by studying and selecting the most important characteristics needed for a Learning Management Systems [1]. The quality of Learning Management Systems is further influenced by the quality of content and how the instructors can successful manipulate educational materials and effectively deliver them to the learners/students [28]. Developers of Learning Management Systems usually focus on enhancing the characteristics that are considered to be the most important, thereby improving the overall system quality [29].

Here, system quality refers to the needed and desired characteristics of open source Learning Management Systems, which are:

- Functionality states to attaining the user’s anticipated necessities [30], [31].
- Reliability talks about acceptance and duration. Acceptance states whether any guidance available for the system in use. Duration refers to the duration of the system in the potential market. Fault tolerance deals with the support related issues regarding possible faults in the system.
- Usability refers to how easy it is to learn, accessibility, user interface and operability. Learnability addresses to the ease of control and understand the system without referring to the user manual. Operability is associated to the ease of operating of the system. Accessibility refers to ease of access to the system without requiring any other software or plug in.
- Efficiency refers to maximum performance, be easy to install, configure and operate within a short time. Efficiency is the most important feature when it comes to open source Learning Management Systems.

III. Methodology

Learning Management Systems have become an important part of higher education. Due to this reason, this study aims to look at the system quality and features of the top 5 highly rated open source Learning Management Systems using the ISO/IEC 9126 quality model and compare the results. Learning Management Systems provide certain usability features that enable education instructors to interact and provide educational materials and resources to students during the learning process. There is vast increase in the number of Learning Management Systems especially open source LMS because they are easy to modify and customize to suit different preferences. Due to this, the purpose of this study is to evaluate and compare the best 5 open source Learning Management Systems to find out which among them has a higher system quality and usability. Initial background research was conducted on 35 open source Learning Management Systems.

Then, the best 5 amongst them were selected and compared using the external characteristics of ISO/IEC 9126 quality evaluation model, according to the system quality characteristics as proposed by [19] and [7]. This is based on the overall system quality of the Learning Management Systems. The importance of system quality in every application and in Learning Management Systems in particular cannot be over emphasized because system quality automatically translates to usability, that is, the higher the system quality, the higher the usability vice versa. Due to the increasing number and usage of Learning Management Systems over the years, it has become necessary to carry out an evaluation studies on the top 5 open source LMSs based on the following system quality characteristics; Functionality, Reliability, Usability and Efficiency. The Learning Management Systems selected for this study are Moodle, ATutor, Eliademy, Forma lms and Dokeos.

Two experts were asked to conduct a research on both the websites and while using the selected top 5 best open source Learning Management Systems in order to get the relevant information on the latest versions of their systems. The background of the experts are; expertise on educational technology and Information Systems and in computer engineering. They are well experienced in the processes of designing and developing various types of management information systems. After the experts have gotten access and obtained required information, they evaluated the selected open source Learning Management Systems using the software quality characteristics of Learning Management Systems were identified using the model proposed by [27]. Table I shows the corresponding questions considered by the experts during evaluation process. The flowchart in Fig. 2 represents the evaluation procedure of the study.

[3] ISO/IEC 19796-1:2005 Information technology — Learning, education and training — Quality management, assurance and metrics —General approach. [online] Available at: https://www.iso.org/standard/33934.html [Accessed March 16, 2018]

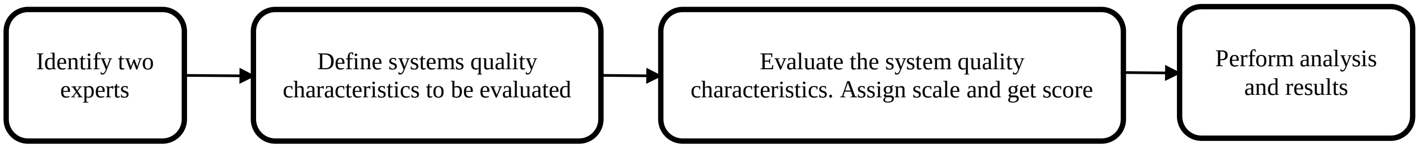

Identify two experts
Define systems quality characteristics to be evaluated
Evaluate the system quality characteristics. Assign scale and get score
Perform analysis and results

Fig 2. Evaluation Workflow.

IV. Results

Two experts were required to use the top five open source Learning Management Systems and allocate scores to them based on their how they function effectively. The Ideal Value, which is 1 represents the highest level of functionality for each usability sub-criteria as shown below in the tables. Therefore, in order to get the result of sub-criteria, the score is divided over the Ideal Value.

In order to arrive at the final result, the score of each sub-criteria as allocated by both experts is added up and their average is obtained, then the final score (i.e. the average from the score from both experts) is compared against the Ideal Value. If the score is equal to the Ideal Value, then the sub-criteria is set to be complete (i.e. fully functional) but if the score is not equal to the Ideal Value, then the sub-criteria is not complete (i.e. not fully functional). The evaluation was done on the external characteristics of each of the selected open source Learning Management Systems. After both the experts had carried out their evaluations separately, the values of their results for the characteristics of each open source Learning Management System will be compared and their average will be taken in order to determine the final score of each characteristic. Tables II to VI represent the usability characteristics with sub criteria evaluation for each open source LMS. Tables VII to XI indicate the overall scores for each LMS. Table XII depicted the overall result with combined assessments of the experts.

TABLE II. Usability characteristics for Moodle

Metric Name	Sub-criteria	Sub-criteria description	Expert number 1			Expert number 2		
			Score	**Ideal value**	**Formula/ Result**	**Score**	**Ideal value**	**Formula/ Result**
Functionality	Suitability	Can the software perform required tasks?	1	1	1/1	1	1	1/1
	Accuracy	Is the expected result achieved?	1	1	1/1	1	1	1/1
	Interoperability	Does the system interact with other systems?	0.8	1	0.8/1	1	1	1/1
	Security	Does the system stop unauthorized access?	1	1	1/1	1	1	1/1
Reliability	Maturity	Have the faults of the software been eliminated?	1	1	1/1	1	1	1/1
	Fault Tolerance	Does the system handle errors?	1	1	1/1	1	1	0.8/1
	Recovery	Does the system still work after data loss?	1	1	1/1	1	1	1/1
Usability	Understandability	Is the system easy to use?	1	1	1/1	0.8	1	0.8/1
	Learnability	How easy can the user learn to use the system?	1	1	1/1	1	1	1/1
	Operability	Can the system be used with less effort?	1	1	1/1	1	1	1/1
	Attractiveness	Does the user interface look good?	1	1	1/1	0.8	1	0.8/1
Efficiency	Time Behavior	How quickly does the system respond?	1	1	1/1	1	1	1/1
	Resource Utilization	Does the system utilize resources?	1	1	1/1	1	1	0.8/1

TABLE III. USABILITY CHARACTERISTICS FOR ATUTOR

Metric Name	**Sub-criteria**	**Sub-criteria description**	**Expert number 1**			**Expert number 2**		
			Score	**Ideal value**	**Formula/ Result**	**Score**	**Ideal value**	**Formula/ Result**
Functionality	Suitability	Can the software perform required tasks?	1	1	1/1	0.8	1	0.8/1
	Accuracy	Is the expected result achieved?	0.8	1	0.8/1	1	1	1/1
	Interoperability	Does the system interact with other systems?	0.6	1	0.6/1	0.8	1	0.8/1
	Security	Does the system stop unauthorized access?	1	1	1/1	1	1	1/1
Reliability	Maturity	Have the faults of the software been eliminated?	0.8	1	0.8/1	0.8	1	0.8/1
	Fault Tolerance	Does the system handle errors?	0.6	1	0.6/1	0.6	1	0.6/1
	Recovery	Does the system still work after data loss?	1	1	1/1	1	1	1/1
Usability	Understandability	Is the system easy to use?	1	1	1/1	1	1	1/1
	Learnability	How easy can the user learn to use the system?	1	1	1/1	1	1	1/1
	Operability	Can the system be used with less effort?	1	1	1/1	1	1	1/1
	Attractiveness	Does the user interface look good?	0.8	1	0.8/1	1	1	1/1
Efficiency	Time Behavior	How quickly does the system respond?	0.6	1	0.6/1	0.8	1	0.8/1
	Resource Utilization	Does the system utilize resources?	0.8	1	0.8/1	0.6	1	0.6/1

TABLE IV. USABILITY CHARACTERISTICS FOR ELIADEMY

Metric Name	**Sub-criteria**	**Sub-criteria description**	**Expert number 1**			**Expert number 2**		
			Score	**Ideal value**	**Formula/ Result**	**Score**	**Ideal value**	**Formula/ Result**
Functionality	Suitability	Can the software perform required tasks?	0.8	1	0.8/1	1	1	1/1
	Accuracy	Is the expected result achieved?	0.8	1	0.8/1	0.8	1	0.8/1
	Interoperability	Does the system interact with other systems?	0.4	1	0.4/1	0.4	1	0.4/1
	Security	Does the system stop unauthorized access?	1	1	1/1	0.8	1	0.8/1
Reliability	Maturity	Have the faults of the software been eliminated?	0.8	1	0.8/1	0.8	1	0.8/1
	Fault Tolerance	Does the system handle errors?	0.8	1	0.8/1	0.8	1	0.8/1
	Recovery	Does the system still work after data loss?	0.6	1	0.6/1	0.8	1	0.8/1
Usability	Understandability	Is the system easy to use?	0.8	1	0.8/1	1	1	1/1
	Learnability	How easy can the user learn to use the system?	0.6	1	0.6/1	0.8	1	0.8/1
	Operability	Can the system be used with less effort?	0.8	1	0.8/1	0.6	1	0.6/1
	Attractiveness	Does the user interface look good?	0.6	1	0.6/1	1	1	1/1
Efficiency	Time Behavior	How quickly does the system respond?	1	1	1/1	0.8	1	0.8/1
	Resource Utilization	Does the system utilize resources?	0.8	1	0.8/1	0.6	1	0.6/1

TABLE V. USABILITY CHARACTERISTICS FOR FORMA LMS

Metric Name	Sub-criteria	Sub-criteria description	Expert number 1			Expert number 2		
			Score	Ideal value	Formula/ Result	Score	Ideal value	Formula/ Result
Functionality	Suitability	Can the software perform required tasks?	0.8	1	0.8/1	1	1	1/1
	Accuracy	Is the expected result achieved?	1	1	1/1	1	1	1/1
	Interoperability	Does the system interact with other systems?	0.6	1	0.6/1	0.8	1	0.8/1
	Security	Does the system stop unauthorized access?	1	1	1/1	0.8	1	0.8/1
Reliability	Maturity	Have the faults of the software been eliminated?	1	1	1/1	0.8.	1	0.8/1
	Fault Tolerance	Does the system handle errors?	0.6	1	0.6/1	0.4	1	0.4/1
	Recovery	Does the system still work after data loss?	0.4	1	0.4/1	0.8	1	0.8/1
Usability	Understandability	Is the system easy to use?	0.8	1	0.8/1	0.8	1	0.8/1
	Learnability	How easy can the user learn to use the system?	0.8	1	0.8/1	1	1	1/1
	Operability	Can the system be used with less effort?	1	1	1/1	0.8	1	0.8/1
	Attractiveness	Does the user interface look good?	0.8	1	0.8/1	0.6	1	0.6/1
Efficiency	Time Behavior	How quickly does the system respond?	0.8	1	0.8/1	0.6	1	0.6/1
	Resource Utilization	Does the system utilize resources?	0.8	1	0.8/1	0.2	1	0.2/1

TABLE VI. USABILITY CHARACTERISTICS FOR DOKEOS

Metric Name	Sub-criteria	Sub-criteria description	Expert number 1			Expert number 2		
			Score	Ideal value	Formula/ Result	Score	Ideal value	Formula/ Result
Functionality	Suitability	Can the software perform required tasks?	0.8	1	0.8/1	0.8	1	0.8/1
	Accuracy	Is the expected result achieved?	1	1	1/1	1	1	1/1
	Interoperability	Does the system interact with other systems?	0.8	1	0.8/1	0.8	1	0.8/1
	Security	Does the system stop unauthorized access?	0.6	1	0.6/1	0.6	1	0.6/1
Reliability	Maturity	Have the faults of the software been eliminated?	0.8	1	0.8/1	0.8	1	0.8/1
	Fault Tolerance	Does the system handle errors?	0.6	1	0.6/1	0.6	1	0.6/1
	Recovery	Does the system still work after data loss?	0.8	1	0.8/1	0.4	1	0.4/1
Usability	Understandability	Is the system easy to use?	0.4	1	0.4/1	0.8	1	0.8/1
	Learnability	How easy can the user learn to use the system?	0.8	1	0.8/1	0.6	1	0.6/1
	Operability	Can the system be used with less effort?	0.8	1	0.8/1	0.6	1	0.6/1
	Attractiveness	Does the user interface look good?	0.6	1	0.6/1	0.8	1	0.8/1
Efficiency	Time Behavior	How quickly does the system respond?	0.6.	1	0.6/1	0.2	1	0.2/1
	Resource Utilization	Does the system utilize resources?	0.2	1	0.2/1	0.6	1	0.6/1

TABLE VII. SCORE OF CHARACTERISTICS FOR MOODLE

Characteristic	**expert number 1**			**expert number 2**		
	Score	**Ideal value**	**Description**	**Score**	**Ideal value**	**Description**
Functionality	3.8	4	Not complete	4	4	Complete
Reliability	3	3	Complete	3	3	Complete
Usability	4	4	Complete	3.6	4	Not complete
Efficiency	2	2	Complete	2	2	Complete

TABLE VIII. SCORE OF CHARACTERISTICS FOR ATUTOR

Characteristic	**expert number 1**			**expert number 2**		
	Score	**Ideal value**	**Description**	**Score**	**Ideal value**	**Description**
Functionality	3.4	4	Not complete	3.6	4	Not complete
Reliability	2.4	3	Not complete	2.6	3	Not complete
Usability	4	4	Complete	4	4	Complete
Efficiency	1.4	2	Not complete	1.4	2	Not complete

TABLE IX. SCORE OF CHARACTERISTICS FOR ELIADEMY

Characteristic	**expert number 1**			**expert number 2**		
	Score	**Ideal value**	**Description**	**Score**	**Ideal value**	**Description**
Functionality	3	4	Not complete	3	4	Not complete
Reliability	2.2	3	Not complete	2.4	3	Not complete
Usability	2.8	4	Not complete	3.4	4	Not complete
Efficiency	1.8	2	Not complete	1.4	2	Not complete

TABLE X. SCORE OF CHARACTERISTICS FOR FORMA LMS

Characteristic	**expert number 1**			**expert number 2**		
	Score	**Ideal value**	**Description**	**Score**	**Ideal value**	**Description**
Functionality	3.4	4	Not complete	3.6	4	Not complete
Reliability	2	3	Not complete	2	3	Not complete
Usability	3.4	4	Not complete	3.2	4	Not complete
Efficiency	1.6	2	Not complete	0.8	2	Not complete

TABLE XI. SCORE OF CHARACTERISTICS FOR DOKEOS

Characteristic	**expert number 1**			**expert number 2**		
	Score	**Ideal value**	**Description**	**Score**	**Ideal value**	**Description**
Functionality	3.2	4	Not complete	3.2	4	Not complete
Reliability	2.2	3	Not complete	1.8	3	Not complete
Usability	2.6	4	Not complete	2.4	4	Not complete
Efficiency	0.8	2	Not complete	0.8	2	Not complete

TABLE XII. EVALUATION RESULT

Evaluation result for Moodle				
Characteristics	**Score**	**Ideal Value**	**Description**	Moodle has very high system quality. As shown in the table, Moodle has 3 complete characteristics, which are Reliability, Usability and Efficiency. Apart from these three, it also has a very high functionality, which makes it a very usable and dependable open source Learning Management System to use
Functionality	3.9	4	Not complete	
Reliability	3	3	Complete	
Usability	4	4	Complete	
Efficiency	2	2	Complete	
Evaluation result for ATutor				
Characteristics	**Score**	**Ideal Value**	**Description**	ATutor rather has a very good system quality. It seems to have complete usability, which means it is fully usable. It is moderately efficient as well, indicating that it has an average performance. However, its functionality is high and reliability is average, which indicates that improvement is needed
Functionality	3.5	4	Not complete	
Reliability	2.5	3	Not complete	
Usability	4	4	Complete	
Efficiency	1.4	2	Not complete	
Evaluation result for Eliademy				
Characteristics	**Score**	**Ideal Value**	**Description**	Eliademy has an average reliability. The functionality and usability are also average, which suggests a need for improvement is mandatory However, it has below average usability, which indicates that there is a possibility that users find it hard to use the system partly due to interface deficiency and other usability factors.
Functionality	3	4	Not complete	
Reliability	2.3	3	Not complete	
Usability	3.2	4	Not complete	
Efficiency	1.6	2	Not complete	
Evaluation result for Forma LMS				
Characteristics	**Score**	**Ideal Value**	**Description**	Forma LMS has a very good functionality and usability. Also, the efficiency is average suggesting that there is a need to improve the overall performance of the Learning Management System
Functionality	3.5	4	Not complete	
Reliability	2	3	Not complete	
Usability	3.3	4	Not complete	
Efficiency	1.2	2	Not complete	
Evaluation result for DOKEOS				
Characteristics	**Score**	**Ideal Value**	**Description**	DOKEOS has a high functionality. The reliability and usability however, are on the average level, which suggests improvements are needed especially on certain perceived faults and the time it takes for the system to responds to certain requests. It has a very poor efficiency, which suggests that there are improvements needed in the performance and in response time.
Functionality	3.2	4	Not complete	
Reliability	2	3	Not complete	
Usability	2.5	4	Not complete	
Efficiency	0.8	2	Not complete	

V. DISCUSSION

Only the external characteristics of ISO/IEC 9126 are used in this evaluation study. This is because they are the only characteristics where evaluators have access to as the others; Maintainability and Portability are internal characteristics, which means only the developers and admin has access to [32]. Since the scores allocated to the characteristics of each open source Learning Management Systems was as a result of the information derived.

The Learning Management Systems mentioned above all possess different levels of system quality characteristics. From the results obtained, Moodle seems to have a higher system quality among the Learning Management Systems selected for this study. This is why it is available in 78 different languages and is being used in 216 countries compared to the others that are available in far lesser languages and countries (ATutor: 20 languages, 58 countries; Eliademy: 26 languages, 53 countries; Forma LMS: 15 languages, 17 countries; and DOKEOS: 11 languages, 18 countries).

The completeness of the reliability of Moodle, which means it can be used effectively in small and large-scale environments, which is difficult for the other Learning Management Systems to achieve. This also, is another reason why Moodle has a larger penetration and used in more countries. Also, Moodle has a complete efficiency as well, which show that its response time to user requests, speed and simplicity of installation are performing at maximum. With regards to efficiency, Forma LMS and DOKEOS both have relatively low efficiency due to their inability to be utilized in large environments, profiling and management competencies and installation deficiencies.

Moodle and ATutor both have complete usability, which indicates the level of user satisfaction while using the systems. It also shows the level of usable the customization features, adaptability and accessibility they both possess. Eliademy has a low usability due to low integration and interoperability. The functionality of Moodle, Forma LMS and DOKEOS shows high functionality features due to the availability of plug-ins, add-ons and core functionality features reside in the core of the systems. These features gave them a high functionality while using the open source LMSs, the internal characteristics could not be accessed and studied.

There exists fuzzy decision making techniques that uses at least one expert in the evaluation process in a similar fashion with this study [33].

VI. Conclusion

The aim of this study was to evaluate and help identify the best open source Learning Management Systems from the top 5 current available for use. The study gained access to previous works as reference points and used the external characteristics of ISO/IEC 9126 standard quality evaluation model as the guideline for the evaluation study.

Focusing on the system quality of the open source Learning Management Systems alone, the study realized that Moodle is the best open source Learning Management Systems that is presently in use, due to its availability and accessibility in many languages and countries as well as in the availability and plug-ins and add-ons, which greatly improves its functionalities among other features.

VII. Suggestions

This study evaluates most widely used open source learning systems using external quality characteristics of ISO/IEC 9126 quality standards by two experts. In the future, further research can be conducted using different approaches or models to see whether the same result can be achieved or not. Increasing the number of experts, using heuristic evaluation,- and applying fuzzy decision making processes using same ISO/IEC 9126 quality models were discussed in earlier studies [34], [35] that could also be used in the evaluation of open source learning management systems and risk assessments could also be beneficial for further improvements of such open source learning systems.

References

[1] M. Sarrab, M. Elbasir, and S. Alnaeli, “Towards a quality model of technical aspects for mobile learning services: An empirical investigation,” Computers in Human Behavior, vol. 55, pp. 100–112, Feb. 2016

[2] N. B. Awang and M. Y. B. Darus, “Evaluation of an Open Source Learning Management System: Claroline,” Procedia - Social and Behavioral Sciences, vol. 67, pp. 416–426, Dec. 2012.

[3] M. Cheng and A. H. K. Yuen, “Student continuance of learning management system use: A longitudinal exploration,” Computers & Education, vol. 120, pp. 241–253, May 2018.

[4] M. Sarrab, N. Alalwan, O. Alfarraj, and A. Alzahrani, “An empirical study on cloud computing requirements for better mobile learning services,” International Journal of Mobile Learning and Organisation, vol. 9, no. 1, p. 1, 2015

[5] S. Hussain, Zhaoshun Wang, and Chang-ai Sun, “A comparative study of open-source learning management systems,” 2011 IEEE International Workshop on Open-source Software for Scientific Computation, Oct. 2011.

[6] E. Pecheanu, D. Stefanescu, L. Dumitriu, and C. Segal, “Methods to evaluate open source learning platforms,” 2011 IEEE Global Engineering Education Conference (EDUCON), Apr. 2011.

[7] M. Sarrab and O. M. H. Rehman, “Empirical study of open source software selection for adoption, based on software quality characteristics,” Advances in Engineering Software, vol. 69, pp. 1–11, Mar. 2014.

[8] Pappas, C. (2018, March 13). The Top 8 Open Source Learning Management Systems. Retrieved March 16, 2018, from https://elearningindustry.com/top-open-source-learning-management-systems

[9] T. Volery and D. Lord, “Critical success factors in online education,” International Journal of Educational Management, vol. 14, no. 5, pp. 216–223, Sep. 2000.

[10] V. Aldrin, “Evaluating e-Learning: Guiding Research and Practice. By Rob Phillips, Carmel McNaught, and Gregor Kennedy. New York, N.Y.: Routledge, 2012. xxvii + 207 pages. ISBN 978-0-415-88194-4. $42.95.,” Teaching Theology & Religion, vol. 16, pp. e4–e4, Jul. 2013.

[11] E. Islas, M. Pérez, G. Rodriguez, I. Paredes, I. Ávila and M. Mendoza, E-learning tools evaluation and roadmap development for an electrical utility. Journal of Theoretical and Applied Electronic Commerce Research, vol. 2, no. 1, pp. 63–75, 2007.

[12] R. D. Johnson, S. Hornik, and E. Salas, “An empirical examination of factors contributing to the creation of successful e-learning environments,” International Journal of Human-Computer Studies, vol. 66, no. 5, pp. 356–369, May 2008.

[13] S. W. Kim and M. G. Lee, “Validation of an evaluation model for learning management systems,” Journal of Computer Assisted Learning, vol. 24, no. 4, pp. 284–294, Jul. 2008.

[14] N. Phongphaew and A. Jiamsanguanwong, “Usability Evaluation on Learning Management System,” Advances in Usability and User Experience, pp. 39–48, Jun. 2017.

[15] N. F. D. Filho and E. F. Barbosa, “A requirements catalog for mobile learning environments,” Proceedings of the 28th Annual ACM Symposium on Applied Computing - SAC ’13, pp.121-130, 2013.

[16] B. Behkamal, M. Kahani, and M. K. Akbari, “Customizing ISO 9126 quality model for evaluation of B2B applications,” Information and Software Technology, vol. 51, no. 3, pp. 599–609, Mar. 2009.

[17] C. Alves, X. Franch, J. P. Carvallo, and A. Finkelstein, “Using Goals and Quality Models to Support the Matching Analysis During COTS Selection,” Lecture Notes in Computer Science, pp. 146–156, 2005.

[18] B. Little, “Issues in mobile learning technology,” Human Resource Management International Digest, vol. 21, no. 3, pp. 26–29, Apr. 2013.

[19] E. Ossiannilsson and L. Landgren, “Quality in e-learning - a conceptual framework based on experiences from three international benchmarking projects,” Journal of Computer Assisted Learning, vol. 28, no. 1, pp. 42–51, Aug. 2011.

[20] S. Kitanov and , D. Davcev, “Mobile cloud computing environment as a support for mobile learning” In Cloud Computing 2012, The Third International Conference on cloud computing, GRIDs, and Virtualization pp. 99-105, 2012.

[21] H. Movafegh Ghadirli and M. Rastgarpour, “A Paradigm for the Application of Cloud Computing in Mobile Intelligent Tutoring Systems,” International Journal of Software Engineering & Applications, vol. 4, no. 2, pp. 63–73, Mar. 2013.

[22] J. Wishart, D. Green, and Joint Information Services Committee. "Identifying emerging issues in mobile learning in higher and further education: A report to JISC." University of Bristol, 2010.

[23] D. S. K. Seong, “Usability guidelines for designing mobile learning portals,” Proceedings of the 3rd international conference on Mobile technology, applications & systems - Mobility ’06, 2006.

[24] 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, no. 1, p. 1, 2013.

[25] S. Rabi'u, S.A. Ayobami, and H. Okere. "Usability characteristics of mobile applications." Proceedings of International Conference on Behavioural & Social Science Research (ICBSSR), Kampar, Malaysia.(Indexed by Thomson Reuters). Vol. 2. 2012.

[26] M. A. Hamdeh and A. Hamdan. "Using analytical hierarchy process to measure critical success factors of m-learning." European, Mediterranean & Middle Eastern Conference on Information Systems. Abu Dhabi. 2010.

[27] R. Djouab and M. Bari, “An ISO 9126 Based Quality Model for the e-Learning Systems,” International Journal of Information and Education Technology, vol. 6, no. 5, pp. 370–375, 2016.

[28] P. Pocatilu and C. Boja. "Quality characteristics and metrics related to m-learning process." Amfiteatru Economic vol. 11, no. 26, pp.346-354, 2009.

[29] X. Zhang, P. O. de Pablos, and Q. Xu, “Culture effects on the knowledge sharing in multi-national virtual classes: A mixed method,” Computers in Human Behavior, vol. 31, pp. 491–498, Feb. 2014.

[30] D. Taibi, L. Lavazza, and S. Morasca, “OpenBQR: a framework for the assessment of OSS,” IFIP — The International Federation for Information Processing, pp. 173–186, 2007.

[31] V. del Bianco, L. Lavazza, S. Morasca, D. Taibi, and D. Tosi, “An Investigation of the Users’ Perception of OSS Quality,” Open Source Software: New Horizons, pp. 15–28, 2010.

[32] A. Stefani and M. Xenos, “E-commerce system quality assessment using a model based on ISO 9126 and Belief Networks,” Software Quality Journal, vol. 16, no. 1, pp. 107–129, Oct. 2007.

[33] S. Başaran, “Multi-Criteria Decision Analysis Approaches for Selecting and Evaluating Digital Learning Objects,” Procedia Computer Science, vol. 102, pp. 251–258, 2016.

[34] S. Başaran and O. J. Aduradola, “A Multi-Criteria Decision Making to Rank Android based Mobile Applications for Mathematics,” International Journal of Advanced Computer Science and Applications, vol. 9, no. 7, 2018.

[35] S. Başaran and Y. Haruna, “Integrating FAHP and TOPSIS to evaluate mobile learning applications for mathematics,” Procedia Computer Science, vol. 120, pp. 91–98, 2017.

Evaluation Criteria for RDF Triplestores with an Application to Allegrograph

Khadija Alaoui[1], Mohamed Bahaj[2]
LITEN Lab, Faculty of Sciences and Techniques
Hassan I University
Settat, Morocco

***Abstract*—Since its launching as the standard language of the semantic web, the Resource Description Framework RDF has gained an enormous importance in many fields. This has led to the appearance of a variety of data systems to store and process RDF data. To help users identify the best suited RDF data stores for their needs, we establish a list of evaluation and comparison criteria of existing RDF management systems also called triplestores. This is the first work addressing such topic for such triplestores. The criteria list highlights various aspects and is not limited to special stores but covers all types of stores including among others relational, native, centralized, distributed and big data stores. Furthermore, this criteria list is established taking into account relevant issues in accordance with triplestores tasks with respect to the main issues of RDF data storage, RDF data processing, performance, distribution and ease of use. As a study case we consider an application of the evaluation criteria to the graph RDF triplestore AllegroGraph.**

Keywords—RDF; RDFS; SPARQL; triplestore; big data; NoSQL; AllegroGraph

I. Introduction

The primary goal of the W3C (World Wide Web Consortium) standardized ontology language RDF (Resource Description Framework, [24]) and its query language SPARQL (SPARQL Protocol and RDF Query Language, [25]) is to enrich the Web with semantics by structuring data through linking. This goal was set up with the aim to transform the web from a web of documents to a web of intelligent data in order to allow applications to easily extract semantics from data. With the web of documents, there is a difficulty to intelligently follow the semantics of the data because of the lack of structure in the documents content ([9]). For these reasons, there has been a massive use of RDF for publishing data on the web during the last decade. The use of RDF has paved the way for new features and use by scientists and businesses. RDF has indeed been used for modeling and publishing of data in various fields such as health services [3], smart city services [7], Internet of Things [8] and Geography Information Systems (GIS) [23]. This use of RDF has also been accompanied by a rapid development of a multitude of data management systems, also called triplestores, for the storage and processing of RDF data. In the first years of RDF, storage and processing solutions for RDF data were developed based on the use of relational based management systems because of the successful developments of such systems that had been reached over many years. However, these relational solutions present many limitations because of multiple problems such as, among others, SPARQL to SQL (Structured Query Language) query conversion overhead for RDF data querying, complex joins processing imposed by the relational schema proposals for modeling RDF data, integration of other data sources and the handling of big amounts of data. To come up with solutions to the relational problems with regards to RDF data handling, various RDF data management systems have been proposed during the past decade ranging from NoSQL (Not only SQL) based systems through native triplestores to Big Data solutions.

The aim of this work is to give a complete list of evaluation and comparison criteria for RDF management systems. To this end, we first give a summarized categorization of existing triplestores while considering the motivations behind their use for handling RDF data. We identify the benefits of each identified category of systems and the challenges they are facing. In a second step, we establish and motivate an extended evaluation criteria list for triplestores taking into account their associated categorization and relevant aspects with respect to their tasks for handling RDF data.

With the established criteria list, we aim to provide users with detailed insights of the various RDF management systems and comparison aspects with regards to the various relevant issues of dealing with RDF data. Users will be able to differentiate between RDF management systems and identify the best suited triplestore to their data for their specific use cases.

Contrary to existing comparison works that mainly focus on response times of query processing for a limited number of RDF storage systems (e.g. [29], [32], [13]) our list of evaluation criteria for triplestores considers a large variety of aspects. Indeed, based on the categorization details we are considering, various issues related among others to storage models, data organization and data recovery, query processing, query optimization, concurrency, dynamicity, scalability, reasoning, data integration, data exchange, data portability, scalability, visualization and support of analytical functionalities. The detailed criteria list provides users with means to focus on the triplestores aspects that better fulfill their objectives while comparing triplestores. Users can indeed choose the right criteria to identify the drawbacks or the positive aspects of these triplestores.

The following sections are structured as follows. Section 2 presents the W3C standards RDF, RDFS (RDF Schema, [10]), OWL (Web Ontology Language, [18]) and SPARQL as well as a summary about triplestores categories. Sections 3 to 7 present the main categories of comparison and evaluation criteria with motivations behind their associated criteria. Section 8 discusses the case of Allegrograph and Section 9 concludes this work.

II. Semantic Web Standards and Related Work

In this section we present aspects of the semantic web standards RDF, RDFS, OWL and SPARQL ([24], [10], [18], [25]) as well as of associated existing management systems that help in guiding the identification of evaluation criteria for such systems. We also give an overview of research works that deal with comparison and evaluation of triplestores.

A. RDF and SPARQL

RDF semantic language revolutionized the research domain of creation, engineering and processing of ontologies for sharing information on the web. It uses a flexible model where statements in RDF are simply modeled as a set of triples having the form of (S,P,O):=(Subject, Predicate, Object) where a subject represents a resource, an object can be either resource or a literal value and the relation between the Subject and Object is expressed by the Predicate. An object may also be a set of either resources or literals grouped together using RDF grouping constructs such as “RDF:bag”, “RDF:seq” for an ordered list or “RDF:list”. Literal values may have a type and XML types may be used as types of literals.

RDF data can be presented in different formats: XML, Turtle, N-Triples and the N3 (Notation 3). Fig. 1 gives an RDF example using N3 and XML formats. RDF resources and predicates may be endowed with URIs (Uniform Resource Identifiers) to separate data into groups and to allow linkage between graphs to get a web of data.

With the RDF representation of data in form of triples such data can be considered as an oriented graph where nodes are either resources or literals and edges are labeled with predicates. There could be of course more than one edge between two nodes of the graph.

As mentioned above, the W3C standardized query language of RDF data is SPARQL (SPARQL Protocol and RDF Query Language, [25]). A SPARQL query has a SELECT clause and a WHERE clause and may have a FILTER clause to filter the results according to some conditions. In the SELECT clause, attributes to look for are given as variables and these variables are used as substitutes of either subjects, predicates or objects in the triples to look for in the WHERE clause.

```
@prefix fsts: <http://www.fsts.ma/studies#> .
fsts:MachL a fsts:Course .
fsts:MachL fsts:coursename "Machine Learning".
```

(a) N3 Format

```
<rdf:RDF xmlns:rdf= "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
             xmlns:fsts="http://www.fsts.ma/studies#">
<rdf:Description rdf:about=”http://www.fsts.ma/studies#MachL”>
 <fsts:coursename>
  Machine Learning</fsts:coursename>
</rdf:Description>
</rdf:RDF>
```

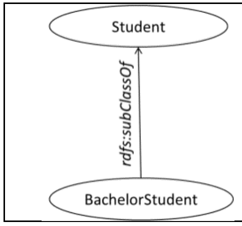
(b) XML Format

Fig 1. Example of an RDF Triple.

B. Schema Languages RDFS and OWL

The RDF schema language RDFS [10]) is the meta-language for RDF data. Statements in RDFS are also RDF triples. RDFS allows RDF resources to be grouped into classes, and allows the declaration of subclasses, properties, subproperties and domains and ranges for properties. An example is given in Fig. 2 where “BachelorStudent” is declared as a subclass of the class “Student”.

Built on top of RDFS, OWL (Web Ontology Language [18]) extends RDFS by adding concepts of classes and properties equivalence, resources equality, symmetric properties, disjoint properties and cardinalities.



```
<rdfs:Class rdf:ID="Student">
</rdfs:Class>

<rdfs:Class rdf:ID="BachelorStudent">
    <rdfs:subClassOf
          rdf:resource="#Student"/>
</rdfs:Class>
```

(a) Graph representation **(b)** XML format

Fig 2. Class Hierarchy in RDFS.

OWL uses "ObjectPropertyDomain" and "DataPropertyDomain" to specify the domains of an object property and a data property. It also offers other inference constructs such as “owl:sameAs”, “owl:inverseOf” and “owl:TransitiveProperty”. Such OWL constructs have the advantage to induce inheritance between classes and similarity between properties and therefore allow reasoning over data through inference.

C. RDF Triplestores

Over the two past decades several systems for the storage and the processing of RDF data have been developed. Those systems called triplestores can be classified into several categories according to the aspects considered for data management [1]. The criteria we are giving in the following section take into account the category of the triplestore chosen for handling RDF data.

RDF management systems can be broadly classified as being relational or non-relational, native or non-native, centralized or distributed and memory or disc based, as well as Map-Reduce based or not relying on Map-Reduce for the case of big RDF data.

Relational RDF stores are solutions that exploit relational database systems to store RDF data. However, the dynamicity of the RDF data is generally not guaranteed by these triplestores. Object relational stores on the other hand provide the link between classic relational databases and object databases. Non-relational RDF stores are those stores that do not rely on relational database systems for handling RDF data. Native triplestores are those systems designed solely for the purpose of handling RDF data. Some of them are disk-based stores (e.g., 4Store [17]) and others are main-memory-based stores (e.g., Cliopatra [35]). NoSQL triplestores are those RDF solutions that use column, document, Key-value or graph

NoSQL databases for handling RDF data. Among NoSQL triplestores we have CumulusRDF that is based on Cassandra ([21]) and SHARD ([28]).

RDF triplestores are also categorized as either centralized or distributed stores. Although centralized triplestores ensure efficient and scalable RDF query processing in a centralized way, they show limitations in storing and processing large amount of data.

RDF management systems can further categorized in cloud based triplestores (e.g., 4Store [17], Amada [4], [11]), mobile solutions designed for mobile devices (e.g., RDF on the Go [22]), and P2P solutions (e.g., Rya [26], Atlas [20], Statustore [33], RAPID+ [27]). Another category of RDF management systems consists of Big Data triplestores that either use Hadoop Map-Reduce (e.g., SHARD [28], HadoopRDF [19], RAPID+ [27], PigSPARQL [30]) or other frameworks such as Spark framework (e.g., S2RDF [31], PRoST [12]).

To be noticed is that a triplestore may belong to one or more of the given categories. The comparison and evaluation criteria given in following sections also considers the categorization of triplestores. Fig. 3 summarizes the list criteria and the classes they belong to.

D. Related Work

As already mentioned, this is the first times a research paper addresses the topic of evaluation and comparison criteria for RDF management systems. Many works mainly dealt with the comparison of some triplestores only with respect to either the amount of RDF data they can store, the loading times of such data or the execution times of SPARQL queries on these data. This is done for example for the comparison of some Big Data and some NoSQL RDF in [6].

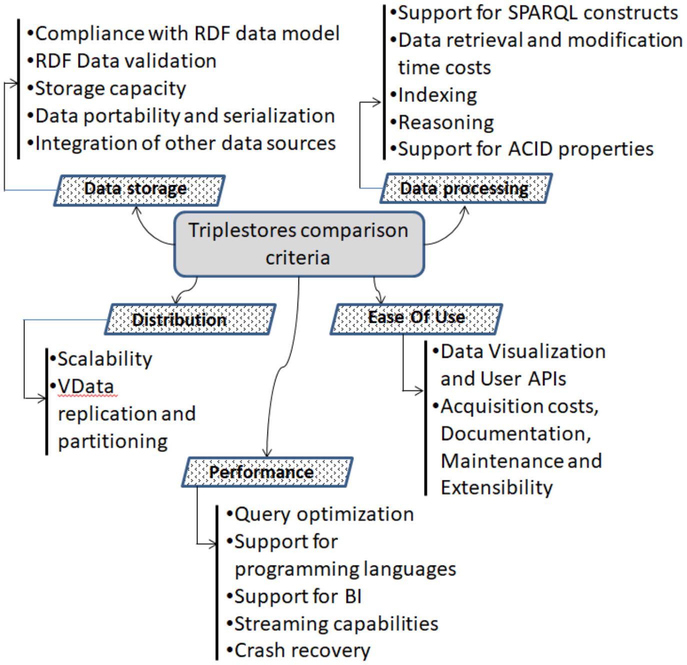

•Compliance with RDF data model
•RDF Data validation
•Storage capacity
•Data portability and serialization
•Integration of other data sources
•Support for SPARQL constructs
•Data retrieval and modification time costs
•Indexing
•Reasoning
•Support for ACID properties
Data storage
Data processing
Triplestores comparison criteria
Distribution
Ease Of Use
•Scalability
•VData replication and partitioning
•Data Visualization and User APIs
•Acquisition costs, Documentation, Maintenance and Extensibility
Performance
•Query optimization
•Support for programming languages
•Support for BI
•Streaming capabilities
•Crash recovery

Fig 3. Comparison Criteria with Associated Categories.

Also such type of comparison has also been done in the context of the specific application domain of smart city services, RDF data loading times and query response times were compared in [7] principally for some NoSQL and relational triplestores using data benchmarks related to smart city services.

III. Criteria Related to RDF Data Storage

In this section we list some important criteria dealing with the capabilities of triplestores to handle RDF data storage. Such criteria involve the respect of RDF data model, RDF data validation, storage capacity, Data portability and serialization and integration of other data sources.

A. Compliance with RDF Data Model

RDF storage solutions have to preserve the flexibility and dynamicity of RDF data. The "-Subject, Object, Predicate" data model and the graph structure of RDF data is beneficial for querying the semantic information and also for adding new predicates without the need to change the schema. It also allows partitioning of the data for the efficient storage and processing of the queries.

B. RDF Data Validation

For triplestores it is also necessary that they provide the possibility for users to validate their RDF data against the constraints and the structures they provide in associated RDFS/OWL schemas. Through validation, not only data conformity with such schemas will be guaranteed but also data exchange and integration will be facilitated.

C. Storage Capacity

The storage capacity for RDF data management systems refers to the possible amount of RDF triples such systems can store and handle. AllegroGraph can handle RDF datasets with more than 1 trillion RDF triples. The Stardog triplestore can handle up to 50 billion triples [31], and GraphDB and Virtuoso triplestores cab handle up to 15 billion [34]. Such information are naturally of great importance for RDF users because of its crucial role in choosing the best suited triplestore for managing their RDF data.

D. Data Portability and Serialization

Data portability would give the opportunity for users to exchange information and content between the services. This requires representation portability mechanisms to be implemented in triplestores. Among such mechanisms, at least export functionalities of RDF data into portable formats such as XML or Json formats are of great importance. In this way, exported RDF data will be machine-understandable and extensible. Furthermore, switching from a triplestore to another one can be easily done.

E. Integration of Other Data Sources

Integration functionalities expected from a triplestore concern above all adding new RDF graphs into the triplestore as well as merging graphs. Also adding non RDF data source is of great importance to allow interoperability with other database systems that are not RDF based. Many existing transformation techniques of other non RDF data sources such as UML, relational and XML already exist and can be incorporated into triplestores to realize such interoperability.

IV. RDF Data Processing Criteria

A. Support for SPARQL Constructs

Without SPARQL support by a triplestore such a triplestore will of course be useless. Triplestores should offer SPARQL querying to extract the desired information in an efficient way while providing support for all constructs of SPARQL 1.1. It is an important requirement to efficiently process queries, especially interactively. Also querying with the use of SPARQL should be possible also for massive amounts of data.

B. Data Retrieval and Modification Time Costs

When considering triplestores, we consider the data, its size and how it is processed. The first thing to consider is how long the triplestore needs to load the data.

Another point to consider is the storage and the retrieval time of the data. Generally, native triplestores are more efficient than existing relational database based triplestores because of the difficulty they face when trying to map the graph based models to SQL.

C. Indexing

The main objective of data indexing is to sort data in order to make its querying easier and faster. Indexing plays an important role especially when managing large amount of data to increase the performance for a large-scale analysis. Indeed, though indexing involves some space overhead, it lets focus only on the portions of data involved by the analysis so that loading these data can be faster and memory space and execution time will be reduced.

The major problem faced by RDF data stores, is how they can build an index data structure over RDF triples. Because of the performance problems related to loading RDF files, or creating suitable indexes, an RDF triplestore must also provide a memory efficient data representation that leaves enough space for the operation of SPARQL querying algorithms.

With regards to indexing, both automatic indexing through the system and the possibility for users to set indexes on specific resources or literal values of triples are of great importance. The former solution will let users not care about indexation and the latter will give them the possibility to index items dependently of their needs.

Triplestores that are relying on relational database management systems have naturally profited from indexing techniques these systems offer.

D. Reasoning

Reasoning allows inferring logical consequences and checking the consistency of a database. It allows a better interpretation and processing of the information for the users.

As mentioned above, RDFS and OWL offer constructs (e.g., “rdfs:subClassOf”, “rdfs:subPropertyOf”, “owl:sameAs”, “owl:inverseOf”, “owl:TransitiveProperty”) for modeling the relations between RDF classes or properties to better structure RDF data in order to avoid problems related for example to redundancies, updates or deletion. However, structuring of information using such constructs will have no sense if the system does not have algorithms for an automatic reasoning that can infer, with the use of such constructs, the hidden information which is implicitly deducible from RDFS/OWL schemas. Concerning AllegroGraph, it allows RDFS reasoning with its built-in reasoned as well as temporal reasoning.

E. Support for ACID Properties

The well-known properties of atomicity, consistency, isolation and durability are of course of great importance for transactions handling [15]. RDF systems that use relational database management systems to store RDF triples have profited from implementations of these properties in these systems. However there is still a lack for support of such properties in non-relational triplestores. Users should therefore be aware of supported properties to ensure that operations of transactions are performed in the right sequence to avoid problems related to inconsistencies, to incomplete executions of such operations or to conflicting operations.

V. Performance Criteria

A. Query Optimization

The query optimizer as a component of triplestores, attempts to find the best way to execute a given query efficiently. It simplifies the query and removes redundant computation. In [5] a methodology using the BGPs and OPTIONALs query optimization techniques for the queries with a mix of UNION and FILTER clauses is proposed.

In term of query optimization, relational based RDF triplestores offer better solutions due to the efforts done on making relational query processing efficient over the last three decades.

B. Support for Programming Languages

It is also to consider if the triplestore serves most modern programming languages (e.g., Java, C++, C#, Python). Within the associated programming APIs RDF Formats and SPARQL query languages should also be supported.

AllegroGraph, for example, offers a Java and Python APIs that implement most of the Sesame and Jena interfaces to access RDF data. It also provides the possibility to Lisp programmers to interact with its RDF repositories.

C. Support for BI

Nowadays, we deal with a huge amount of data and businesses are aware that analyzing and processing those data can generate new opportunities and improvements of the processes. Business intelligence (BI) tools are therefore to be supported by triplestores in order to provide analytical functionalities for users to analyze their data and extract useful information from these data.

D. Streaming Capabilities

Real time processing of data is becoming of high importance, due to the increased sources of real time data (e.g. weather sensors, social networks, IoT tools).

The ability for a triplestore to provide support for streaming will have a crucial role in applications. To this end, the triplestore should support SPARQL querying to be done dynamically over the streams with results given as a continuous streams.

E. Crash Recovery

A important requirement that is be considered, when considering comparison criteria, is the robustness of triplestores toward the system components failure while processing RDF data and the ability to restore the accidently loosed, deleted or corrupted data.

We consider here, as an example, HadoopRDF ([19]) which provides an architecture that stores triples on HDFS. It replicates the triples on multiple machines and decomposes a user query into partial queries with an independent evaluation of these queries without any communication overhead between the partitions.

VI. Criteria for Distributed Triplestores

A. Data Replication and Partitioning

Due to the quick increase of the scale of RDF data, various distributed storage systems have been developed. For such systems partitioning and replication capabilities while handling RDF data is necessary to distribute both data and processing among RDF nodes. For the distribution of data, partitioning techniques should be efficient enough to achieve a reasonable query processing performance together with efficient data transfers between the nodes.

There are two types of data partitioning in existing RDF stores. The first type is the static graph partitioning, which creates partitions with a minimum of edges. The second type is the workload aware partitioning, which faces however the complex problem of choosing the right decisions regarding space and workload [2].

On the other hand, replication refers to the storage of the same data in several different locations. Of course, such replication requires the availability of synchronization mechanisms between the data sources to guarantee consistency between the replicated data. To this end, good strategies are to be provided by triplestores to select the RDF data to be replicated, to control the storage availability and to handle data changes related to updates, insertion or deletion. For example, the distributed triplestore DREAM [16] does not partition data over nodes but simply replicates the whole data in every node, which necessitate updating the same data each time changes occur.

B. Scalability

A distributed triplestore should have the ability to scale either vertically with the possibility to add data resources to the nodes, or horizontally with the possibility to add more nodes to the system. This is a relevant property for handling large amounts of data that are gathered from various sources as well as for integrating data from classical databases (e.g., XML, relational or file systems).

Because of the graph nature of RDF data, good strategies are needed to achieve both arts of scalability in order to achieve efficient SPARQL search, delete or update queries. Indeed, such queries may involve complex joins of subgraphs and therefore an extra time complexity. The development of Big Data technologies and frameworks (e.g., Hadoop, Map Reduce and Spark) has also favored the development of various scalable triplestores based on such technologies (e.g., SHARD [28], HadoopRDF [19], PRoST [12] and CliqueSquare [14]).

VII. Ease of Use Criteria

A. Data Visualization and User APIs

With APIs (Application Programming Interfaces) we mainly mean those APIs that make it easier for triplestores users to interact with their data easily to query their RDF data and to have their data presented in a user friendly way. The list should also include APIs for programming languages or for the use existing RDF/SPARQL programming packages such as the use of Jena.

Visualization of RDF in several ways has also to be taken into account for the understanding of different RDF data structures. Principally, a triplestore, because of the nature of RDF data, should support RDF data presentation in form of graphs.

With regards to APIs, relational databases based triplestores have largely profited of existing APIs developed for relational database systems.

B. Acquisition costs, Documentation, Maintenance and Extensibility

Two other points to consider are the product costs and the learning costs that are associated of a triplestore and its implementation.

Also the development conditions of a triplestore are also to be considered, together with its documentation, maintenance, accessibility and performance.

As stated, it is important to check how long a triplestore is used, and to also get an overview of possible updates, releases development and dedicated extensibility mechanisms. This will provide an idea about the triplestore, if it is an individual initiative, an active or a non-active project, if it is dependent to a third party application and if it is an open source system.

It is also important to consider, if the store is brightly used, in which domains it is used and how long it is being in use. These factors play an important role in the decision regarding the adoption of such a triplestore or not.

VIII. Case Study: AllegroGraph

It is absolutely evident that an evaluation of triplestore should be done in the context of its comparison with other stores belonging to its category using associated established criteria. However for specific applications, the triplestore could also be compared with stores not belonging to its category and in this case such comparison needs to only be conducted with respect to some specific criteria pertaining to the specific use in applications. Both types of comparison will lead to further research papers and constitute one of our future perspectives.

However, to illustrate the application of the established list of criteria, we discuss in this section the case of AllegroGraph triplestore with the NoSQL triplestore XX and with the Big Data triplestores HadoopRDF. As mentioned, a thorough comparison of AllegroGraph with other triplestores from Graph stores and other types of stores using the aforementioned criteria will be the subject of another research paper.

AllegroGraph is an efficient RDF native graph database that uses disk storage, which allows it to scale up to billions of triplets. It was developed to meet RDF standards and It has been continuously further improved since its appearance in 2004. It also offers interfaces for many programming languages such as Java, Python, Ruby, C#, and Scala.

Inference is also supported by AllegroGraph under two angles. On one hand, AllegroGraph offers the so-called "dynamic RDFS++ reasoned" that implements a set of RDFS inference rules and also OWL2resoner. The first reasoner generates inferred triples during inference execution without saving inferred triples. However, the OWL2 reasoner adds generated triples to the considered triples database.

AllegroGraph also has components for the analysis of social network and geospatial data. It also supports visual generation of SPARQL queries as well as visualization of graphs using Gruff. A free, developer and enterprise versions of Allegrograph with storage capacities of respectively 5, 50 and 50+ million triples are provided for users.

In comparison with other Graph oriented triplestores and even to other kind of RDF stores, AllegroGraph fulfills by far many of the criteria mentioned above. We can say that many of such RDF management systems are still at their infancy phase since they are still limited to RDF storage and SPARQL processing functionalities.

For example, HadoopRDF is a Big Data triplestore [19] that uses the Hadoop file system for the distributed storage of RDF data in a cluster of nodes and Map Reduce framework for SPARQL query processing. In comparison of HadoopRDF with AllegroGraph, HadoopRDF also shows a high failure tolerance and reliability. Indeed, Hadoop based triplestores can be easily implemented on clusters of so called commodity computers and the cluster can continue functioning after node failure. Therefore HadoopRDF can also handle very large amounts of RDF data. With regards to RDF querying, processing of SPARQL queries is done in HadoopRDF efficiently since it partitions the RDF data not in a single file but in a set of small files and Map Reduce jobs are simply run on small portions that are of concern [19].

Apart from RDF data modeling compliance, storage and querying, HadoopRDF has not been further developed since its appearance and show strong limitations with respect to the other criteria already listed in comparison with AllegroGraph. However, because of the Hadoop architecture of HadoopRDF, HadoopRDF can also be easily extended and further yields other research perspectives. Indeed, this fact will let HadoopRDF benefice from the analytical technologies and APIs already developed within the framework of Hadoop.

IX. Conclusion

We have established a list of criteria for the comparison and evaluation of RDF triplestores. To achieve this task, we provided a methodology relying on the identification of expected key characteristics for triplestores. This is done by categorizing the criteria according to: - RDF data storage (e.g., Compliance with RDF data model, RDF Data validation, Storage capacity, Data portability and serialization, Integration of other data sources), - RDF data processing (e.g., support for SPARQL constructs, data retrieval and modification times, indexing, reasoning, support for ACID properties), - performance (e.g., query optimization, support for programming languages, support for BI, streaming capabilities, crash recovery), - distribution (e.g., data replication, scalability), - and ease of use (e.g., user APIs, visualization, acquisition costs, documentation, maintenance and extensibility). As an illustration of the criteria list, we considered the case of AllegroGraph triplestore and showed that AllegroGraph fulfills many of these criteria.

The criteria will play an important role in supporting users to make accurate decisions for the adoption of the appropriate triplestore that best suit their objectives and will help in identifying the strength and weaknesses of existing triplestores.

This research work is as far as we know the first work that addresses comparison and evaluation criteria for triplestores. Because of the increasing use of RDF in many application domains, the established list of comparison and evaluation criteria will surely pave the way for more research works that deal with further improvements of the functioning of existing triplestores or with the development of new ones.

References

[1] K. Alaoui. "A Categorization of RDF Triplestores," Proc. Smart City Applications, SCA-2019, October 2–4, 2019, Casablanca, Morocco, ACM, ISBN 978-1-4503-6289-4/19/10, DOI 10.1145/3368756.3369047, 2019.

[2] A. Al-Ghezi and L. Wiese, "Adaptive workload-based partitioning and replication for RDF graphs" Database and Expert Systems Applications, 2018.

[3] S. Anand and A. Verma, "Development of Ontology for Smart Hospital and Implementation using UML and RDF," IJCSI Int. J. of Computer Science Issues, Vol. 7, Issue 5, 2010.

[4] A. Aranda-Andujar, F. Bugiou, J. Camacho-Rodriguez, D. Colazzo, F. Goasdoué, Z. Kaoudi, and I. Manolescu, "Amada: Web data repositories in the Amazon cloud," in Proc. 21st Int. Conf. on Information and knowledge Management, CIKM"12, maui, 29 Octpber-02 November 2012, ACM, pp. 2749-2751, 2012.

[5] M. Atre, "Algorithms and analysis for the SPARQL constructs," arXiv:1805.08037v3 [cs.DB] 23 May 2018.

[6] M. Banane and A. Belangour, "An Evaluation and Comparative study of massive RDF Data management approaches based on Big Data Technologies," Int. J. of Emerging Trends in Engineering Research, vol. 7, no. 7, July 2019.

[7] P. Bellini and P. Nesi, "Performance assessment of RDF graph databases for smart city services," J. Vis. Lang. Comput. 2018, 45, 24–38.

[8] M. Bermudez-Edo, T. Elsaleh, P. Barnaghi, and K. Taylor, "IoT-Lite: A Lightweight Semantic Model for the Internet of Things," in 2016 International IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, July 2016, pp. 90–97.

[9] T. Berners-Lee, J. Hendler, and Ora Lassila. “The Semantic Web,” Scientific American, pages 29–37, May 2001.

[10] D. Brickley and R.V. Guha, "RDF Schema 1.1, W3C Recommendation," https://www.w3.org/TR/rdf-schema/, 2014.

[11] F. Bugiou, F. Goasdoué, Z. Kaoudi, and I.Manolescu, “RDF data management in the Amazon cloud,” in ICDT/EDBT Workshops 2012.

[12] M. Cossu, M. Färber, and G. Lausen, "PRoST: Distributed Execution of SPARQL Queries Using Mixed Partitioning Strategies," in Proc. of the 21st International Conference on Extending Database Technology (EDBT), March 26-29, 2018.

[13] J. V. F. Dombeu, and R. Kwuimi, “Semantic data storage in information systems,” African J. Of Information Systems, 2018.

[14] F. Goasdoué, Z. Kaoudi, I. Manolescu, J.A. Quiané-Ruiz, and S. Zampetakis, “CliqueSquare: Flat plans for massively parallel RDF queries,” in Proc. IEEE 31st International Conference on Data Engineering, 2015, pp. 771–782.

[15] T. Haerder, A. Reuter, "Principles of Transaction-oriented Database Recovery.," ACM Comput. Surv. 15, Nr. 4, 1983, pp. 287–317, doi:10.1145/289.291

[16] M. Hammoud, D.A, Rabbou, and R. Nouri, “DREAM: Distributed RDF Engine with Adaptive Query Planner and Minimal Communication,” VLDB Endowment, 2015.

[17] S. Harris, N. Lamb, and N. Shadbolt, "4store: The design and implementation of a clustered RDF store." In Proc. Scalable SemanticWeb Knowledge Base Systems - SSWS2009. pp. 94–109, 2009.

[18] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, "OWL 2 Web Ontology Language Primer,” https://www.w3.org/TR/owl-primer/, 2012.

[19] M. Husain, J. McGlothlin, M.M. Masud, L. Khan, and B.M. Thuraisingham, “Heuristics-Based query processing for large RDF graphs using cloud computing,” IEEE Trans. on Knowl. and Data Eng., 2011.

[20] Z. Kaoudi, M. Koubarakis, K. Kyzirakos, I. Miliaraki, M. Magiridou, and A. Papadakis-Pesaresi, "Atlas: Storing, updating and querying RDF(s) data on top of DHTS," J. of Web Semantics 8 (4), pp. 271–277, 2010.

[21] G. Ladwig and A. Harth, "CumulusRDF: linked data management on nested key-value stores,” SSWS 30, 2011.

[22] D. Le-Phuoc, J. X. Parreira, V. Reynolds, and M. Hauswirth, "RDF on the go: an RDF storage and query processor for mobile devices," In ISWC Posters&Demos. 2010.

[23] G. Mai, K. Janowicz, B. Yan, and S. Scheider, "Deeply integrating linked data with geographic information systems," Transactions in GIS, 230 (3), pp. 579–600, 2019, doi: 10.1111/tgis.12538.

[24] F. Manola, E. Miller, and B. McBride. “RDF 1.1 Primer," http://www.w3.org/TR/rdf-primer/, 2014.

[25] E. Prud’hommeaux and Andy Seaborne. SPARQL Query Language for RDF, W3C Recommendation. http://www.w3.org/TR/rdf-sparqlquery/, 2008.

[26] R. Punnoose, A. Crainiceanu, and D. Rapp, “Rya: A scalable RDF triple store for the clouds,” in 1st Interna=onal Workshop on Cloud Intelligence (in conjunction with VLDB), 2012.

[27] P. Ravindra, H. Kim, and K. Anyanwu, “An intermediate algebra for optimizing RDF graph patern matching on MapReduce”, in ESWC 2011.

[28] K. Rohloff and R. E. Schantz, "High-performance, massively scalable distributed systems using the MapReduce software framework: the SHARD triple-store," In Programming Support Innovations for Emerging Distributed Applications, pp.1-5, October 17-21, Reno, Nevada, 2010.

[29] S. Sankar, A. Sayed, and J. A. Bani-younis, “A schematic analysis on selective-RDF database stores," Int. J. of Computer Applications 86(11), pp. 21-28, January 2014..

[30] A. Schätzle, M. Przyjaciel-Zablocki, and G. Lausen, ”PigSPARQL: mapping SPARQL to Pig Latinn”, in SWIM 2011.

[31] A. Schätzle, M. Przyjaciel-Zablocki, S. Skilevic, and G. Lausen, “S2RDF: RDF querying with SPARQL on Spark,” arXiv Prepr., pp. 804–815, 2015.

[32] M. Schmidt, O. Görlitz, P. Haase, G. Ladwig, A. Schwarte, and T. Tran, ” FedBench: A benchmark suite for federated semantic data query processing," In Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) Proceedings of the 10th International Semantic Web Conference (ISWC 2011), Bonn, Germany, October 23-27, 2011, Part I. pp. 585–600. Springer 2011.

[33] R. Stein and V. Zacharias, “RDF on cloud number nine,” in 4th Workshop on New Forms of Reasoning for the Seman=c Web: Scalable and Dynamic, AWS SimpleDB, May 2010.

[34] W3C, https://www.w3.org/wiki/LargeTripleStores.

[35] J. Wielemaker, W. Beek, M. Hildebrand, and J. van Ossenbruggen, "ClioPatria: A SWI-Prolog infrastructure for the semantic web," Semantic Web 7(5), pp. 529-541, 2016.

A Review on Honeypot-based Botnet Detection Models for Smart Factory

Lee Seungjin[1], Azween Abdullah[2], NZ Jhanjhi[3]
School of Computer Science and Engineering (*SCE*)
Taylor's University
Subang Jaya, Selangor, Malaysia

***Abstract*—Since the Swiss Davos Forum in January 2017, the most searched keywords related to the Fourth Revolutionary Industry are AI technology, big data, and IoT. In particular, the manufacturing industry seeks to advance information and communication technology (ICT) to build a smart factory that integrates the management of production processes, safety, procurement, and logistics services. Such smart factories can effectively solve the problem of frequent occurrences of accidents and high fault rates. An increasing number of cases happening in smart factories due to botnet DDoS attacks have been reported in recent times. Hence, the Internet of Thing security is of paramount importance in this emerging field of network security improvement. In response to the cyberattacks, smart factory security needs to gain its defending ability against botnet. Various security solutions have been proposed as solutions. However, those emerging approaches to IoT security are yet to effectively deal with IoT malware, also known as Zero-day Attacks. Botnet detection using honeypot has been recently studied in a few researches and shows its potential to detect Botnet in some applications effectively. Detecting botnet by honeypot is a detection method in which a resource is intentionally created within a network as a trap to attract botnet attackers with the purpose of closely monitoring and obtaining their behaviors. By doing this, the tracked contents are recorded in a log file. It is then used for analysis by machine learning. As a result, responding actions are generated to act against the botnet attack. In this work, a review of literature looks insight into two main areas, i.e. 1) Botnet and its severity in cybersecurity, 2) Botnet attacks on a smart factory and the potential of the honeypot approach as an effective solution. Notably, a comparative analysis of the effectiveness of honeypot detection in various applications is accomplished and the application of honey in the smart factories is reviewed.**

Keywords*—*IoT; smart factory; honeypot; Botnets; detection; security; model

I. Introduction

Smart plant strategies are being pushed forward to innovate global manufacturing competitiveness. Germany is undergoing the Industry 4.0 process. It builds manufacturing into an automatic production system through the Internet of Things, Initiated in China 2025 in China, Terrain Manufacturing System in Japan and Seoul is pushing for Manufacturing Innovation 3.0 [1]. Smart factories in the era of the Fourth Industrial Revolution refer to consumer-oriented intelligent factories that incorporate digital new technologies and manufacturing technologies beyond the current level of factory automation (FA). It can produce a variety of products from one production line and is expected to change from mass customization to flexible production systems through modularization. It is possible to save energy by changing from a person-centered working environment to an ICT-oriented one, and it is expected that the productivity of the manufacturing industry will increase [2], Various possibilities for the transition to smart factories are recognized. It is predicted that it will be able to monitor and control manufacturing sites via virtual space, making it easier to manage factories. It will enhance competitiveness in quality and cost [3]. Smart factories are closely linked to data by application of the latest ICT technologies such as AI, Blockchain and hyper-automation, Augmentation as well as IoT as shown in Fig. 1.

Based on that, production processes are controlled on their own, making the industrial control system (ICS, Industrial Control System) more complex and advanced than the ordinary systems. However, due to the complexity of the system and the application of new technologies, the advancement in smart factories raises the risk of new security threats that have not occurred earlier. Specifically, the number of attacks on actual cyber vulnerabilities has increased sharply in recent years on physical equipment and software in power generation, energy, and manufacturing [5].

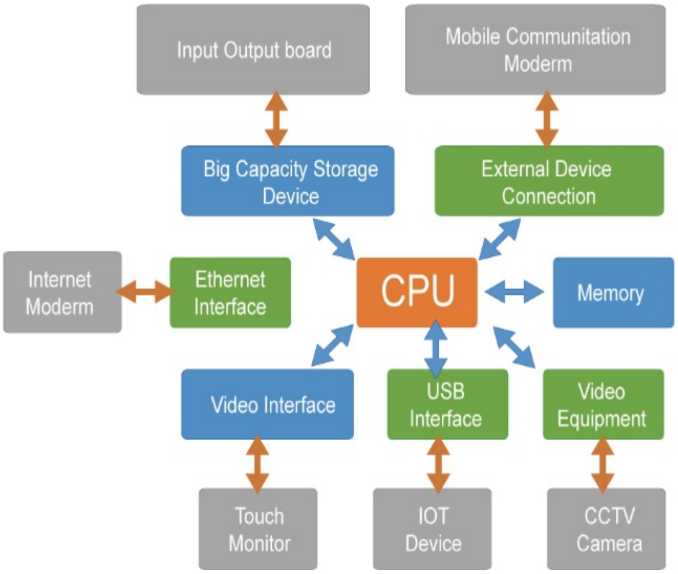


Fig. 1. Smart Factory Function Requirements[4].

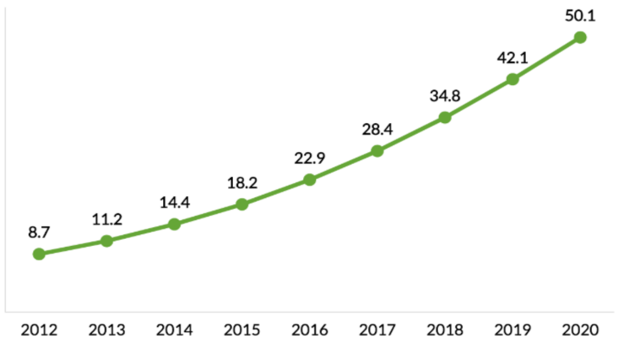


Fig. 2. Growth in the Internet of Things Devices [9].

In order for smart factories to operate and maintain their autonomy, they must analyze themselves and accurately carry out product design as well as quality process management and production management. It utilizes important information such as process know-how, requirements of analysis data, product design diagrams, business secrets and research and development results. The threat of leaking confidential information by exploiting the security vulnerability of wireless communications via remote control or monitoring of systems in a factory using wireless devices could cause severe damage and economic loss.

In the last five years, there is a tremendous increase in the use of IoT devices from 8 billion to 25 billion, noticeably in the application for smart factories [6]. However, one significant limitation of these IoT devices is its instability, as identified in 250 vulnerable features (shown in Fig. 2) [7]. In an analysis of 10 most popular IoT devices such as the open telnet ports, firmware of outdated Linux and the transmission of sensitive data without being encrypted [8]

Due to the instability of the IoT devices on the infrastructure of the Internet, they become an ideal target of the IoT botnet with a surging number of attack cases. A good example is an attack that occurred in October 2016 when the Dyn DNS infrastructure involving 100,000 IoT devices (mainly on CCTV cameras). It was under the attack of a DDoS (distributed denial of service) caused by Mirai botnet. This DDos attack resulted in temporary unavailability in several commercial websites like Amazon, Netflix, Twitter, CNN and PayPal. Another example is the new Mirai source code released in 2017. It rendered increasing DDos attacks. Such Mirai-induced IoT botnets occurred more frequently in recent times and their consequences became more severe. Thus, the identification and mitigation of IoT botnets are urgently needed due to the development of new technologies [10].

With a great potential of machine learning (ML) emerging recently, it offers a new solution for anomaly detection of any malicious internet traffic. Indeed, internet traffic is distinct from other Internet-connected devices (smartphones and computer laptops) in a way that communication between IoT devices is allowed by small endpoints contained in a limited set. In contrast, internet-connect devices use a variety of web servers. Moreover, for the IoT devices, the patterns of network traffic are repetitive in the regularity of network pings with small packets for logging. Interestingly, there is a scarcity of research investigating in development of ML models featured in networks of IoT devices and attacks in IoT traffic using the machine learning approach.

Botnets are a collection of computers associated with the Internet that are compromised and are controlled distantly by intruders through malicious software, normally called bots. Malicious software is generally used by attackers [11]. Mirai botnet. It was recently unveiled as an open source, used the vulnerability of such an IoT device to launch a botnet attack on the DSN provider, DYN. The reason for the significance of a botnet attack is that it was not a computer but an IoT device such as a webcam. Besides, similar incidents are expected to occur due to the open code as well as attacks that have already been carried out. Mirai botnet starts with an attacker approaching a random IP address. An attacker uses a pre-specified ID and password to gain root privileges on the device. An attacker has an ID and a cryptographic list of several IoT devices that are specified when a product is sold, thereby gaining the device's ROOT authority from the contents of the list. This is a vulnerability that occurs because users do not change passwords on their IoT devices even after they purchase the products [10].

If root privileges are obtained in this way, the shell script executes an infection behavior in the device. The infection code is downloaded and executed from the attacker's loader server to infect the device completely. When the surveillance is complete, a standby message is sent to the attacker's C&C server in Fig. 3 [12]. The same routine that infects other IoT devices and increases the bot exponentially. The attacker then sets a target of the attack and commands the bots to execute the DDOS attack as the target of an attack on the attacker server. To prevent the infection of the BOT of a DDOS attack, the code of the attacker, who is connected to find, defend and detect attacks of the same pattern was proposed to be analyzed [13].

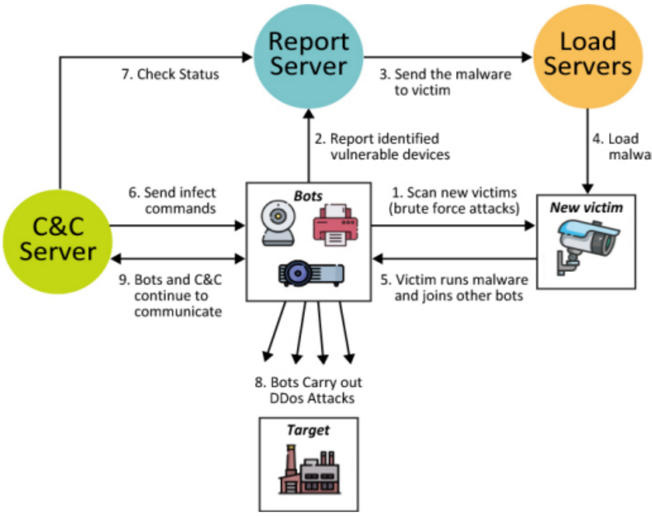

Report Server
Load Servers
7. Check Status
3. Send the malware to victim
2. Report identified vulnerable devices
4. Load malware
C&C Server
6. Send infect commands
Bots
1. Scan new victims (brute force attacks)
New victim
9. Bots and C&C continue to communicate
5. Victim runs malware and joins other bots
8. Bots Carry out DDos Attacks
Target

Fig. 3. Mirai Botnet Operations [12].

A. Review Aim

The review aims to compare various methods of detecting botnet attacks and to evaluate the potential of using honeypot as a solution to botnet attacks in the smart factory situation.

II. Related Work

The work mainly focuses on two sections. The first section presents the threat of cyber-attacks to the IoT based system and narrows the case of the smart factory with the utilization of IoT equipment. The second section addresses the latest security solutions to the botnet attack, especially honeypot detection for a smart factory. A comparative analysis is made among related studies with respect to the advantages and disadvantages of the honeypot and other similar methods. Finally, future work is suggested for developing a botnet detection model using the Honeypot approach.

A. Smart Factory Security

In the Industry 4.0, smart factories raise interest in manufacturing companies and academic researchers [14]. Although smart factories are already constructed and operated in the industry, implementation standards are yet to be established for smart factories [15]. The manufacturing system could be rated in scale based on different perspectives, such as function [16]. A smart factory is conceptualized as adaptive and flexible manufacturing consisting of three aspects, i.e., interconnection, collaboration and execution [17]. Furthermore, systems in the smart factory's architecture for IoT are segregated into four layers arranged hierarchical starting at physical resource layer, networking layer, application layer and interface layer, as shown in Fig. 4 [18] [19]. With the aim of transforming a modern factory into a smart factory, key technologies related to all layers require in-depth research.

The key element of a smart manufacturing system, such as smart factories is intelligence. It is based on network technology and manufacturing data. Additionally, system maintenance and requirements of manufacturing should be incorporated into the implementation of smart factories. Due to such complexity in the design and operation of the smart factories, many technical problems arise and need to be solved [20].

Although the industrial radio sensor network (IWSN) has advantages for industrial manufacturing networks, the Internet of Things (IoT) can be used to apply for the network layer with support of new protocols and new data formats at higher flexibility and scalability. At the data application layer, the cloud platform should be able to perform semantic analysis of various data. Therefore, modeling smart plants should consider the employment of ontology to provide self-organization, self-learning and self-adaptation capabilities. In addition, data analysis and data mining are useful to provide a scientific basis for decision-making and design optimization, respectively. All layers of smart factories are developed and analyzed by focusing on core technology [21].

At the sensing or physical resource layer, the acquisition of real-time information is obtained by physical equipment and transmission of heterogeneous information at high-speed through communication devices. As a result, rapid reconfiguration and adaptability have to be ensured at the workplace by increasing the intelligence of these basic devices/equipment to meet the requirements for the Internet of Things [22]. In the entire operating cycle of smart factories, including the physical resources, the efficiency of smart manufacturing to produce customized products creates new demand for adjusting manufacturing equipment, production lines and data acquisition. Due to the limitation in flexibility and configurability, current manufacturing equipment in the workplace is highly specific and relatively narrow in scope, making it vulnerable to adapt to changes in the manufacturing environment.

Some manufacturing equipment such as robots, mechanical arms, machining centers can be modulated in the manufacturing unit to improve a dynamic scheduling. This leads to a reconfiguration of the controller and extension of the manufacturing equipment function. The assembly capability of the workshop can be improved to be more adaptive and self-configurable at the robotic islands for each modular manufacturing units. Because of this, the capability of flexible manufacturing can be enhanced at the integrated management level [20]. A variety of algorithm was proposed [23]. This is for the reconstruction of grid-based, modular self-reconfiguring robots that dramatically simplify the complexity of robot configurations through a repetitive approach.

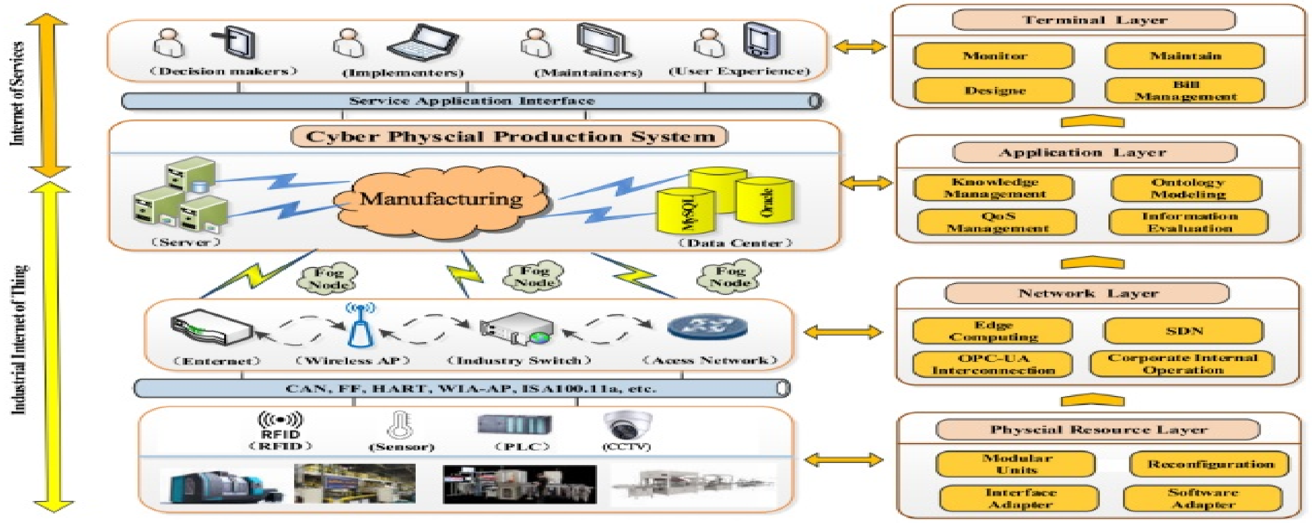


Fig. 4. Function Requirements of Smart Factories [18].

cyber-physical integration of cognitive robots in manufacturing was proposed [24]. Specifically, a humanoid robot is used to integrate with the smart manufacturing plant in coordination with the execution system. These robots can cognitively adjust their own manufacturing behaviors to recognize information uncertainty, changes in schedule management and independently dealing with complex problems of manufacturing. There is a relationship between the level of intelligence of smart factories and the modular maturing units. Therefore, increasing the intelligence of the modular manufacturing units such as robotic units is very important to allow them to work together to accomplish common tasks by mutual recognition and co-working mechanisms. The heterogeneity of interactions should also be considered. It is important to create an optimal or lane combination method because the functions of other modular manufacturing devices can be duplicated for a particular product. Each manufacturing unit not only meets the manufacturing requirements of the product, but it can also systematically improve the smart factory efficiency on its own. However, a deadlock can occur in smart manufacturing as some different products enter the production system in a disordered manner [25]. Currently, a solution to a deadlock in a flexible manufacturing system increasingly becomes a research interest [25].

Smart factories operate in an IoT platform. It collects and processes information from various RFID devices [26]. Fig. 5 shows the system structure of the study that manages various data by integrating middleware and sensors (RFID readers) into a single framework based on RFID security network systems. RFID is a technology that collects information from tags through signals that detect and transmit radio frequencies from tags installed on devices. The application of RFID technology is considered to be the key to establishing an environment for smart factories with ubiquitous security networks. Because of the many advantages in the real-time product, equipment transportation and automatic information collection, various areas such as plant facility management and worker safety management are being applied [27]. Besides that, RFID technology is also used in building security networks of the smart factories by gathering tags and removing duplicated or unnecessary tag information. RFID middleware is needed to provide information only in application programs and to manage security under the framework [27].

The RFID system is constructed. The RFID system consists of a reader, a questionnaire, a tag and a transponder. Readers need to recognize tags in the fast-paced reading range. The reader has a radio communication module and an interface to communicate with the antenna, power supply, math and memory and host systems. The reader is relatively less restrictive in terms of power, computing power and memory size. It can be divided into fixed and mobile types depending on its mobility. The reader sends the identified tag information to the host computer system. It consists of database and application software to processes the tag information [29].

Tags being attached to objects are in the form of microchips with unique allocation IDs that contain information about each object. The tag has relatively big limits on the power, operating functions, adding memory and antenna size required to operate a circuit. Tags can be classified into different types depending on the nature of the internal memory or the presence of batteries. Classification of RFID tags is made according to the characteristics of the internal memory or the availability of batteries [30]

RFID tags, which are widely used in smart factories in IoT platforms, are equipped with a variety of memories, i.e. to read-only or to read/write memory types. The classification of RFID tags is based on battery installation. First, passive tags have no power source for their own functions. Instead, they utilize the power generated by the electromagnetic signals of the readers with similar functions. Secondly, semi-passive tags equipped with their own internal battery are able to be identified at a longer distance. Thirdly, active tags are more advanced than the semi-passive tags. Due to their longer wavelength and their own supply of power, their capacity covers are not only for channel detection but also for collision detection. In comparison between the three RFID tags, the active tag is the most expensive, followed by the semi-active whereas the more economical one is the passive tag [31].

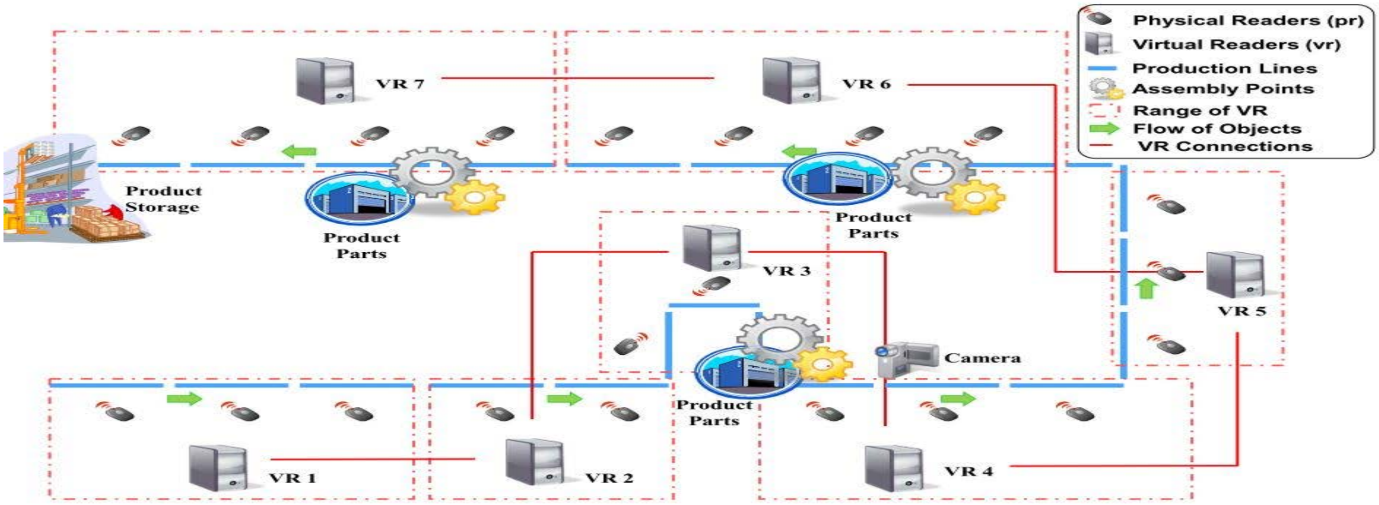


Fig. 5. RFID System Deployment in Smart Factories [28].

Data written on the RFID tags are collected by the process sensors, known as RFID tag readers. These data are then combined to become real-time information for smart factories to analyze in the IoT platform. Such analyzed information is crucial for communication like warning notifications for the management [28].

B. *Botnet Detection*

Serious damages caused by botnet attack occurs in online banks, e-commerce Internet systems and industries as a whole. Such damages become increasingly threatening to users as well as the service providers. Therefore detection is imperative.

One of the severe cybercrimes is botnets, which are termed software robots or boats that operate automatically and systematically. A four-stage bot is typically created and maintained. Zombie computer groups are controlled remotely by attackers who call them botmasters [32].

Stage 1 - Initial infection: The computer can be infected in several different ways. For example, 1) being actively exploited. There are some vulnerabilities in the host, such as DCE-RPC. The malware then runs on the host and exploits the vulnerability, 2) The malware is automatically downloaded while viewing the web page, 3) The malware is downloaded and executed automatically by opening an e-mail attachment, 4) USB autorun.

Stage 2 - Injection: In this stage, the infected host downloads and runs the bot code and then becomes a real bot. Downloading is available via FTP, HTTP and P2P (e.g., Trojan horses).

Stage 3 - Malicious activity: The bot communicates with the controller to get instructions for performing the activity like spam, DDoS and scanning. Currently, there are more sophisticated methods called fast flush service networks. Command communication can avoid single points of failure using IRC-based, HTTP-based, DNS-based, or P2P protocols[33].

Stage 4 - Maintenance and upgrade: Botnets are always classified according to control structures and commands. The bot will continue to upgrade its binary number. At this stage, for example, the Internet Relay Chat (IRC)-based bonnet is an IRC protocol user[34]

Network devices that have a low level of information security and mass distribution of personal computers, as well as IoT devices, are the attractive targets for cybercriminals. The most serious attacks in recent years have been made by a botnet consisting of unsecured IoT devices. Among the botnets, Mirai botnet, as the largest botnet in history, has affected a vast number of IoT devices [35]. The working principle of the Mirai is that performing an IPv4 address space scan to identify vulnerable devices with open port TCP / 23 and TCP / 23233 [36] used by the network service TELNET [37] and then conduct a robbery attack on these ports.

In the work of Nguyen et al., (2020), the mechanism of spreading Mirai botnet was identified and its effects on IoT devices investigated. A combination of more than 60 basic user credentials is employed in Mirai to access the shell of any devices which are open to the public. Once, a smart device becomes a part of the botnet, other connected devices which are subjected to vulnerability as it will scan for other IPv4 address spaces. It would be subsequently identified and damaged. Despite becoming a botnet, receiving the order of the energy to perform malice, the botnet infected devices are still able to carry out the default activities set by the manufacturer. Such attacks caused by Mirai botnet has laid a foundation for the rising of botnet targeting IoT devices such as botnet list and botnet amnesia. For example, the attack that happened to Telnet and SSH services caused by most botnets resulted from gaining unauthorized access to IoT devices. Another example is the unauthentic access to nearly 400,000 IoT devices via two services reported in the Cybersecurity Survey on IoT [38]. Therefore, due to the availability of numerous network devices that are vulnerable to protection, botnets remain one of the main concerns of cyberspace. To become a useful part of the botnet, vulnerable network devices go through the step sequence.

The first step of the lifecycle compromises with vulnerable devices that are considered potential boats. In the second step, the malicious code that is required to communicate with the botmaster is downloaded and installed. The third step is to connect to the Command and Control (C&C) server and receive instructions from the botmaster. The next step is malicious activity, which assumes malicious behavior in accordance with the instructions of the botmaster of the infected host. The final step consists of upgrades and maintenance. This step is essential for botmasters to effectively monitor infected hosts as long as possible and modify their behavior as follows: installing malware updates to prevent the loss of large-scale malicious activity by botnets, breaking the representative chain at any stage.

Other detection methods are for after botnet penetration. In particular, post-intrusion measures are much less effective in terms of detection rates. Contrary to previous research, honeypot protects IoT equipment inside the smart factory by installing the trap in advance, not after the botnet intrusion. The following studies compare the pros and cons.

Signature-based detection is a method of analyzing, scaling and detecting botnets based on their knowledge. A typical one is Snort. The signature-based detection method has the advantage of high detection and low false detection for botnets and malware, as previously found. However, if a new botnet attacks, it is not able to detect it because it does not have a signature. For example, in Lishi, a bot is judged by the IRC name of the bot. The IRC name of the boat was thought to be much different from the nickname of the end-user. However, making IRC nicknames similar to end-users is difficult to detect and impossible to detect without an IRC-based botnet [39].

Anomalies-based detection is a way to suspect and detect strange things that behave differently than normal users in the network traffic, such as intensive traffic or abnormal port use. Yeung’s study presented a method for detecting botnets by analyzing the data flow data of the transport layer. The data suspected from the bot is extracted separately from the data flow and scores are calculated and determined by the bot if the score exceeds the threshold. This method is detectable, even if botnet communication is encrypted and has a low error rate. It is also highly scalable and can help to measure the size of a botnet. Again, only IRC-based botnets can be detected [40].

Bots communicate through command and control, bots and botmasters exchange messages to regularly perform certain tasks. Botnet relies heavily on C&C servers and provides low-latency communication [41]. The botsniffer, HTTP-based botnets can also be detected, and abnormal detection technologies can be applied to stop all bots. In addition, servers as channels are detected in similar types of behavior i.e., flexible in the substation of C&C server addresses. It also provides the information needed to detect hybrid botnet structures[42].

Mirai botnet detection using binary code is a classic method developed by Lee Jun-soo. First, the binary code of the malicious code is analyzed and the structure of the binary file is determined before it is used [43]. For example, the binary code is described as a portable file structure that runs in a Windows environment. PE format should be implemented based on the nature of the detection, i.e., compatibility in various operating systems (OS) to facilitate detection. So, it was named the “Easy Movement” format. This is because this format is a file format for executable files used in Windows, such as Dynamic-Link Library (DLL), Object Code (object code) and FON-type font files. Similar to PE, there are executable files and connection formats (ELF) and Mac OS X formats, which are x86-based UNIX and UNIX systems[44].

Honeypot is a system that is installed with a purpose to detect abnormal access and it also serves to track down attackers and gather information. To deceive an attacker, a trap is created, as if the attacker had infiltrated into a normal system. And then a bot is caught and analyzed. Based on the analysis results, software disguised as a bot is created and traffic exchanged with the software is analyzed to find a botmaster or botnet. One advantage is that botnets can be detected at a high detection rate without existing knowledge [45].

Kippo: Medium Interaction SSH Honeypot can add or delete files through a fake file system and save files separately on the host system when downloading them. It also configures simple command execution and setup files. It can record Burt Force attacks and malicious user behavior [46].

Cowry: Similar to Kippo’s Honeypot, Telnet service, SFTP and SCP are added to allow the collection of uploaded files via Telnet and SSH attack houses, SFTP, and SCP [47].

Dionaea: uses libemu to detect shellcodes with Python honeypot. It supports IPv6 and TLS to collect malicious codes by providing vulnerabilities to malicious users[48].

Telnet-IoT-Honipat: Honeypot for collecting Telnet attacks is written in Python and mainly collects malicious botnet codes. Then the collected malicious code to Virus Total is uploaded [49].

IoTPOT: jointly developed by German and Japanese universities. It consists of Honeypot and Sandboxes against Telnet attacks. It provides Telnet service for various IoT devices and consists of two parts. The front end provides a low level of interaction and the back end provides a high level of interaction through an environment called the IoTBox. The IoTBoX integrates eight CPU architectures, including Mips and ARM, to provide a variety of virtual environments commonly used in the systems. However, the use is limited until it is released, not for open-source [50]

C. Potential of Botnet Attack to Smart Factories and the Honeypot Approach

The network environment in the smart factory will require both the new honeycomb system and the IDS method to be deployed if the honeypot detection system is applied. It also designs scalable honeypot clients that perform and interact efficiently. The purpose of their study will be to increase capture capacity and establish in-depth analysis. The autonomous version of the honeypot implementation was addressed by [51].

There is a high possibility that some major attacks will target smart manufacturers, especially those smart factories using IoT technologies. Typical IoT nodes can be directly attacked by individuals within the radio range, such as relatively low-power processors and wireless networking functions. This undermines the security model in which borders and devices (e.g., firewalls and intrusion detection systems) are defined. Instead, each device needs to be self-secured, at least in part and this is a task that becomes more difficult due to the reduced processing power of typical IoT nodes. Normally, manufacturers may not be aware of large-scale attacks that do not adequately secure individual devices.

Botnet Mirai [35], the biggest cause of distributed denial-of-service attacks, is the best example and hypothesis of the failure of the smart manufacturing defense. Operation of the Miribot Net allows Mirai to identify vulnerable IoT devices that can be accessed using the Internet.

Once these devices are identified, an attack is carried out using a simple pre-attack (composed of factory default user names and passwords belonging to users such as administrators) [5]. The boat sees the identified IP address of the vulnerable device (1), reposts it to the server (2) and then deploys the vulnerable device to the load server (3). The load server loads the malware associated with the operating system (4). When the device executes malicious code, bot (5) appears and receives new commands from the command and control server (C&C server) (6). Mirai also has the ability to eradicate other malware processes by closing all processes using SSH, Telnet and HTTP ports, searching for and removing other botnet processes that may be running on the device. The C&C server communicates with the report server to keep an eye on the infected devices (7). The boat carries out distributed denial-of-service attacks on targets (8), continues to scan and infect new victims and receives further instructions from C&C Server (9) [12].

It will take advantage of the lack of security in IoT devices and carries out a successful approach that can cause production downtime and negatively affect the company’s reputation due to equipment failure or attacks on other systems.

The operation of smart factories on the IoT platform reveals some features. It became vulnerable to the Botnet attack.

Web interface insecure: Loss or damage to data can be caused by unsafe web interfaces [52].

Lack of accountability or denial of access can lead to a complete device takeover. (security impact) [53].

Lack of transport encryption: Depending on the data exposed, user accounts or devices lose data or become completely corrupted (e.g., sending unencrypted credentials and data) [52].

Privacy concerns: Data collection of the smart factories, along with a lack of data protection, can lead to a compromise of a user's personal data.

The threat of botnet attacks on smart factories can possibly be encountered by applying honeypot as a detection method. This is because the honeypot approach presents the following special features: Able to capture attack into log files. And log analysis allows for details about exploitation and attack patterns to be found. Able to capture anything that interacts with them, including tools or tactics which have not seen before (0-Days).

Only deal with incoming malicious traffic. So, the collected information is smaller and has a higher value. Fewer false positives compared to other security solutions since only attack traffic is detected (no legitimate traffic). Require minimal resources with no additional budget for the companies. Simple to understand, to configure and to install. Do not require known attack signatures (unlike IDS). Able to detect an IPv6-based attack the same way it does with an IPv4 attack. Besides the good features, limitations of the honeypot detection method are also presented: it suffers from fingerprinting: it is easy for an experienced hacker to differentiate between a fake system and a real one. The risk of being hijacked by the honeypot system (if not adequately designed) may be used to attack other systems. Limited visions: a honeypot can only capture data involving directly interacting hacking activity.

Industrial manufacturers need to maximize production and plant management efficiency. It is important to understand and resolve issues that occur in the manufacturing process. Finding security-related issues is critical in running the operating system smoothly. In addition, concerning the management of smart factories involving the use of various IoT equipment, the recent threat of botnet has become a problem because it has caused considerable damage to production. Since botnet attacks are becoming increasingly more serious, it is an urgent matter for producers to detect botnet. Problems with data transfer between botnets, sensors, CCTVs, PLC equipment, and main database servers may be affected by data leakage in the smart factory network which resulted in data updates being exploited by unauthorized users who may cause unexpected impact on smart factory operations. Indeed, real-time detection is of paramount importance, especially in smart manufacturing environments [11]. Various methods of detecting botnet are compared, as shown in Table I. Honeypot and honeynet can respond to attacks in real-time and attract attackers to deceptive assets rather than real assets. For binary, anomaly and C&C detection methods, reaction to real-time is slower than honeypot and honeynet method [43], [57], [58]. Although binary detection is simple in structure, the detection processing is too slow for smart manufacturing environments that seek real-time detection. In terms of cost-effectiveness, honeypot has an advantage in being capable of responding to attack in real-time at relatively low cost for construction and management. It is suitable for smart manufacturing environment [55], [56] However, processing botnet information by the honeypot is slow for analysis. It results in a decrease in accuracy and processing speed [59], [60]. Notably, an attempt of using machine learning techniques to combine with honeypot has not studied so far.

TABLE I. COMPARISON OF HONEYPOT VS. OTHER DETECTION METHOD

Division	**Honeypot**	**Honeynet**	**Binary detection**	**Anomaly detection**	**C&C detection**
Configuration form	One host	General host Security solution Honeypot Network	BINARY	Heuristic Rules[54]	SERVER
Advantages	High Efficiency of Collective Data Efficient Packet Data Processing Install, apply, operate, and manage [55]	Application flexibility Excellent data collection and warning Applicable to various systems and applications[56]	Suitable computer application Only 1 & 0 are being used, implementation becomes easy[43]	Systems have the capability to detect zero-day attacks as well [57]	It has the advantage of being able to detect and expand HTTP-based botnets.[58][41]
Disadvantages	Intrusion into the network Information analysis slow [59]	Difficult to set up and build [60]	Long-term processing [61]	Not simple structure high false-positive [62]	In the case of botnets with large delays, the detection rate drops, and the false detection rate increases.[63][42]
Research Gap	Botnet is highly efficient in collecting data and easy to build and manage detection However, information analysis is slow.	Data collection and alerting against botnet attacks are quick. It takes a lot of time to build a system for the first time.	The structure of the system is very simple, and the computer is highly recognizable. However, the process is too slow. So not suitable for smart factory environment	It is good for detecting attacks that attack vulnerabilities such as zero-day, but the program structure is complex, and the probability of failure is high.	There are many IoT devices used for the smart factory. If there is a high probability of a delay in IoT communication devices such as RFID, the detection rate and error rate will automatically increase. Not suitable for the smart factory environment.

TABLE II. HONEYPOT STAGE OF INTERACTION

Interaction stage	**Pros**	**Cons**	**Honeypot types**
Low	Ability to log huge amounts of attack data. No simulation needed for the actual OS used for interaction	Time-consuming . The highest probability of risk. Complex [70]	Honeynet [71]
Medium	Offer better. Simulated services. More difficult for attackers to identify [72]	Increasing subject to the security vulnerability Longer time for implementation and expertise required [73]	Kippo [74]
High	Easy to install and price effective. Low risk. Require little to no expertise [75]	Require to have a complete set of features Give limited information about the specific attacks [76].	Honeyd [77]

Honeypot is a program designed for cybersecurity to defend against virus attacks. Attackers are attracted by the exploits contained in the honey software to extract data from the network of an organization with the intention of causing malice. Frank Cohen was the designer of the first honeypot known as The Deception Toolkit. It was used to effectively against the automated attacks on a system. With a variety of vulnerabilities as a form of deception, attackers are lured to the system. This deception toolkit has an important feature to create alert for the administrator against the deceptions, given that information used to hack into the system was often under a particular service like sending an e-mail [64]. The invention of honeypot created a breakthrough in the field of security networks and computers. Due to this, honeypots have been widely used in the year of 2000s. Meanwhile, some computer programs. It can self-replicate were spread rapidly within the Internet network, causing an outbreak of worms. The spread of such computer worms poses a danger to network traffic and thus, increases the latency of the whole network. The idea of capturing these worms for analysis was not considered. Instead, the optimal solution to collect these worms was by trapping them into a honeypot [65].

A variety of honeypot solutions has been used within the organization or applied to specific industries and services. It was found that honeypots are categorized into two main dimensions: level of interaction with attackers and service provision, as shown in Table II. Analysis requires a variety of data to be collected, requiring a higher level of interaction, with reference to the research conducted by Ronald and Keshnee [66].

Some applications of honeypots in distinct research fields were studied. A good example of this is the design of a system called “Sweetbait”, using the honeypot approach for capturing fast worms with the purpose of automated analysis and signature generation. Continuous updates of signatures were sent to both network and host-based IDS/IPS all over the parts of the Internet [67].

Machine learning as an alternative solution to the conventional detection methods using in smart factories currently. So far, smart manufacturing has been using rule-based intrusion detection methods that use signal DB or SNORT to analyze security data. The detection method is mainly based on anomaly-based detection. It determines normal and abnormal conditions compared with normal network conditions.

This makes it difficult to respond quickly to unknown attacks or attacks on manufacturing IoT using a botnet. There is a hassle to manually set the rule for attack patterns. And it often results in poor efficiency of security personnel input. In addition, the rule-based methods, based on expert knowledge, have difficulties in flexibly responding to changes in the external attacks and maintain consistency in high detection performance. This is because such rule-based detection makes it impossible to detect intrusions outside of pre-designed rules and it can vary the quality of the rules depending on the experts and system environment. Therefore, research related to the development of the detection model using machine learning has recently been attempted to overcome the shortcomings of the rule-based detection methodology. Indeed, machine learning (ML) is a procedure that teaches computers or devices to perform automatic processes. Network machine learning will learn from the network environment for a period of time. In addition, ML is based on mathematical modeling. Thus, it is better than signature-based and rule-based detection methods, so it will be able to cope with the continuously advanced and sophisticated attacks [68][69].

D. Critical Literature Review

Taxonomies of the botnet detection and smart factories IoT security are illustrated in Fig. 6 and 7. Table IV gives a summary of the selected papers used in these reviewed papers. These studies were conducted with the effort to detect botnet, the honeypot approach to detect botnets and narrowly focused on botnet detection using honeypot for smart factories. They are then grouped in Fig. 8, 9 and 10 accordingly.

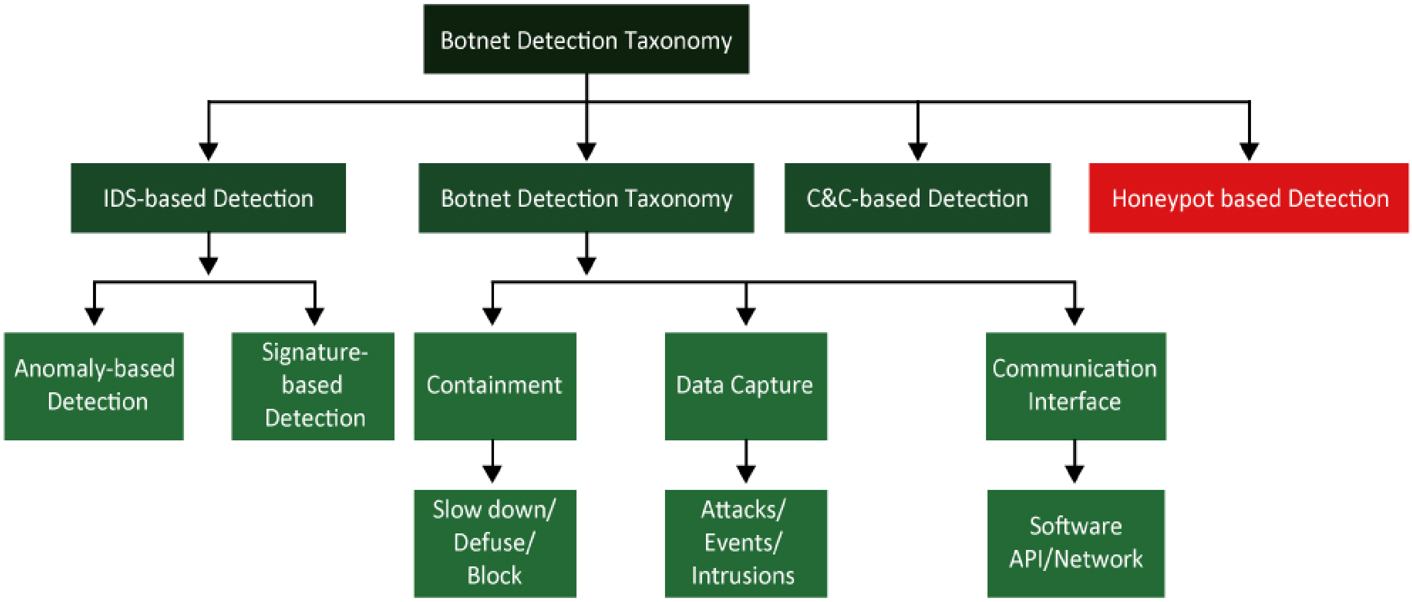


Fig. 6. Botnet Detection Taxonomy.

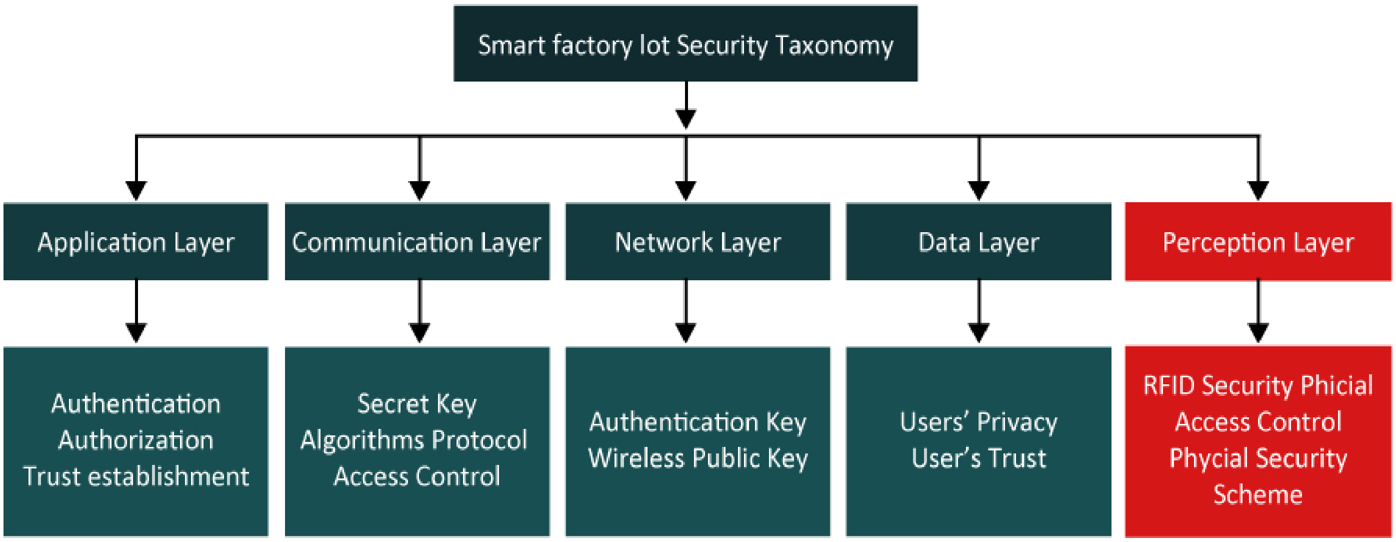

Smart factory Iot Security Taxonomy
Application Layer
Communication Layer
Network Layer
Data Layer
Perception Layer
Authentication Authorization Trust establishment
Secret Key Algorithms Protocol Access Control
Authentication Key Wireless Public Key
Users' Privacy User's Trust
RFID Security Phicial Access Control Phycial Security Scheme

Fig. 7. Smart Factory IoT Security Taxonomy.

This section provides a review of the literature related to work conducted by the researcher in smart factories detection and machine learning areas. The task scheduling algorithm is detailed in the two taxonomies to clearly understand and classify the basic approaches currently in use. Recently, there has been a significant increase in Botnet threats. In particular, it can be seen that botnet's attack on IoT platform has increased dramatically. Table III summarizes some important studies in using the Honeypot approach for detecting botnet in smart factories. A study in IoT botnet detection suggested that it is easy to monitor, if the IoT devices are infected through web services [79]. Some restrictions have pointed out that monitoring algorithms for IoT devices are simple to implement and it is scalable enough for smart factories using IoT equipment [79]. The capacity for the IoT devices has clear limits. This approach was first designed with a hypothesis that botnet contacts IoT devices were used to invent a detection model based on the binary. The botnet detection uses a machine learning approach that shows high accuracy of 99.94%. The combination of flow-based with graph-based detection and machine warning has a high accuracy of detecting a botnet attack as an advantage. The disadvantage of this approach is that it is harder to detect quickly in the randomized number of packets. Thus, the appropriation of applying this approach for smart manufacturing needs more research in real-time and it is time-consuming [80].

TABLE IV. A COMPARATIVE SUMMARY OF STUDIES IN BOTNET DETECTION FOR SMART FACTORIES USING THE HONEYPOT APPROACH

Ref.	Approaches	Strengths	Weaknesses	Research gap
[78]	Smart factory detection using machine learning	Cost reduction	Low detection rate, high complexity and uncertainty	Intrusion detection systems deployed in this study are implemented by deep neural networks, requiring intrusion detection systems through convolutional neural network, recurrent neural network, deep brief network and deep q-networks applied to various systems in this study.
[79]	Botnet , IoT botnet	Web service is available for easy monitoring of IoT device health and is useful for smart factories with many IoT devices.	Limited capacity.	Develop a strategy to simplify and optimize the binaries that implement this security technology to broaden the application of the results in this study.
[80]	Botnet detection using Machine learning	It shows the effect of bookmarks. Hybrid analysis of flow-based and graph-based traffic behavior achieves 99.94% detection accuracy, surpassing individual detectors	Randomly specify the number of bytes per packet and the number of packets per flow so that they are not detected.so flow-based detectors are not easy to apply quickly.	Improved detection results show new botnet detection through effective graph-based features and botmark effects.
[80]	Detection using IoT Honey pot	The speed of information gathering is rapid. Because it implements only part of the system, it consumes fewer resources.	Unnecessary data piles up	To support high protocols, the company plans to expand the IoT and expand sandboxes with features that can further activate the architecture and environment commonly used in IoT devices.
[68]	Detection using Honey pot Machine learning	It has developed a honeypot-based solution for botnet detection using a machine-learning detection framework. The use of honeypot ensures logging of newly released malware functions.	The function varies greatly depending on the difference in the system performance.	The honeypot approach should be expanded. Cloud servers should also be employed to handle IoT devices with minimal resources.

For smart factories, botnet detection using honeypot integrated with IoT (IoT honeypot) was studied [81]. There is a stochastic basis compared to the machine learning approach with superstitious running. Although the IoT honeypot has stopped scalability by simply applying it to sandboxes IoT, it aims to apply for common expansion in more situations and environments [14.] In the detection system using honey pot machine botnet, the learning logging for detection and tracking are so accurate. The system, in accordance with the system different but most standard equipment, is suitable for performance smart factories. Hence, it is likely to be adopted in the future. The cloud server approach uses the proposal [68]. As for IDS used by the smart factory, although the machine learning approach can reduce costs, significant imperfections such as low detection rates, highly complex and unsustainable systems were observed [78]. Three studies in IDS, IoT botnet and honeypot machine learning showed some application results for smart factories. Such solutions are possible to trace through logging at low cost and are most cost-saving for the IoT devices. Thus, botnet detection for smart factories using machine learning based on honey pot detection needs more in-depth research.

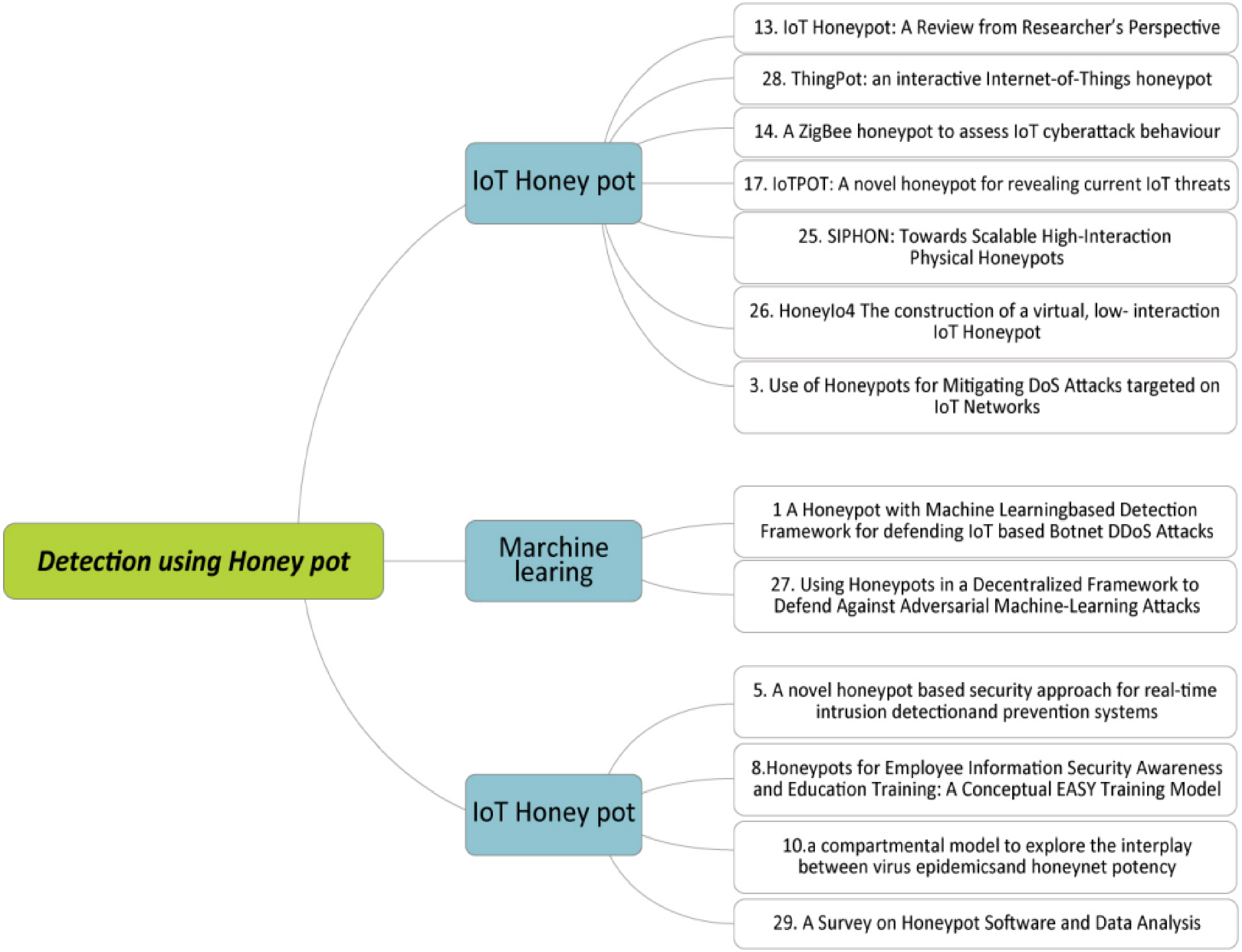

Detection using Honey pot
IoT Honey pot
13. IoT Honeypot: A Review from Researcher's Perspective
28. ThingPot: an interactive Internet-of-Things honeypot
14. A ZigBee honeypot to assess IoT cyberattack behaviour
17. IoTPOT: A novel honeypot for revealing current IoT threats
25. SIPHON: Towards Scalable High-Interaction Physical Honeypots
26. Honeylo4 The construction of a virtual, low- interaction IoT Honeypot
3. Use of Honeypots for Mitigating DoS Attacks targeted on IoT Networks
Marchine learing
1 A Honeypot with Machine Learningbased Detection Framework for defending IoT based Botnet DDoS Attacks
27. Using Honeypots in a Decentralized Framework to Defend Against Adversarial Machine-Learning Attacks
IoT Honey pot
5. A novel honeypot based security approach for real-time intrusion detectionand prevention systems
8.Honeypots for Employee Information Security Awareness and Education Training: A Conceptual EASY Training Model
10.a compartmental model to explore the interplay between virus epidemicsand honeynet potency
29. A Survey on Honeypot Software and Data Analysis

Fig. 8. Grouping of Honey Pot based Botnet Detection Model for Smart Factory (A).

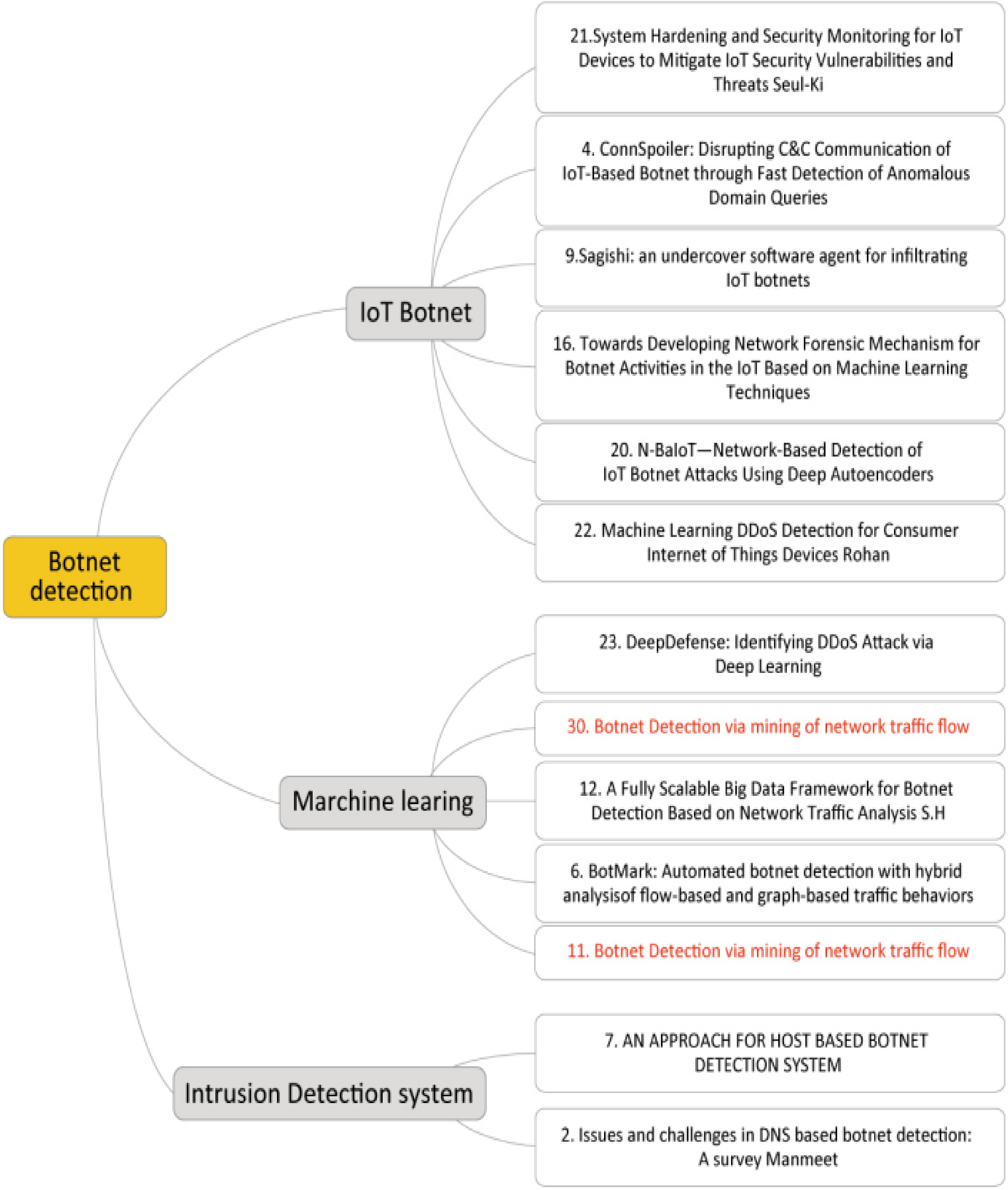

Botnet detection
IoT Botnet
21.System Hardening and Security Monitoring for IoT Devices to Mitigate IoT Security Vulnerabilities and Threats Seul-Ki
4. ConnSpoiler: Disrupting C&C Communication of IoT-Based Botnet through Fast Detection of Anomalous Domain Queries
9.Sagishi: an undercover software agent for infiltrating IoT botnets
16. Towards Developing Network Forensic Mechanism for Botnet Activities in the IoT Based on Machine Learning Techniques
20. N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders
22. Machine Learning DDoS Detection for Consumer Internet of Things Devices Rohan
Marchine learing
23. DeepDefense: Identifying DDoS Attack via Deep Learning
30. Botnet Detection via mining of network traffic flow
12. A Fully Scalable Big Data Framework for Botnet Detection Based on Network Traffic Analysis S.H
6. BotMark: Automated botnet detection with hybrid analysisof flow-based and graph-based traffic behaviors
11. Botnet Detection via mining of network traffic flow
Intrusion Detection system
7. AN APPROACH FOR HOST BASED BOTNET DETECTION SYSTEM
2. Issues and challenges in DNS based botnet detection: A survey Manmeet

Fig. 9. Grouping of Honey Pot based Botnet Detection Model for Smart Factory (B).

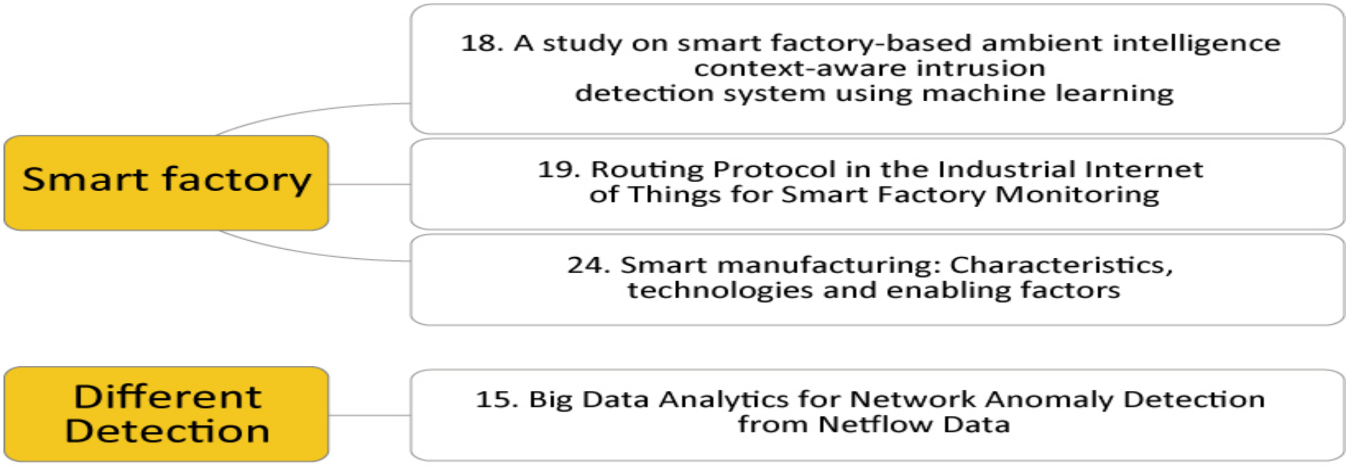

Smart factory
18. A study on smart factory-based ambient intelligence context-aware intrusion detection system using machine learning
19. Routing Protocol in the Industrial Internet of Things for Smart Factory Monitoring
24. Smart manufacturing: Characteristics, technologies and enabling factors
Different Detection
15. Big Data Analytics for Network Anomaly Detection from Netflow Data

Fig. 10. Grouping of Honey Pot based Botnet Detection Model for Smart Factory (C).

TABLE V. SUMMARY OF LITERATURE REVIEWED

No	Contribution	Research gap
[68]	. It has developed a honeypot-based DDoS detection solution that utilizes a real-time machine learning detection framework. Using beehives can ensure logging of newly released malware functions that can be utilized as ML-based detection.	IoT honeypot structure is heterogeneous because IoT device types are different. But the original Honeypot structure is similar. So there is a big difference between traditional honey and IoT honeypot
[82]	Propose an in-depth category for DNS techniques within a category.	While various papers talk about DNS-based botnet detection, each detection technique does not provide efficient classification and does not think about parameters.
[83]	Implemented in real-time environments with a variety of microcontrollers that interface with central servers. The idea of deploying a honeypot to handle DoS attacks can also be extended by deploying a honeypot system that can handle DDoS attacks using botnets.	Suggest a Honeypot Model to Mitigate DoS Attacks Started with IoT Devices
[84]	This paper provides a lightweight sensing system. NXDomain's ConSpoiler Works in Nuclear IoT-based botnets in response mode, NXDomain's compiler works the weeks. Also, on a DNS train collected from two other large ISP networks, Conspoiler can have this peculiarity that computers have been developed by the city.	Information gathered from ISP networks requires an identifiable assessment of devices infected with botnets.
[85]	The system developed is a type of honeypot-based IDPS that allows real-time animation of server network traffic. Zero-day attacks can be detected more effectively than other IDS. And you can save resources by reducing the amount of data on IDS.	A good hybrid honey pot is reminiscent of high interaction and low interaction. Perform Honeypot. To effectively analyze data in real-time, The developed Honeypot server application works by combining with IDS.
[80]	A variety of experiments outstrip flow-based or graph-based detection by achieving 99.94% accuracy.	A combination of flow-based network traffic behavior and graph-based and network traffic behavior for botnet detection suggests an automatic model of botmark. Network anomalies detect and detect botnets through flow-based traffic analysis using optical communication patterns in traditional methods. However, simple flow-based traffic analysis and graph-based analysis increasingly increase the failure rate due to the evolution and precision of botnet attacks.
[86]	The system host interface is used as an aberrant detection technique by analyzing the traffic of genetic algorithm variation. The experimental results analyze each algorithm and show that it is relevant. Future work provides additional detection techniques as an extension of system functions. Anomaly signature-based techniques such as the addition of data integrity analysis are integrated to enable rapid delivery.	The host-based chip-in-detection system approach is based on the modification of the algorithm It can be used in the event of an attack. Based on the approach of anomaly detection. The cause of the current system's malicious code is the use of bots. This approach analyses the cause of botnet attacks.
[74]	By placing Honey pot in a company or institution in the future, employees are aware of the importance of security awareness and increase the risk of detection by increasing the security culture within the organization.	Analyze the honeypot data to analyze how employee information security awareness and systems can be utilized by reviewing cybercriminals (including vendors and malicious insiders).
[87]	In this work, the concept of active honeypots was introduced to	Mapping and classifying boats that form part of the existing IoT botnet

	mimic real bots so that information can be leaked, usually only to infect machines. It also described how active honeypots are integrated into the proposed software architecture that allows users to receive malware samples and extract botnet functions for the purpose of penetrating botnets.	and how it spreads through the Internet are difficult. Forensic had previously sought to solve the problem by installing, reporting and securing remote IoT devices that had been changed to boats. The activity was based on honey pot, which can only “detect” the Internet to guide and classify the IoT bot according to its behaviour.
[71]	Here are some ideas for improving the design of honeynet and for the constitution of the spread of the virus. Because the honeynet has a low power index, it is easy to obtain virus samples and it is easy to provide a number of numerical examples for theoretical analysis.	Within a systematic framework, honeynet’s potency was not evaluated theoretically.
[88]	The existing detection system shows limits in preventing botnet attacks on IoT. Because botnet keeps evolving, but the study shows that ML technology has advanced in detecting specific threats to IoT networks.	One method of detection is botnet processing power, amount of energy consumption, IoT environment requirements, etc. Often it does not meet the potential to address security threats. So the IoT network and Ddos attacks botnet often cause major security problems.
[89]	Extensible botnet detection methods can be integrated by providing computing infrastructure for building big data frameworks using crowd service providers in large networks. In addition, skilled human resources include the cost of ownership because they build a framework	The scalability problems of botnet detection systems arise from a variety of problems, such as bottlenecks in the detection process, storage, data collection and analysis.
[67]	The design of IoT honeypots can be expanded to provide intelligent responses based on interaction in the combination of login/passwords, distribution of attacks, types of devices being attacked and IP addresses by country of distribution. It can be used as an additional review of IoT honeypots for IoT devices.	Implementing a strong security mechanism leaves little room for security implementation of IoT devices and limited hardware functions, making it very difficult to implement.
[81]	Zigby is one of the wireless technologies used for IoT. Large global malware uses SSH as an entry point to continue pre- and violent attacks. Therefore, deploying this Honeypot in SSH to collect large data sets to identify awareness and interest in the ZigBee network makes it easy to collect and recognize automated attack types.	There is a lot of inconvenience by limiting attack statistics to the physical scope of zigbee communications to collect cyber attackers’ paddles.
[40]	The accuracy of high anomaly detection should provide high-quality service and communication, even as the complexity of the attack and analysis processes increases. Anomaly attacks and singularities are naturally rare. Propose innovative algorithms achieved in future studies with more data and anomalies.	Big Data Anomalies Detection Security is key to continuing and long-term cyber-attacks. With constant changes in the distribution of network data, detection becomes more difficult
[90]	In DT, ANN, NB and AN machines, the machine learning technique proved to be superior to other techniques in false alarm rate and accuracy, showing excellent ability in identifying and investigating botnet without errors.	To used ML techniques to investigate botnet activity and apply detection to bonnet attacks, but there were many problems and high false alarm rates to make perfect detection. So I found that there are many problems in training and verification research of the detection model of ML technique.
[91]	Five malign program families were identified, all of which are actively used in DDoS attacks.	The botnet has done a lot of damage to devices powered by Microsoft and Sony’s IoT through spam e-mails and we learned that the main target point of the attack came from IoT devices. IoTPop analyses the samples of malware captured by honeypots and analyses them on Telnet-based This research proposes this method because it will be easy to track and analyze attacks.
[78]	Machine learning-based intrusion detection systems have reduced the frequency of incorrect replacement of existing devices, resulting in significant savings. It showed 33% to 1.33% process performance and 29% to 1.29% abnormalities but showed an effective scoring architecture.	In a smart manufacturing environment, real-time detection is important enough to have a significant impact. But there is a limit to real-time detection. The efficiency of IoT development and application should be increased within smart manufacturing.
[92]	This paper aimed at proposing and analysing the efficient framework of industrial IoT and providing the latest approach to industrial applications. This paper also dealt with the adoption of Laura.	Industrial IoT requires more thorough security to prevent data leakage, transmission errors and data injection due to communication between hundreds of devices. Access should also be naturally controlled by IoT devices. Real-time industrial IoT device security needs to be monitored.
[93]	While the IoT device, part of the botnet, was launched in a damaged state, it demonstrated the ability to detect the exact and immediate manner of attack that we proposed.	A new network-based abnormality detection method called IoT N-BaIoT, which detects abnormal network traffic by taking snapshots of network operations using deep automatic encoders from damaged IoT devices, is needed. This is because the number of IoT-based botnet attacks should be rapidly increasing and the threat of botnet attacks should be mitigated by detecting IoT-based attacks that last for hours and milliseconds.
[94]	The proposed technique is expected to be useful in managing	The weak aspects of IoT device security through real-time security

	numerous IoT devices such as the smart factory. The IoT market is rapidly changing and IoT devices are widely adopted in various fields. It suggested a system and operation method. It can easily apply security functions to IoT devices. This study further validated the usefulness of the proposed technique by developing a prototype.	monitoring In order to minimize the threat of attack, there is a threat of secondary attacks or malware attacks on damaged IoT devices. So I propose to build various security functions and strengthen the system.
[7]	The network traffic pattern of IoT devices classifies Ddos detection general and Ddos attack traffic, thus using a limited set of functions that are important for real-time intermediate box layout. So the study of machine learning at the packet level shows that it depends on the hypothesis.	To remain limited due to memory constraints in IoT devices, caching adds to the delay time and complexity. Therefore, the optimal algorithm should store flow information only for a short period.
[95]	Reducing the error rate from 7.517% to 2.103%, the in-depth learning approach was automatically extracted from the high level of characteristics and then patterned from the sequence of network traffic, making it easier to track network attacks. It has shown enough that the new model is superior to the existing one.	Traditional detection solutions have failed to defend against fatal threats and have shown limitations in monitoring network traffic based on statistical variances against rapidly growing DDoS attacks. However, if performance identification based on machine learning is improved, there is also a potential for the development of statistical characteristics.
[4]	Smart manufacturing ontology feature that can be used to provide a platform for active technology and factors identified, discussions and clusters. Overall, situational awareness, modularity, bilateral, inter-operability and configuration of five characteristics are considered.	Intelligent manufacturing, many items that appear to be indistinct redundancy to smart factories, one of the most advanced manufacturing, need to be established as a foundation for manufacturing ontology.
[96]	Potential weaknesses in IoT devices and Internet attacks have always been threats to IoT equipment. SIPHON can expose IoT devices to the Internet to enable clear monitoring of test beds and enhance honeypot's reality. The limitations of simple security test mechanisms can be overcome.	An inefficient way for an attacker to move benefits before solving the vulnerability problem of finding vulnerabilities in IoT devices. In traditional IT security, we understood accounts that were critical to the dynamic threat environment without hacking and potentially conducted honeypot attempts to establish unauthorized connections. There were attacks in realistic ways, such as log-in shellfish.
[97]	It detects Honeypot with IoT devices and provides detailed information to attackers. HoneyIo4 can run on both CLI and GUI and for both experienced and inexperienced users. Although its initial performance is limited, it succeeded in detecting IoT OS. Honey can be improved by adding more features to this basic core.	Honeypot usually deceives the system in a limited way. They also have less risk to the network if the honeypot is damaged, but the information collected for attackers or attackers is also very limited. Preventive, detection and response mechanisms should be provided to facilitate maintenance and protection on the organization's network.
[98]	A decentralized defense framework that prevents opponents from degrading the learning model, suggesting a network of high-interaction honeypots (HIHP). To achieve a goal by preventing an attacker from learning the label correctly and by approximating the structure of a black box system. Attracting attackers, using adobe honey to generate calculations that are not feasible for the enemy, for the Decoy model and for the enemy.	Limited access to input and output labels of data can be used to confuse input learning. However, the market is increasingly in demand for machine learning services. Naturally, there is a possibility of exposure to a variety of complaints by increasing threats.
[[99]	The IoT platform and device attackers are caught. In particular, five types of attacks have been discovered.	IoT device signals clearly sent to IoT for the exploitation of security vulnerability by those who want it. So it may be possible to secure IoT, but it is important to identify an attack strategy.
[100]	Honey pot that analyses the data methodology, the ethical and legal issues are discussed.	The survey provides a broad overview of honeybees. This includes not only honeypot software but also methodologies for analyzing honeypot data.
[101]	Apply deep learning optimization to handle the high false cost of the algorithm, integrating high-level new detection model. So by classifying random filters, effectively achieve botnet defense.	Many researchers have many botnet detection models in the past, but most of them have not found botnets these days with both high probability and good memory and time efficiency.

Table IV elaborated literature review critically to the botnet issues with smart factories, and other related domains. This issue equally impedes to the smart homes as well [102]. In addition, studies elaborated that the phishing works and supports for botnet attacks [103]. Further, these botnet attacks help to the attacker by providing them ground, where they can launch different attacks and make possible intrusion of the network [104], and later these attacks could help the criminals [105-106] for their different activities.

III. Conclusion

Throughout this review of literature, the field of IoT-based smart factory using honeypot approach to detect botnet is a potential area for the research to explore. Conventionally, smart factories have been using three methods i.e. signature-based, rule-based and anomaly-based for detection.

However, these conventional methods were recognized to have a limitation in the responding time which is desired to be quicker in detection. It took a long time to detect the botnet, exposing the vulnerability of smart factory. Honeypot is an approach that has been examined for its effectiveness to trap botnet in some studies. The honeypot approach can overcome the limitation of the conventional methods in terms of quick detection, while the botnet is easy to spread in the IoT based environment.

So far, there is a scarcity of studies in applying the honeypot approach for botnet detection designed for smart factories. However, this paper suggests the possibility. If honeypot botnet detection is applied in Smart factory IoT environment, it can improve the productivity of Smart factory and fasten the production time.

IV. Future Work

Future research should look into developing honeypot models and algorithms that can be applied in smart factory IoT environment. And if the failure rate and detection time can be reduced through the metadata score, the model performance will be improved dramatically.

Acknowledgment

I would like to thank my supervisors, Associate Prof. Dr Azween Abdullah & Dr. Noor Zaman Jhanjhi, for their patient guidance, encouragement and continuous support provided throughout my postgraduate studies. I have been extremely fortunate to have supervisors, who cared so much about my work, and professional growth.

References

[1] L. Barreto, A. Amaral, and T. Pereira, “Industry 4.0 implications in logistics: an overview,” Procedia Manuf., vol. 13, pp. 1245–1252, 2017.

[2] B. Huang, W. Wang, S. Ren, R. Y. Zhong, and J. Jiang, “A proactive task dispatching method based on future bottleneck prediction for the smart factory,” Int. J. Comput. Integr. Manuf., vol. 32, no. 3, pp. 278–293, 2019.

[3] I. Mistry, S. Tanwar, S. Tyagi, and N. Kumar, “Blockchain for 5G-enabled IoT for industrial automation: A systematic review, solutions, and challenges,” Mech. Syst. Signal Process., vol. 135, p. 106382, 2020.

[4] S. Mittal, M. A. Khan, D. Romero, and T. Wuest, “Smart manufacturing: Characteristics, technologies and enabling factors,” Proc. Inst. Mech. Eng. Part B J. Eng. Manuf., vol. 233, no. 5, pp. 1342–1361, 2019.

[5] R. A. Rojas and E. Rauch, “From a literature review to a conceptual framework of enablers for smart manufacturing control,” Int. J. Adv. Manuf. Technol., vol. 104, no. 1–4, pp. 517–533, 2019.

[6] M. S. Smith, “Protecting Privacy in an IoT-Connected World.,” Inf. Manag. J., vol. 49, no. 6, pp. 36–39, 2015.

[7] R. Doshi, N. Apthorpe, and N. Feamster, “Machine learning DDoS detection for consumer internet of things devices,” Proc. - 2018 IEEE Symp. Secur. Priv. Work. SPW 2018, no. Ml, pp. 29–35, 2018.

[8] A. D. Manyika James, Chui Michael, Bisson Peter, Woetzel Jonathan, Dobbs Richard, Bughin Jacques, “Unlocking the potential of the Internet of Things | McKinsey &amp; Company,” McKinsey, pp. 1–4, 2015.

[9] E. (TOPICAL C. Casalinuovo, “Thematic Investment Opportunity – Internet of Things,” no. March, pp. 3–6, 2019.

[10] M. Ozcelik, N. Chalabianloo, and G. Gur, “Software-Defined Edge Defense Against IoT-Based DDoS,” IEEE CIT 2017 - 17th IEEE Int. Conf. Comput. Inf. Technol., pp. 308–313, 2017.

[11] M. A. Rajab, “My Botnet is Bigger than Yours (Maybe, Better than Yours) : why size estimates remain challenging,” HotBots’07 Proc. first Conf. First Work. Hot Top. Underst. Botnets, no. USENIX Association Berkeley, CA, USA ©2007, pp. 5–5, 2007.

[12] N. Tuptuk and S. Hailes, “Security of smart manufacturing systems,” J. Manuf. Syst., vol. 47, no. November 2017, pp. 93–106, 2018.

[13] Constantinos Kolias, Georgios Kambourakis, Angelos Stavrou, and Jeffrey Voas, “DDoS in the IoT: Mirai and Other Botnets,” Computer (Long. Beach. Calif)., p. 79, 2017.

[14] J. Wan et al., “Software-Defined Industrial Internet of Things in the Context of Industry 4 . 0,” vol. 16, no. 20, pp. 7373–7380, 2016.

[15] N. E. Vaskenly and M. Dhanya, “Smart factories: An Indian scenario,” Int. J. Pure Appl. Math., vol. 118, no. Special Issue 9, pp. 505–509, 2018.

[16] F. Galati and B. Bigliardi, “Computers in Industry Industry 4 . 0 : Emerging themes and future research avenues using a text mining approach,” Comput. Ind., vol. 109, pp. 100–113, 2019.

[17] Z. Zhang, Y. Zhang, J. Lu, X. Xu, F. Gao, and G. Xiao, “CMfgIA : a cloud manufacturing application mode for industry alliance,” 2018.

[18] B. Chen, J. Wan, L. Shu, P. Li, M. Mukherjee, and B. Yin, “Smart Factory of Industry 4.0: Key Technologies, Application Case, and Challenges,” IEEE Access, vol. 6, no. c, pp. 6505–6519, 2017.

[19] X. Li, D. Li, J. Wan, C. Liu, and M. Imran, “Adaptive Transmission Optimization in SDN-Based Industrial Internet of Things With Edge Computing,” vol. 5, no. 3, pp. 1351–1360, 2018.

[20] M. Ghobakhloo, “The future of manufacturing industry: a strategic roadmap toward Industry 4.0,” J. Manuf. Technol. Manag., vol. 29, no. 6, pp. 910–936, 2018.

[21] M. Salhaoui, A. Guerrero-González, M. Arioua, F. J. Ortiz, A. El Oualkadi, and C. L. Torregrosa, “Smart industrial iot monitoring and control system based on UAV and cloud computing applied to a concrete plant,” Sensors (Switzerland), vol. 19, no. 15, 2019.

[22] Y. Liu, Y. Peng, B. Wang, S. Yao, and Z. Liu, “Review on cyber-physical systems,” IEEE/CAA J. Autom. Sin., vol. 4, no. 1, pp. 27–40, 2017.

[23] J. Bourgeois et al., “Programmable matter as a cyber-physical conjugation,” 2016 IEEE Int. Conf. Syst. Man, Cybern. SMC 2016 - Conf. Proc., pp. 2942–2947, 2017.

[24] A. Valente, S. Baraldo, and E. Carpanzano, “Smooth trajectory generation for industrial robots performing high precision assembly processes,” CIRP Ann. - Manuf. Technol., vol. 66, no. 1, pp. 17–20, 2017.

[25] Nguyen, “Study on realtime control system in IoT based smart factory Interference awareness, architectural elements, and its application,” pp. 1–4, 2017.

[26] J. Feng, F. Li, C. Xu, and R. Y. Zhong, “Data-Driven Analysis for RFID-Enabled Smart Factory : A Case Study,” IEEE Trans. Syst. Man, Cybern. Syst., vol. PP, pp. 1–8, 2018.

[27] S. Lu, C. Xu, R. Y. Zhong, and L. Wang, “A RFID-enabled positioning system in automated guided vehicle for smart factories,” J. Manuf. Syst., vol. 44, pp. 179–190, 2017.

[28] B. Hameed, F. Rashid, F. Dürr, and K. Rothermel, “Self-calibration of RFID reader probabilities in a smart real-time factory,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7319 LNCS, pp. 253–270, 2012.

[29] L. Shen, Q. Zhang, J. Pang, H. Xu, and P. Li, “PRDL: Relative Localization Method of RFID Tags via Phase and RSSI Based on Deep Learning,” IEEE Access, vol. 7, pp. 20249–20261, 2019.

[30] K. H. Wang, C. M. Chen, W. Fang, and T. Y. Wu, “On the security of a new ultra-lightweight authentication protocol in IoT environment for RFID tags,” J. Supercomput., vol. 74, no. 1, pp. 65–70, 2018.

[31] S. U. Rehman, R. Liu, H. Zhang, G. Liang, Y. Fu, and A. Qayoom, “Localization of moving objects based on RFID tag array and laser ranging information,” Electron., vol. 8, no. 8, 2019.

[32] J. M. Ceron, K. Steding-Jessen, C. Hoepers, L. Z. Granville, and C. B. Margi, “Improving iot botnet investigation using an adaptive network layer,” Sensors (Switzerland), vol. 19, no. 3, pp. 1–16, 2019.

[33] S. Maeda, A. Kanai, S. Tanimoto, T. Hatashima, and K. Ohkubo, “A Botnet Detection Method on SDN using Deep Learning,” 2019 IEEE Int. Conf. Consum. Electron. ICCE 2019, pp. 1–6, 2019.

[34] S. Baruah, “Botnet Detection : Analysis of Various Techniques Sangita Baruah a,” Int. J. Comput. Intell. IoT Proc., vol. 2, no. 2, pp. 461–467, 2019.

[35] C. V. F. Jr, “Mirai Bot Scanner Summation Prototype,” 2019.

[36] J. Margolis, T. T. Oh, S. Jadhav, Y. H. Kim, and J. N. Kim, “An In-Depth Analysis of the Mirai Botnet,” Proc. - 2017 Int. Conf. Softw. Secur. Assur. ICSSA 2017, pp. 6–12, 2018.

[37] F. F. Giordani, “Consiglio nazionale delle ricerche,” How long does it Tak. before a new Internet node is contacted very first time?, vol. 3, no. 6, p. 413, 2018.

[38] H. Nguyen, Q. Ngo, D. Nguyen, and V. Le, "PSI-rooted subgraph: A novel feature for IoT botnet detection using classifier algorithms Huy-Trung," ICT Express, 2020.

[39] A. Mathematics, "DETECTION AND ERADICATION OF BOTNETS IN ONLINE BANKING," vol. 116, no. 10, pp. 73–77, 2017.

[40] C. H. Yeung, "Big Data Analytics for Network Anomaly Detection from Netflow Data Duygu," Int. J. Androl., 2017.

[41] R. S. and A. Thakral, A Review of Various Mechanisms for Botnets Detection. Springer, Singapore, 2018.

[42] S. Mulik and A. Patil, "Botnet Detection using Traffic Analysis and Defenses," vol. 6, no. 2, pp. 108–115, 2019.

[43] A. Aziz, "A soft-decision fusion approach for multiple-sensor distributed binary detection systems," IEEE Trans. Aerosp. Electron. Syst., vol. 47, no. 3, pp. 2208–2216, 2011.

[44] F. Gerstmayer, J. Hausladen, M. Kramer, and M. Horauer, "Binary protection framework for embedded systems," 2017 12th IEEE Int. Symp. Ind. Embed. Syst. SIES 2017 - Proc., 2017.

[45] J. Zhen and Z. Liu, "New honeypot system and its application in security of employment network," Proc. - 2012 IEEE Symp. Robot. Appl. ISRA 2012, pp. 627–629, 2012.

[46] A. Pauna, I. Bica, F. Pop, and A. Castiglione, "On the rewards of self-adaptive IoT honeypots," Ann. des Telecommun. Telecommun., vol. 74, no. 7–8, pp. 501–515, 2019.

[47] R. K. S. authorBazila B. Hota, Attack Detection and Forensics Using Honeypot in IoT Environment Rajesh, vol. 2, no. Dec. Springer International Publishing, 2018.

[48] A. Amjad, A. Griffiths, and M. Patwary, "QoI-Aware Unified Framework for Node Classification and Self-Reconfiguration Within Heterogeneous Visual Sensor Networks," IEEE Access, vol. 4, pp. 9027–9042, 2016.

[49] M. Wang, "Understanding Security Flaws of IoT Protocols through Honeypot Technologies," J. Opt. Soc. Am., 2017.

[50] D. Ramirez, J. I. Uribe, L. Francaviglia, P. Romero-Gomez, A. Fontcuberta i Morral, and F. Jaramillo, "IoTCandyJar: Towards an Intelligent-Interaction Honeypot for IoT Devices," J. Mater. Chem. C, vol. 6, no. 23, pp. 6216–6221, 2017.

[51] E. Pricop, J. Fattahi, N. Dutta, and M. Ibrahim, Recent Developments on Industrial Control Systems Resilience. 2020.

[52] H. Sharma and K. Govindan, Advances in Computing and Intelligent Systems. 2019.

[53] A. Umamaheswari and B. Kalaavathi, "Honeypot TB-IDS: trace back model based intrusion detection system using knowledge based honeypot construction model," Cluster Comput., vol. 4, pp. 1–8, 2018.

[54] H. Hadeli, R. Schierholz, M. Braendle, and C. Tuduce, "Leveraging determinism in industrial control systems for advanced anomaly detection and reliable security configuration," ETFA 2009 - 2009 IEEE Conf. Emerg. Technol. Fact. Autom., pp. 1–8, 2009.

[55] N. C. Rowe, "Honeypot Deception Tactics," Auton. Cyber Decept., pp. 35–45, 2019.

[56] A. Noaman, A. Abdel-Hamid, and K. Eskaf, "A novel honeynet architecture using software agents," 2019 Int. Conf. Innov. Intell. Informatics, Comput. Technol. 3ICT 2019, pp. 1–6, 2019.

[57] P. Duessel, C. Gehl, U. Flegel, S. Dietrich, and M. Meier, "Detecting zero-day attacks using context-aware anomaly detection at the application-layer," Int. J. Inf. Secur., vol. 16, no. 5, pp. 475–490, 2017.

[58] G. Fedynyshyn, M. C. Chuah, and G. Tan, "Detection and classification of different botnet C&C channels," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6906 LNCS, pp. 228–242, 2011.

[59] S. Morishita et al., "Detect me if you... Oh wait. An internet-wide view of self-revealing honeypots," 2019 IFIP/IEEE Symp. Integr. Netw. Serv. Manag. IM 2019, no. 1, pp. 134–143, 2019.

[60] C. Dalamagkas et al., "A Survey on honeypots, honeynets and their applications on smart grid," Proc. 2019 IEEE Conf. Netw. Softwarization Unleashing Power Netw. Softwarization, NetSoft 2019, pp. 93–100, 2019.

[61] D. Gur, H. E. Rockette, and A. I. Bandos, "'Binary' and 'non-binary' detection tasks: are current performance measures optimal?," Acad. Radiol., vol. 14, no. 7, pp. 871–876, 2007.

[62] L. Fernandez Maimo, A. L. Perales Gomez, F. J. Garcia Clemente, M. Gil Perez, and G. Martinez Perez, "A Self-Adaptive Deep Learning-Based System for Anomaly Detection in 5G Networks," IEEE Access, vol. 6, no. c, pp. 7700–7712, 2018.

[63] C. Han and Y. Zhang, "CODDULM: An approach for detecting C&C domains of DGA on passive DNS traffic," Proc. 2017 6th Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2017, vol. 2018-Janua, pp. 385–388, 2018.

[64] H. Šemić and S. Mrdovic, "IoT honeypot: A multi-component solution for handling manual and Mirai-based attacks," 2017 25th Telecommun. Forum, TELFOR 2017 - Proc., vol. 2017-Janua, pp. 1–4, 2018.

[65] C. Tzagkarakis, N. Petroulakis, and S. Ioannidis, "Botnet attack detection at the IoT edge based on sparse representation," Glob. IoT Summit, GIoTS 2019 - Proc., pp. 1–6, 2019.

[66] R. M. Campbell, "A Survey of Honeypot Research : Trends and Opportunities," pp. 208–212, 2015.

[67] M. F. Razali, G. Muruti, M. N. Razali, N. Jamil, and F. Z. Mansor, "IoT honeypot: A review from researcher's perspective," 2018 IEEE Conf. Appl. Inf. Netw. Secur. AINS 2018, pp. 93–98, 2019.

[68] R. Vishwakarma, "A Honeypot with Machine Learning based Detection Framework for defending IoT based Botnet DDoS Attacks," 2019 3rd Int. Conf. Trends Electron. Informatics, no. Icoei, pp. 1019–1024, 2019.

[69] K. P. Murphy, "Machine Learning - A Probabilistic Perspective - Table-of-Contents," MIT Press, 2012.

[70] Z. Wang et al., "Honeynet construction based on intrusion detection," ACM Int. Conf. Proceeding Ser., 2019.

[71] J. Ren and Y. Xu, "A compartmental model to explore the interplay between virus epidemics and honeynet potency," Appl. Math. Model., vol. 59, pp. 86–99, 2018.

[72] A. Belqruch and A. Maach, "SCADA security using SSH honeypot," ACM Int. Conf. Proceeding Ser., vol. Part F1481, 2019.

[73] A. Vetterl, R. Clayton, and I. Walden, "Counting outdated honeypots: Legal and useful," Proc. - 2019 IEEE Symp. Secur. Priv. Work. SPW 2019, no. 2001, pp. 224–229, 2019.

[74] L. Christopher, K. K. R. Choo, and A. Dehghantanha, Honeypots for Employee Information Security Awareness and Education Training: A Conceptual EASY Training Model. Elsevier Inc., 2016.

[75] E. P. Joshi and P. S. Barth, "Honeypots and Honeynets: Level of Interaction and Issues in Privacy," vol. 21, no. 16, pp. 881–885, 2019.

[76] S. Sekhar, D. K. Vijayakumar, B. Ketan, P. Swagatam, and D. Editors, Advances in Intelligent Systems and Computing 517 Artificial Intelligence and Evolutionary Computations in Engineering Systems, vol. 517. 2016.

[77] R. Breuk, "A visual analytic approach for analyzing SSH honeypots," 2012.

[78] S. T. Park, G. Li, and J. C. Hong, "A study on smart factory-based ambient intelligence context-aware intrusion detection system using machine learning," Journal of Ambient Intelligence and Humanized Computing, vol. 0, no. 0, Springer Berlin Heidelberg, p. 0, 2018.

[79] S. K. Choi, C. H. Yang, and J. Kwak, "System hardening and security monitoring for IoT devices to mitigate IoT security vulnerabilities and threats," KSII Trans. Internet Inf. Syst., vol. 12, no. 2, pp. 906–918, 2018.

[80] W. Wang, Y. Shang, Y. He, Y. Li, and J. Liu, "BotMark: Automated botnet detection with hybrid analysis of flow-based and graph-based traffic behaviors," Inf. Sci. (Ny)., vol. 511, pp. 284–296, 2020.

[81] S. Dowling, M. Schukat, and H. Melvin, "A ZigBee honeypot to assess IoT cyberattack behaviour," 2017 28th Irish Signals Syst. Conf. ISSC 2017, pp. 0–5, 2017.

[82] M. Singh, M. Singh, and S. Kaur, "Issues and challenges in DNS based botnet detection: A survey," Comput. Secur., vol. 86, pp. 28–52, 2019.

[83] M. Anirudh, S. Arul Thileeban, and D. J. Nallathambi, "Use of honeypots for mitigating DoS attacks targeted on IoT networks," Int.

Conf. Comput. Commun. Signal Process. Spec. Focus IoT, ICCCSP 2017, pp. 8–11, 2017.

[84] L. Yin, X. Luo, C. Zhu, L. Wang, Z. Xu, and H. Lu, “ConnSpoiler : Disrupting C & C Communication of IoT-Based Botnet through Fast Detection of Anomalous Domain Queries,” IEEE Trans. Ind. Informatics, vol. PP, no. c, p. 1, 2019.

[85] M. Baykara and R. Das, “Journal of Information Security and Applications A novel honeypot based security approach for real-time intrusion detection and prevention systems,” J. Inf. Secur. Appl., vol. 41, pp. 103–116, 2018.

[86] Y. ALEKSIEVA, H. VALCHANOV, and V. ALEKSIEVA, “An approach for host based botnet detection system,” 2019 16th Conf. Electr. Mach. Drives Power Syst., no. June, pp. 1–4, 2019.

[87] A. Oliveri and F. Lauria, “Sagishi: an undercover software agent for infiltrating IoT botnets,” Netw. Secur., vol. 2019, no. 1, pp. 9–14, 2019.

[88] R. Alhajri, R. Zagrouba, and F. Al-Haidari, “Survey for Anomaly Detection of IoT Botnets Using Machine Learning Auto-Encoders,” Int. J. Appl. Eng. Res., vol. 14, no. 10, pp. 2417–2421, 2019.

[89] S. H. Mousavi, M. Khansari, and R. Rahmani, “A fully scalable big data framework for Botnet detection based on network traffic analysis,” Inf. Sci. (Ny)., 2019.

[90] N. Koroniotis, N. Moustafa, and E. Sitnikova, Towards Developing Network Forensic Mechanism for Botnet Activities in the IoT Based on Machine Learning Techniques, vol. 235. Springer International Publishing, 2018.

[91] Y. M. P. Pa, S. Suzuki, K. Yoshioka, T. Matsumoto, T. Kasama, and C. Rossow, “IoTPOT: A novel honeypot for revealing current IoT threats,” J. Inf. Process., vol. 24, no. 3, pp. 522–533, 2016.

[92] M. Boer, M. Friedrich, M. Krämer, P. Noack, J. N. Weiss, and A. Zimmermann, Routing Protocol in the Industrial Internet of Things for Smart Factory Monitoring Abdellah. Springer Singapore, 2019.

[93] Y. Meidan et al., “N-BaIoT-Network-based detection of IoT botnet attacks using deep autoencoders,” IEEE Pervasive Comput., vol. 17, no. 3, pp. 12–22, 2018.

[94] S. K. Choi, C. H. Yang, and J. Kwak, “System hardening and security monitoring for IoT devices to mitigate IoT security vulnerabilities and threats,” KSII Trans. Internet Inf. Syst., vol. 12, no. 2, pp. 906–918, 2018.

[95] X. Yuan, C. Li, and X. Li, “DeepDefense: Identifying DDoS Attack via Deep Learning,” 2017 IEEE Int. Conf. Smart Comput. SMARTCOMP 2017, pp. 1–8, 2017.

[96] J. Guarnizo et al., “SIPHON: Towards scalable high-interaction physical honeypots,” CPSS 2017 - Proc. 3rd ACM Work. Cyber-Physical Syst. Secur. co-located with ASIA CCS 2017, pp. 57–68, 2017.

[97] A. Guerra Manzanares, “HoneyIo4 The construction of a virtual, low-interaction IoT Honeypot Treball Final de Grau,” 2017.

[98] P. Mesa and A. Rodr, B SIEM-IoT : A Blockchain-Based and Distributed SIEM. 2019.

[99] M. Wang, J. Santillan, and F. Kuipers, “ThingPot: an interactive Internet-of-Things honeypot,” 2018.

[100] M. Nawrocki, M. Wählisch, T. C. Schmidt, C. Keil, and J. Schönfelder, “A Survey on Honeypot Software and Data Analysis,” 2016.

[101] L. Mathur, M. Raheja, and P. Ahlawat, “Botnet Detection via mining of network traffic flow,” Procedia Comput. Sci., vol. 132, pp. 1668–1677, 2018.

[102] Z.A. Almusaylim and N. Zaman, “A review on smart home present state and challenges: linked to context-awareness internet of things (IoT) Wireless Networks”, 25 (6), 3193-3204.

[103] Alyssa Anne Ubing, Syukrina Kamilia Binti Jasmi, Azween Abdullah, NZ Jhanjhi and Mahadevan Supramaniam, “Phishing Website Detection: An Improved Accuracy through Feature Selection and Ensemble Learning” International Journal of Advanced Computer Science and Applications(IJACSA), 10(1), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100133

[104] S.H. Kok, A. Abdullah, NZ. Jhanjhi and M. Supramaniam, “A Review of Intrusion Detection System using Machine Learning Approach”, International Journal of Engineering Research and Technology 12 (1), 8-15.

[105] M. Lim, A. Abdullah, N. Jhanjhi, M. Khurram Khan and M. Supramaniam, "Link Prediction in Time-Evolving Criminal Network With Deep Reinforcement Learning Technique," in IEEE Access, vol. 7, pp. 184797-184807, 2019. doi: 10.1109/ACCESS.2019.2958873

[106] Lim, M.; Abdullah, A.; Jhanjhi, N.; Supramaniam, M. Hidden Link Prediction in Criminal Networks Using the Deep Reinforcement Learning Technique. Computers 2019, 8, 8.

AUTHORS’ PROFILE

Mr. Lee Seungjin is currently pursuing his Master’s Degree in Taylor’s University. Previously, he was employed as part of a cold-air manufacturing team under Samsung Electronics in South Korea

Dr. Azween Abdullah is a professional development alumni of Stanford University and MIT and his work experiences include thirty years as an academic in institutions of higher learning and as the Director of Research and Academic Affairs at two institutions of higher learning, Vice-President for educational consultancy services, 15 years in commercial companies as a Software Engineer, Systems Analyst and as a computer software developer and IT/MIS consultancy and training.

Dr. Noor Zaman has awarded as top reviewer 1% globally by WoS/ISI (Publons) recently for the year 2019. He has edited/authored more than 13 research books with international reputed publishers, earned several research grants, and a great number of indexed research articles on his credit. He has supervised several postgraduate students, including master’s and PhD. Dr Noor Zaman Jhanjhi is an Associate Editor of IEEE ACCESS, moderator of IEEE TechRxiv, Keynote speaker for several IEEE international conferences globally, External examiner/evaluator for PhD and masters for several universities, Guest editor of several reputed journals, member of the editorial board of several research journals, and active TPC member of reputed conferences around the globe.

Routing Protocol based on Floyd-Warshall Algorithm Allowing Maximization of Throughput

Kohei Arai
Graduate School of Science and Engineering
Saga University, Saga City
Japan

***Abstract*—Routing protocol based on Floyd-Warshall algorithm which allows maximization of throughput is proposed. The metric function in the proposed routing protocol is throughput including not only send packets but also retransmission packets in order for improving effectiveness and efficiency of the network in concern. Through simulation studies, it is found that the proposed routing protocol is superior to the conventional Open Shortest Path First: OSPF based on Dijkstra algorithm for shortest path determination from the point of view of maximizing throughput. A routing protocol for Virtual Private Network (VPN) in Autonomous System (AS) based on maximizing throughput is proposed. Through a comparison between the proposed protocol and the existing protocols, OSPF (Widely used), it is also found that the required time for transmission of packets from one node to another node of the proposed protocol is 56.54% less than that of the OSPF protocol.**

Keywords—Network routing protocol; virtual private network; autonomous system; open shortest path first; Floyd-Warshall algorithm; Dijkstra algorithm; throughput

I. Introiduction

There are two types of adaptive protocols (Interior Gateway Protocol: IGP) [1] used for path control in an autonomous system (Autonomous System: AS)[2]: one based on vector distance and one based on link state[3]. The latter is better. One of the latter is the Routing Information Protocol[4] (RIP)[5] , which is frequently used at present, which stores a table (Routing Table: RT) that indicates the next route to be connected every 30 seconds. This is a method of rewriting in accordance with the number of nodes (hop number) passing before reaching the network node.

As a result, it is possible to cope with a case where a failure occurs in the network or a case where the scale of the network is changed. In recent years, Open Shortest Path First (OSPF)[6] has been put to practical use and used frequently[7]. This is basically one of the adaptive protocols like RIP, and rewrites RT based on the link state (shortest distance). RIP has the danger of rewriting the table infinitely.

For example, when sending a packet from node A to C via B, if the bus to C fails, B's RT cannot be sent to C, so it is rewritten to send to A. In addition, even when the network topology changes, it takes time to transition to a stable state, and there is a risk of falling into an infinite loop. Further, there is a problem that a load is imposed on all nodes at the time of broadcasting. RIPv2, which compensates for this shortcoming, has also been proposed, in which information that avoids the above-mentioned problems is added to the header of the file indicating the link status.

A packet indicating the link state is broadcast to all nodes other than the target node. Based on this, the shortest route is obtained by Dijkstra Algorithm[8]: DA [1], the shortest route to all target nodes is found, and the method of rewriting RT is Link State Protocol: LSP. However, in LSP, the broadcast interval is 30 seconds, so if there is a change during that time, the tracking may be delayed.

If this interval is shortened, the following speeds up, but only the link state packet may occupy the network. Therefore, it has been considered to avoid this problem by adding time information to the link state bucket (Time Stamping[9]). OSPF is a device that adds such a device and supports the netmask (Variable-Length Sub-Net Masking: VLSM[10]. On the other hand, when constructing a virtual private network between earth observation data supply organizations and exchanging attribute information (metadata), etc. of acquired data with each other, a large amount of data is transmitted from a specific organization to a specific organization occurs.

In such a case, if the shortest path from any node to all the target nodes is obtained by the DA, the shortest path may not be between two specific nodes. Therefore, instead of DA, we introduce Floyd-Warshall [2], which guarantees the shortest path between two nodes. OSPF also believes that the shortest path leads to the shortest required time, the shortest delay time, and the lowest packet loss probability. However, here the link communication capacity, network throughput (link packet occupancy) and an optimal path control method considering the processing capability of the computer of the node in the evaluation function [3].

For the proposed route control method, the network topology was assumed, the packets were generated by random numbers, and the time required for the packets to arrive at the

[1] https://en.wikipedia.org/wiki/Interior_gateway_protocol
[2] https://en.wikipedia.org/wiki/Autonomous_system_%28Internet%29
[3] http://www.cla.kobe-u.ac.jp/Jouhou/95/Ishizaki/suuri.html
[4] https://ja.wikipedia.org/wiki/Routing_Information_Protocol
[5] http://www.acc.com/internet/whitepapers/iprouting.html
[6] https://ja.wikipedia.org/wiki/Open_Shortest_Path_First
[7] http://www.ecse.rpi.edu/Courses/S98/35696/illrip/index.htm
[8] https://en.wikipedia.org/wiki/Dijkstra%27s_algorithm
[9] https://en.wikipedia.org/wiki/Trusted_timestamping
[10] http://www.vlsm.org/rms46/imho-e-vlsmtjt.html

destination node and the delay time at that time were evaluated. The results are compared with OSPF, and the superiority of the proposed method is confirmed.

The following section describes research background. Then the proposed method is described followed by experiment. After that, conclusion is described together with some discussions.

II. Related Research Works

Approach of improved topology development protocol in Ad Hoc network minimizing the number of hops and maintaining connectivity of mobile terminals which move from one to the others is proposed and attempt [4]. Then, routing approach with immediate awareness of adaptive path while minimizing the number of hops and maintaining connectivity of mobile terminals which move from one to the others is proposed and evaluated [5].

On the other hand, agent based approach of routing protocol minimizing the number of hops and maintaining connectivity of mobile terminals which move from one area to the other is attempted [6] together with approach of improved topology development protocol in ad-hoc network minimizing the number of hops and maintaining connectivity of mobile terminals which moves from one to the others [7].

Backup communication routing through Internet Satellite, WINDS, for transmission of disaster relief data is proposed and well reported [8] together with backup communication routing through Internet satellite WINDS for transmission of disaster relief data [9]. Also, back-up communication routing through Internet satellite WINDS for transmitting of disaster relief data is proposed and evaluated [10].

Meanwhile, service robot with communication aid together with routing controlled by human eyes proposed and well report with experiments [11]. More recently, with the spread of mobile terminals equipped with wireless LAN as standard equipment, research on mobile ad-hoc networks (MANET), which builds a network with only wireless terminals, is drawing attention. Many ad hoc routing protocols proposed in MANET do not consider the traffic state on the route when generating the route, so even if the number of relay hops is short, the route with relatively high load is selected. There is a problem of being lost. An ad-hoc routing protocol that is capable of route generation considering traffic on the route by extending OLSR (Optimized Link State Routing) is proposed by Akira Morisaki et al.[12].

SrcRR: A High Throughput Routing Protocol for 802.11 Mesh Network is also proposed by Daniel Aguayo, John Bicket, Robert Morris [13]. SrcRR uses its own transmit bit-rate selection algorithm based on medium-term loss rate measurements, replacing the algorithm built into the radio firmware. Mobile Ad hoc Networks (MANET) are wireless networks consisting of a collection of mobile nodes with no fixed infrastructure. Due to their decentralized, self-configuring and dynamic nature, MANETs offer many advantages and are easy to install. In the recent years, a lot of researches are going on in the area of Mobile Ad hoc Networks (MANETs). This network is an infrastructure -less network where nodes communicate with each other without any aid of centralized administration. In this paper, we are analyzing Throughput of DSR routing protocol [14].

High Throughput Cryptocurrency Routing in Payment Channel Networks is proposed by Vibhaalakshmi Sivarama et al. [15]. They proposed Spider, a routing solution that “packetizes” transactions and uses a multi-path transport protocol to achieve high-throughput routing in PCNs. Packetization allows Spider to complete even large transactions on low-capacity payment channels over time, while the multi-path congestion control protocol ensures balanced utilization of channels and fairness across flows.

III. Proposed Routing Protocol

Considering the throughput *Ti* as “*Ni*”, the number of packets arriving at an arbitrary node *i* in a unit time, the time t_1 from the arrival of an arbitrary packet at an arbitrary node i to another arbitrary node *j* is determined by nodes *i*, *j* If the communication capacity of the link between is α, it can be expressed by equation (1).

$$t_l = Ci / Ni = Ci / Ti \tag{1}$$

The processing power of the node's computer is directly related to the time required for a packet to pass through the node. Considering the processing capacity *Ai* as the number MZ of packets passing through an arbitrary node *i* in a unit time, and considering the delay time *Di* at the node as the length of the packet queue [4],

$$t_n = (Ti + Di\ ) / Az \tag{2}$$

The shortest path that minimizes the time *ti* required to pass through these nodes and links is determined.

$$ti = ti + tn = \{\mathrm{CiAi} + Ti^2 + TiDi\} / TiAi \rightarrow \mathrm{min} \tag{3}$$

Here, the load *Li* is defined as follows.

$$\mathrm{Li} = \mathrm{Ni} / \mathrm{Ci} = 1/\mathrm{t}_\mathrm{l} = \mathrm{Ti} / \mathrm{Ci} \tag{4}$$

The route control method proposed here seeks an optimal route between two nodes that optimizes the weighted average m of the load and the distance.

That is, it is a value obtained by normalizing the throughput by the communication capacity. The route control method proposed here seeks an optimal route between two nodes that optimizes the weighted average m of the load and the distance.

$$m = \omega_\mathrm{L} Li + \omega_\mathrm{t}\ ti \rightarrow \mathrm{min.} \tag{5}$$

Here, $\omega_L + \omega_t = 1$. DA is a kind of polynomial time algorithm in dynamic programming [4], which guarantees the shortest path for all destination nodes. Here, an optimization method that guarantees the shortest path between any two nodes is necessary, and the algorithm of Floyd-Warshall: WF (see Appendix) is adopted as an algorithm that satisfies this.

In other words, using the link state information of the target node written in the RT, find the shortest path to the target node based on the processing capacity of the node and the communication capacity of the link known in advance, rewrite RT, and This is a method in which information is broadcast as link state packets every 30 seconds as in OSPF.

The differences from OSPF are summarized below.

- In OSPF, the shortest path from any node to all nodes is obtained by DA, but in the proposed method, the shortest path between two nodes from any node to the target node is obtained by WF.
- Find the optimal route based on the combination of link communication capacity and node processing capacity not considered in OSPF.
- The concept of throughput is newly introduced, and the optimized route is obtained by optimizing the weighted average of the load normalized by the communication capacity and the distance considered by OSPF.

IV. Simulation Experiments

A. *Method for Simulation*

Since a universal simulation seems to be impossible, a special example (a network topology with 8 nodes) was considered here and the simulation was performed in the following order.

1) Randomly determine the network topology (8 and 16 nodes), source and destination nodes. At that time, the communication capacity of the link and the processing capacity of the node are determined at random.

2) A packet is generated from each node using a uniform random number during 2000 time slots (unit time) that does not hinder the calculation of throughput and communication distance, and the distance between all nodes and the load are calculated in advance.

3) At this time, the unit time is the length of the packet. This is equivalent to generating a fixed-length packet in a slotted time, such as an ATM (Asynchronous Transfer Mode[11]) cell. The occurrence probability is the same for each node and can be set arbitrarily. Furthermore, when the queue of each node exceeds 20, it is assumed that the node is disconnected until the next unit time.

4) After that, the weight of the distance and the load is optimized, and the optimal route between any two nodes is obtained by the WF algorithm.

5) Compare the result and the shortest path between all nodes determined by the DA algorithm based on only the distance with the path between the two nodes (corresponding to the shortest path obtained by OSPF).

B. *Simulation Results*

Fig. 1 shows the network configuration of the eight nodes used in this example.

In the figure, the processing capacity of the node and the communication capacity of the link are shown. Fig. 2 and 3 show the simulation results of the distance and load for the configuration of 8 nodes and 16 nodes. The load in the figure is the total load of 2000 time slots. In the case of 16 nodes, a new 8 node is added to the existing 8 nodes. At this time, the packet occurrence probability was set to 0.1. Here, the average number of packets generated by each node in 2000 time slots is 200. Table I shows the case of the shortest path control of OSPF in the case of 8 nodes, and the difference of the path by the minimum load path control proposed here and the optimal path control with the shortest distance and the minimum load weighted.

In the table, the notations of 50:50, 90:10, and 99:1 correspond to load and distance weights of 0.5: 0.5, 0.9: 0.1, and 0.99: 0.01, respectively. In the case of the shortest route, the DA algorithm, the minimum load, and the weighted optimal route are obtained by the WF algorithm. When the weight is 0.5: 0.5, almost the same route as OSPF is selected. Although only 4-8 is different, in this case, the distance is the same for 4-1-2-6-8 and 4-7-8, but the load is the natural consequence of the latter being smaller.

When the weight is set to 0.9: 0.1, the route is different from OSPF. This situation is not so different from the case where the weight is set to 0.99: 0.01, so 0.9: 0.1 is adopted here as the weight. It was confirmed that this result was also applicable to the case of 16 nodes.

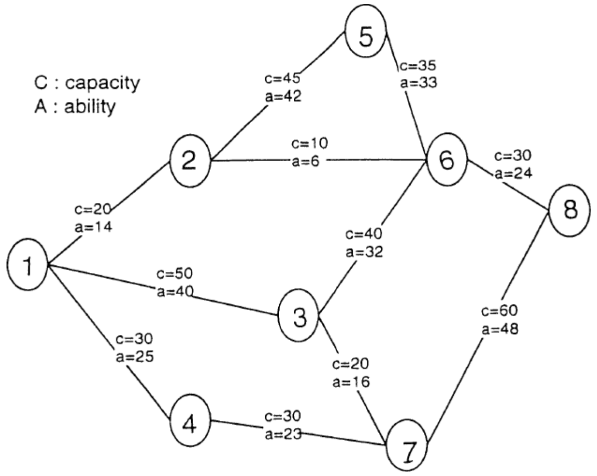


Fig. 1. Network Configuration of Interest (8 Nodes).

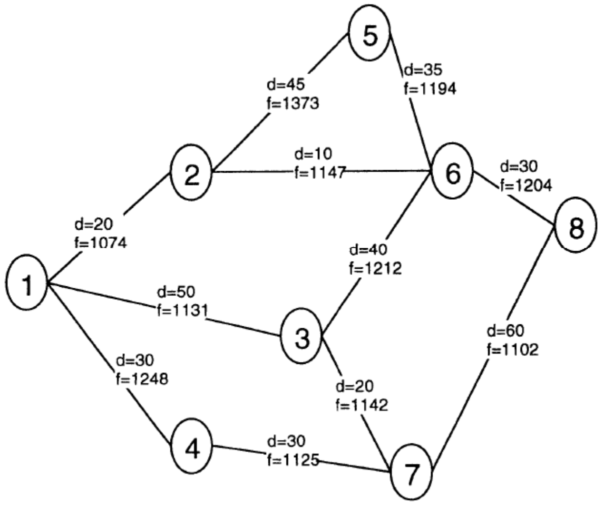


Fig. 2. The Distance (d) and Accumulated Load (f) Between Nodes Over 2000 Time Slots for 8 Nodes of Network Configuration.

[11] https://ja.wikipedia.org/wiki/Asynchronous_Transfer_Mode

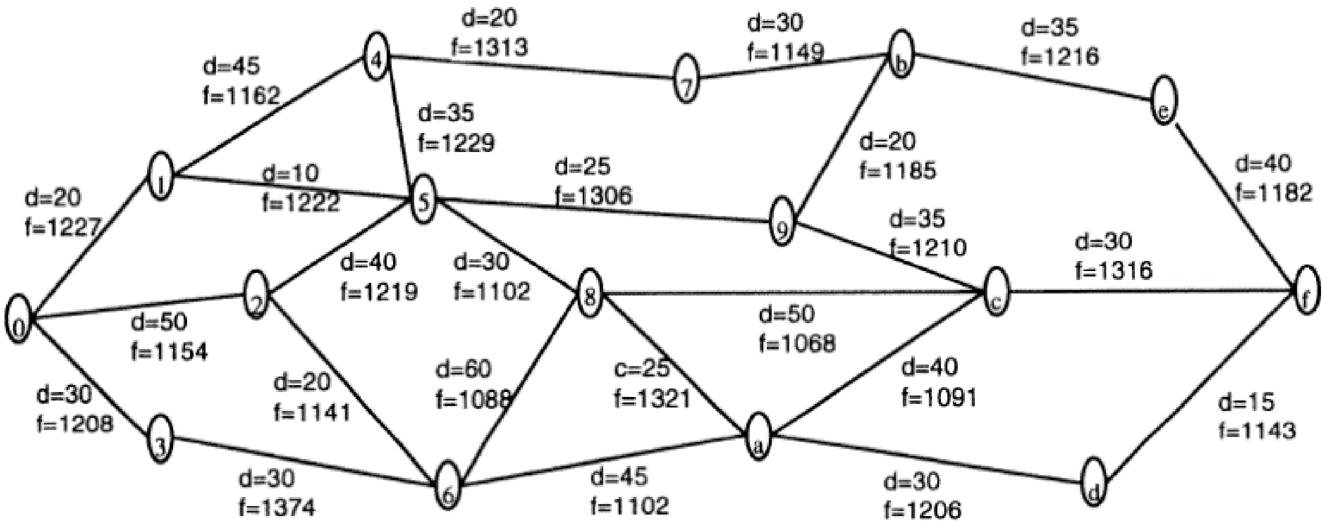


Fig. 3. The Distance (d) and Accumulated Load (f) Between Nodes Over 2000 Time Slots for 16 Nodes of Network Configuration.

TABLE I. A COMPARISON OF OPTIMUM ROUTE AMONG OSPF, MINIMIZING LOAD AND MINIMIZING WEIGHTED AVERAGE BETWEEN DISTANCE AND LOAD FOR 8 NODES NETWORK CONFIGURATION

start-goal	OSPF	Min. Load	Weight(0.5,0.5)	0.9:0.1	0.99:0.01
1-2	1-2	1-2	1-2	1-2	1-2
1-3	1-3	1-3	1-3	1-3	1-3
1-4	1-4	1-4	1-4	1-4	1-4
1-5	1-2-5	1-2-5	1-2-5	1-2-5	1-2-5
1-6	1-2-6	1-2-6	1-2-6	1-2-6	1-2-6
1-7	1-4-7	1-3-7	1-4-7	1-3-7	1-3-7
1-8	1-2-6-8	1-3-7-8	1-2-6-8	1-2-6-8	1-3-7-8
2-3	2-6-3	2-1-3	2-6-3	2-1-3	2-1-3
2-4	2-1-4	2-1-4	2-1-4	2-1-4	2-1-4
2-5	2-5	2-5	2-5	2-5	2-5
2-6	2-6	2-6	2-6	2-6	2-6
2-7	2-1-3-7	2-6-8-7	2-1-3-7	2-1-3-7	2-6-8-7
2-8	2-6-8	2-6-8	2-6-8	2-6-8	2-6-8
3-4	3-7-4	3-7-4	3-7-4	3-1-4	3-7-4
3-5	3-6-5	3-6-5	3-6-5	3-6-5	3-6-5
3-6	3-6	3-6	3-6	3-6	3-6
3-7	3-7	3-7	3-7	3-7	3-7
3-8	3-6-8	3-7-8	3-6-8	3-7-8	3-7-8
4-5	4-1-2-5	4-1-2-5	4-1-2-5	4-1-2-5	4-1-2-5
4-6	4-1-2-6	4-7-8-6	4-1-2-6	4-1-2-6	4-7-8-6
4-7	4-7	4-7	4-7	4-7	4-7
4-8	4-1-2-6-8	4-7-8	4-7-8	4-7-8	4-7-8
5-6	5-6	5-6	5-6	5-6	5-6
5-7	5-6-3-7	5-6-8-7	5-6-3-7	5-6-3-7	5-6-8-7
5-8	5-6-8	5-6-8	5-6-8	5-6-8	5-6-8
6-7	6-3-7	6-8-7	6-3-7	6-3-7	6-8-7
6-8	6-8	6-8	6-8	6-8	6-8
7-8	7-8	7-8	7-8	7-8	7-8

TABLE II. MEAN ARRIVAL TIME FOR EACH 100 OF TIME-SLOT (APPROXIMATELY 56.54% OF IMPROVEMENT OF THE TIME REQUIRED FOR TRANSMISSION OF PACKETS IS EXPECTED FOR THE PROPOSED PROTOCOL COMPARED TO THE OSPF IN CASE OF 8 NODES

Time set	OSPF(min. distance)	Min.Load	Optimum weight for distance and load
100	9.78	9.78	7.96
200	7.87	Fail	7.87
300	16.23	16.23	16.23
400	Fail	Fail	9.94
500	16.57	16.57	7.34
600	10	35.98	7.84
700	9.2	9.2	9.19
800	Fail	Fail	7.45
900	8.42	26.73	8.42
1000	11.27	11.27	6.75
1100	10.32	10.32	7.08
1200	10.06	10.06	6.63
1300	41.03	41.03	10.85
1400	Fail	40.4	9.44
1500	7.77	7.77	7.77
1600	Fail	66.84	8.87
1700	16.08	16.08	7.12
1800	7.98	26.81	7.98
1900	24.9	24.9	17.88
2000	8.39	8.39	8.12
average	20.79	26.42	9.14

Furthermore, when the average time required for every 100 time slots from the generation of a bucket to the arrival of the packet when the packet generation node (Start) is 1 and the destination node (Goal) is 8 is as shown in Table II.

Therefore, comparing only the shortest path and the minimum load of OSPF, OSPF is superior except in the case of time slots 1400 and 1600. OSPF is inferior when the queue is broken and exceeds 20, and is avoided by minimum load control. The proposed route control with weight 0.9: 0.1 is not inferior to OSPF in all cases. It was confirmed that this conclusion was the same for the case of 16 nodes.

The biggest difference between OSPF and the proposed route control is that OSPF uses DA to minimize the distance from any node to all nodes, whereas the proposed method uses WF to determine the distance between any nodes. The point is to minimize it. To confirm the effect, we examined the time required to transmit a packet from node 1 to 8 for 8 nodes and from node 0 to f for 16 nodes. As a result, as shown in Table III, it was confirmed that the average arrival time over 2000 time slots can be reduced by 36.58% and 13.69% for 8 nodes and 16 nodes, respectively, by the proposed method.

TABLE III. AVERAGED PACKET ARRIVAL TIME FOR THE NETWORK WITH 8 NODES (FROM NODE 1 TO 8) AND FOR THE NETWORK WITH 16 NODES (FROM NODE 0 TO F) OVER 2000 TIME SLOTS

Network Node	OSPF	Proposed	Percent Improvement
8 Nodes from Node 1 to 8	5.56	3.52	36.38
16 Nodes from 1 to f	7.23	6.24	13.69

V. CONCLUSION

Routing protocol based on Floyd-Warshall algorithm allowing maximization of throughput is proposed. Through simulation studies, it is found that the proposed routing protocol is superior to the conventional Open Shortest Path First: OSPF based on Dijkstra algorithm for shortest path determination from the point of view of maximizing throughput.

A routing protocol for Virtual Private Network (VPN) in Autonomous System (AS) based on maximizing throughput is proposed. Through a comparison between the proposed protocol and the existing protocols, OSPF (Widely used), it was found that the required time for transmission of packets from one node to another node of the proposed protocol is 56.54% less than that of the OSPF protocol.

The proposed route control method in an autonomous system is based on a similar method (OSPF) that is already widely used, and introduces a new concept of throughput in the shortest distance that OSPF is based on, and calculates the weighted average of them. Is the new norm. By this, it was confirmed that it was possible to search for an optimal route with a light load, although not the shortest distance, but the distance was rather short.

In this paper, it is confirmed that 36.58 and 13.69% reduction in arrival time can be achieved with the network configurations of 8 and 16 nodes, respectively.

VI. FURTHER RESEARCH WORKS

This time, the author has only given an example of the packet occurrence probability, but we plan to examine more cases in the future. In addition, automatic estimation of the optimal weight of distance and load is also a subject for the future.

VII. APPENDIX (FLOYD-WARSHALL'S ALGORITHM)

The algorithm generates a series of matrices by successively including new nodes that give the shortest path to each node pair. In the matrix *Dk*, the path between any two nodes uses only the nodes 1 to *k*. Therefore, if there are *N* nodes in total, a true shortest path is obtained when *k* = *N*. The algorithm is as follows:

1. Assuming that the total number of nodes is *N*, assign a number from 1 to *N* to each node in the network.
2. Define an $N \times N$ matrix M so that the element a_{ij} indicates the length of the path connecting the node *i* to the node *j*.
3. If there is no path connecting *i* and *j*, a_{ij} is set to infinity (∞) or set to a sufficiently large value (Max) for the purpose of calculation.
4. An $N \times N$ matrix *Dk* ($k = 0,..., N$) is sequentially generated as follows.
5. The matrix D^k is for all node pairs i, *j* such that the path between node *i* and node j number may include some node from the set (1,2, ..., *k*) as well as node *i*, *j* This can be explained as giving the shortest path length. (Therefore, matrix *D*- consists only of the nodes at the end of each path, which is similar to matrix *M*.) Matrix D^{k+1} determines whether or not to include node $k + 1$ in the already existing shortest path. It is obtained from the matrix $D^{k.}$

$$d_{i,j}^{k+1}=\min(d_{i,j}^{k1}, d_{i,k+1}^{k1}+d_{k+1,j}^{k1}) \quad (6)$$

Here, the function min takes the minimum value among its arguments. That is, in the shortest path for each node, if the path including the node of $k + 1$ is shorter than the path of the matrix D^K, the node of $k + 1$ is included.

6. The matrix D^N gives the shortest path.

ACKNOWLEDGMENT

This research was recalled when one of the authors was the chair and vice-chair of the Network Working Group at the International Committee on Earth Observation Satellites (CEOS), and was given the opportunity to attend the conference. Special thanks to all members of the group. In addition, this study was discussed regularly.

Special thanks to Dr. Shinichi Sobue of NASDA and Dr. Richard des Jordin of NASA. Also, the author would like to thank Prof. Dr. Hiroshi Okumura, Prof. Dr. Osamu Fukuda and Mr. Hirofumi Etoh of Saga University for their valuable comments and suggestions.

REFERENCES

[1] Masao Fukushima, Introduction to Mathematical Programming (in Japanese), Asakura Shoten, Tokyo, 1996

[2] Toshihide Ibaraki, Masao Fukushima, Methods of Optimization (in Japanese), Kyoritu Shuppan, Tokyo, 1993. [Chinese edition: Translated by D.Z. Zeng, published by World Publishing Corporation, Beijing, 1997

[3] D. R. Cox and Walter L. Smith. Queues, Methuen, London; Wiley, New York, 1961.

[4] Kohei Arai, Lipur Sugiyanta, Approach of improved topology development protocol in Ad Hoc network minimizing the number of hops and maintaining connectivity of mobile terminals which move from one to the others, Anthony V. Stavros Edt., Advances in Communication and Media Research, Vol.8, ISBN 978-1-61324-794-5, 2011

[5] Kohei Arai, Lipur Sugiyanta, Routing Approach with Immediate Awareness of Adaptive Path While Minimizing the Number of Hops and Maintaining Connectivity of Mobile Terminals Which Move from One to the Others, International Journal of Computer Science and Information security, 9, 2, 94-101, 2011.

[6] Kohei Arai and Lipur Sugiyanta, Agent based approach of routing protocol minimizing the number of hops and maintaining connectivity of mobile terminals which move from one area to the other, Proceedings of the International Conference on Computational Science and Its Applications (ICCSA2010), LNCS part-III, 305-320,2010.

[7] Kohei Arai, Lipur Sugiyanta, Approach of improved topology development protocol in ad-hoc network minimizing the number of hops and maintaining connectivity of mobile terminals which moves from one to the others, Journal of Communication and Networks, 2, 6, 190-204, 2011.

[8] Kohei Arai, Kiyotaka Fujisaki, Hiroaki Ikemi, Masato Masuya, Terumasa Miyahara, Backup communication routing through Internet Satellite, WINDS, for transmission of disaster relief data, Proceedings of the International Symposium on WINDS Application Experiments, 2010.

[9] Kohei Arai, Backup communication routing through Internet satellite WINDS for transmission of disaster relief data, Proceedings of the International Symposium on WINDS, 2010

[10] Kohei Arai, Back-up communication routing through Internet satellite WINDS for transmitting of disaster relief data, International Journal of Advanced Computer Science and Applications, 2, 9, 21-26, 2011.

[11] Kohei Arai, Service robot with comminunication aid together with routing controlled by human eyes, Journal of Image Laboratory, 25, 6, 24-29, 2014

[12] Akira Morisaki et al., http://www.wata-lab.meijo-u.ac.jp/file/convention/2009/200907-DICOMO-Akira_Morisaki.pdf

[13] Daniel Aguayo, John Bicket, Robert Morris, http://pdos.csail.mit.edu/~rtm/srcrr-draft.pdf

[14] GirishTiwari1,*Ram Shiromani Gupta, Throughput Based Analysis of DSR Routing Protocol in MANET, IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 12, Issue 6, Ver. II P 56-62, 2017.

[15] Vibhaalakshmi Sivarama et al., High Throughput Cryptocurrency Routing in Payment Channel Networks, https://arxiv.org/pdf/1809.05088.pdf

AUTHOR'S PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 55 books and published 620 journal papers as well as 450 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.html

Knowledge Sharing Framework for Modern Code Review to Diminish Software Engineering Waste

Nargis Fatima[1], Sumaira Nazir[2,] Suriayati Chuprat[3]
Razak Faculty of Technology and Informatics
University Technology Malaysia (UTM), Kuala Lumpur, Malaysia[1, 2, 3]
Faculty of Engineering and Computer Science
National University of Modern Languages (NUML), Islamabad, Pakistan[1, 2]

***Abstract*—Modern Code Review (MCR) is a quality assurance technique that involves massive interactions between team members of MCR. Presently team members of MCR are confronting with the problem of waiting waste production, which results in their psychological distress and project delays. Therefore, the MCR team needs to have effective knowledge sharing during MCR activities, to avoid the circumstances that lead the team members to the waiting state. The objective of this study is to develop the knowledge sharing framework for MCR team to reduce waiting waste. The research methodology used for this study is the Delphi survey. The conducted Delphi survey intended to produce the finalized list of knowledge sharing factors and to recognize and prioritize the most influencing knowledge sharing factor for MCR activities. The study results reported 22 knowledge sharing factors, 135 sub-factor, and 5 categories. Grounded on the results of the Delphi survey the knowledge sharing framework for MCR has been developed. The study is beneficial for software engineering researchers to outspread the research. It can also help the MCR team members to consider the designed framework to increase knowledge sharing and diminish waiting waste.**

Keywords—Knowledge sharing; modern code review; software engineering wastes; waiting waste; lean software development

I. Introduction

Software engineering is known as a systematic application of engineering approaches to the development of software [1]. It highly involves social interaction among stakeholders for the development of cost-effective software [2]. It includes sub-activities such as software requirement recognition, software modeling, software testing, inspections, and Modern Code Review (MCR) [3]. These activities yield wastes for instance rework, defect, needless composite solution, waiting, extra or erroneous feature, and mental distress [3], [4]. The various perception of wastes available in the literature are given in Table I.

MCR, a lightweight software engineering activity, has its origin from Fagan's review process [5], [6] and is largely known since 2013 [5], [7]. Fagan's review process is a heavyweight code inspection that requires face to face communications between team members [8]. While MCR is informal, easy-going, and supported through review tools [5], [9]. MCR aims to improve software quality through the improvement of source code quality [5], [10], [11]. It is being practiced by numerous organizations, for instance, Microsoft, Google, etc. [9], [12].

Though MCR has overcome the inadequacies of Fagan's review process [16] and is aimed to enhance the source code quality through widespread knowledge sharing between team members of MCR [5], [9], [12], [10], however, the MCR produces waiting waste due to lack of knowledge sharing [4], [9], [17], [18], [19], [20].

Even though the existing research has paid attention to knowledge sharing concerning software engineering activities [21], [22], [23] however, knowledge sharing in the context of MCR warrants attention from the researchers [9], [10], [12], [24], regarding explorations of knowledge sharing factors for MCR activities [25]. No, schematized inquiries are available about that knowledge sharing facet concerning MCR to decrease waiting waste. Thus, to minimize waiting waste, this study aims to develop a knowledge sharing framework for MCR.

This study is the an extension of our previous work that involved the identification of knowledge sharing factors for MCR through Systematic Literature Review (SLR) and expert review [25]. The result of SLR and expert review are reported in [24], [25]. In our previous studies, the SLR [24], [25] was performed to identify the knowledge sharing factors from the literature and the expert review [25] has been performed to validate the identified list of knowledge sharing factor. In this study, the Delphi survey has been conducted with experts from the industry to finalize the list of knowledge sharing factors, sub-factors and categories for their practicality concerning the industry, to identify and prioritize the most influential knowledge sharing factors for MCR activities, to get suggestion about naming conventions, grouping, and sub-grouping of provided knowledge sharing factors, sub-factors, and categories, to recognize new industry-based knowledge sharing factors, with their associated sub-factors, and categories in the context of MCR. The results of the Delphi survey have been utilized to develop Knowledge sharing framework for MCR to minimize waiting waste.

The remaining paper is organized as Section II describes the research background. The research methodology is discussed in Section III while Section IV introduces the results of the Delphi study. Section V highlights the study conclusion. Section VI highlights future work suggestions. Section VII highpoints the study contribution.

TABLE I. DEFINATION OF OF WASTES FROM LITERATURE

Definition	Reference
“All activities and work products that do not contribute to customer value” or “Everything that is not consider valuable” or “Non- efficient way of working” or “Everything that does not make it to the release i.e. product feature/qualities not delivered and were a waste of time to investigate and or develop”	[2]
“Activities that absorb resources and increase cost without adding value”.	[13]
“Any Bottlenecks” or “Waste is anything that does not add value to a product, value as perceived by the customer”.	[2], [14]
“Something happens against the flow”.	[2], [15]

II. BACKGROUND

Software engineering is a well-disciplined approach to develop quality software [26]. It is social as well as a technical activity that integrates additional activities [3], [27] such as software requirement recognition, software modeling, software testing, inspections, and MCR. These activities generate several wastes [2], [3], [4]. Waste may lead to mental distress, project delays, and software failure. The research on waste recognition and reduction has been started in the 1980s when Toyota revolutionized the automobile industry with a “Lean Manufacturing” [4], [14]. In the year 2000, the lean manufacturing concept was shifted from manufacturing to software engineering domain [28] and was named as lean software development. Since then numerous researches have been reported in the software engineering domain focusing on waste recognition and reduction [2], [3], [4].

In the software engineering domain several wastes have been identified for instance extra or erroneous features, “task switching, defects, “relearning and handoff”, needless composite solutions, rework, “extraneous cognitive load, and waiting [2], [3], [4]. Though each software engineering activity includes different software engineering actions, therefore, each activity can generate various distinct wastes [3]. MCR is a critical software engineering activity to improve code quality [5], [29], [30]. In this activity, the reviewer reviews the source code, prior to sending it to the code repository. MCR is supported with the aid of review tools, for instance, Code flow, such as Gerrit, Review board, Phabricator, etc. [5], [9], [10], [12], [31]. Fig. 1 represents the MCR process overview.

It is claimed that waste such as extra or erroneous features, defects, needless composite solutions, rework, and waiting are generated during MCR [2], [3], [4]. It is also conveyed that if the organization needs to minimize one waste, then the organization must emphasize waiting waste [2], [15], [28]. Waiting waste deals with delay between two consecutive activities [3], [4]. For instance, in the case of MCR time delay between source code submission for review by the author and receiving feedback from the reviewer [9], [10]. It is stated that one of the reasons behind waiting waste in MCR is a lack of knowledge sharing [4], [9], [18], [19], [32]. The waiting waste affects the efficiency and productivity of the developers [2], [3], [9], [18], [19], [32], and it also leads to project delays [2].

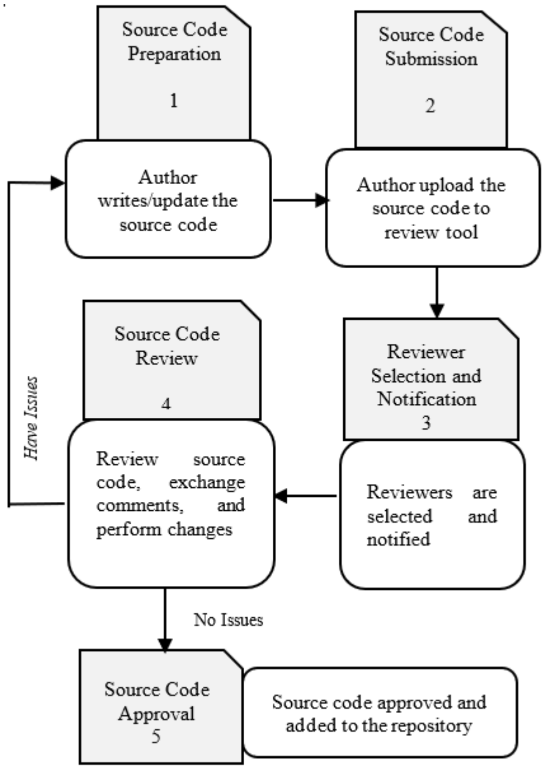

Source Code Preparation
1
Author writes/update the source code
Source Code Submission
2
Author upload the source code to review tool
Reviewer Selection and Notification
3
Reviewers are selected and notified
Source Code Review
4
Review source code, exchange comments, and perform changes
Have Issues
No Issues
Source Code Approval
5
Source code approved and added to the repository

Fig. 1. MCR Process Overview [10].

To diminish the waiting waste it is mandatory to focus on knowledge sharing [2], [3], [4], [33] in MCR. It is reported that knowledge sharing among team members can be augmented by recognizing the factors that can influence knowledge sharing [9], [10], [12], [24]. Considering those factors can aid in knowledge sharing among MCR team members.

Limited researches have been performed concerning knowledge sharing in MCR highlighting the significance of knowledge sharing [10], [12], [34]. For instance, Sadowski et al., (2018), quantify knowledge sharing by looking at comments and files edited or reviewed. They reported that developers build experience through knowledge sharing while working at Google [12]. Similarly, Bosu et al., (2017), stated that code review allows senior developers to mentor newcomers. They conveyed that experienced developers can also enhance their skills while sharing knowledge [10]. Likewise, Rigby and Bird (2013) have explored knowledge sharing facet in code review. They measure the amount of knowledge shared among the MCR team through the number of files known to the developer before and after the source code review [34]. The literature shows that, although existing studies [10], [12], [34] have provided attention towards knowledge sharing in MCR, however, no framework or guidelines are available for effective knowledge sharing in MCR that can help MCR team to reduce software engineering waiting waste. Therefore, this study aims to develop the knowledge sharing framework for MCR to reduce software engineering waiting waste.

III. Research Methodology

Delphi survey has been performed as a research methodology for this study. This methodology is a less costly and relatively competent way to get consensus from the opinions of the experts [35]. It typically involves iterative questionnaires directed to individual experts in such a way that their anonymity is preserved. Feedback received in the Delphi survey after each questionnaire iteration continues until consensus is achieved. The Delphi output is the consensus among the experts along with their observations on the questionnaire items.

A. Objective of Delphi Survey Conduction

A two-round Delphi study has been performed 1) to evaluate the practicality of the identified knowledge sharing factors, sub-factors and their categories in the context of MCR with industry 2) to recognize and prioritize the most influential knowledge sharing factors concerning MCR activities 3) to get suggestion about naming conventions, grouping, and sub-grouping of provided knowledge sharing factors, sub-factors and categories 4) to recognize new industry-based knowledge sharing factors, with their associated sub-factors, and categories in the context of MCR. Delphi survey was conducted based on the guidelines given by [36]. The steps involved in the Delphi survey are detailed in subsections.

B. Delphi Experts' Selection

The selection of the experts to participate in the Delphi study is a very important and critical aspect as the output of the Delphi survey relies on the experts' opinions [36]. Based on the experts' selection requirement conveyed in literature [36], in this study, the experts were selected based on the criteria such as (1) Expert have experience of more than 8 years in the software industry, (2) Expert should have experience in MCR, (3) Expert should have knowledge of wastes in context of software engineering and knowledge sharing. Other selection criteria involve their willingness to participate in the survey as well as enough time to provide feedback [36].

C. Delphi Panel Size

Panel size deals with the number of experts to participate in the study. The panel size varies from a few to hundreds of experts [36]. The size of the panel for the Delphi study is variable. It is conveyed that with a homogenous group of people, ten to fifteen experts might be enough [37]. We requested fifteen experts to participate in the survey. Ten experts showed their interests and willingness to participate.

D. Delphi Rounds

The conducted Delphi survey involved two rounds. The expert's input was collected through questionnaires. The experts were explained each provided knowledge sharing factor, sub-factor, and category to make sure that all of the experts have a shared understanding of knowledge sharing factors. It is conveyed that in the Delphi study, most convergence of panel responses occurs between round one and two [38]. In this study, the consensus among the experts was achieved in two rounds.

E. Delphi Questionnaire Plan

The questionnaire for Round 1 involved four sections. Section A aimed to collect demographic information from the experts. Section B of the questionnaire was composed of a list of knowledge sharing factors, related sub-factors, and categories generated as a result of our previous study based on SLR and expert review [24], [25]. In Section B the experts were also questioned to score the knowledge sharing factors for their practicality and level of influence for MCR activities. Section C was designed to obtain new knowledge sharing factors, sub-factors or categories that should be included in the list. Section C also aimed to collect suggestions about naming conventions, grouping, and sub-grouping of the provided knowledge sharing factors, related sub-factors, or categories. Section D aimed to obtain information about recent real project examples for which the experts had performed MCR activities and experienced the factors influencing knowledge sharing. This section was specifically designed for generating the scenario that was later used in the experiment to validate the developed knowledge sharing framework.

The questionnaire for Round 2 involved three Sections. Section A aimed to evaluate the practicality of knowledge sharing factors finalized after Delphi survey Round 1. The finalized list contains the changes made based on the recommended suggestion of the experts in Round 1. This Section also aimed to evaluate the influence level of listed factors for each MCR activity. Section B aimed to get any new factors, related sub-factors, and categories that should be included in the list. In Section B the experts were also requested to mention the suggestions about the naming conventions grouping and sub-grouping of the provided knowledge sharing factors, sub-factors, and categories. Section C aimed to obtain information about recent real project examples for which the experts had performed MCR activities and experienced the factors influencing knowledge sharing.

F. Pilot Study

The questionnaires were evaluated by five software engineering researchers for their understanding and clarity as it is conveyed that if the questionnaires are used in research, then they should be pretested for length, clarity, and overall adequacy [39]. In the pilot test of this study, the received response was positive and no changes were suggested.

G. Data Analaysis Procedure

Descriptive statistics have been performed in this study as it is a rudimentary analytical approach. These give a basic quantitative strategy for examination and produce a general overview of the outcomes [40].

To score the practicality of knowledge sharing factors and to evaluate the level of influence of knowledge sharing factors for each MCR activity, a five-point Likert scale that is from 1 to 5 (Very High- 5, High - 4, Moderate - 3, Low- 2, Very Low – 1) was provided. For calculating the practicality of knowledge sharing factors and to recognize the most influential Knowledge sharing factors for MCR activities, the mean values were grouped into the discrete categories as shown in Table II for MCR activities.

TABLE II. GROUPING OF MEAN VALUES TO MEASURE PRACTICALITY

Mean Score =X	Practicality Level	Influence Level
4.0≤X≤ 5.0	Very High	Most Influential
3.0≤X< 4.0	High	Influential
2.0≤X< 3.0	Moderate	Moderate
1.0≤X< 2.0	Low	Weakly Influential
0≤X< 1.0	Very Low	Not Influential

The mean practicality and mean influential values of sub-factors were premeditated initially and then the found mean values were further transformed into a single composite mean value showing composite mean practicality and composite mean influence value for the associated knowledge sharing factors.

To get the consensus of the practicality and the influential values of knowledge sharing factors we used the standard deviation as shown in Table III. Initially, we calculated the standard deviation of the sub-factors that were further transformed into a single composite standard deviation for the associated knowledge sharing factor. Based on the obtained composite standard deviation of the knowledge sharing factors we come up with the consensus level among the experts. We formulated equation (1) based on guidelines given by [41] [41] to calculate the composite standard deviation of knowledge sharing factors.

$$SD(KSF) = \sqrt{\frac{(SD\ (SbF_1))^2 + \cdots + \ (SD\ (SbF_k))^2}{k}} \quad (1)$$

Where '*SD*' denotes to standard deviation, '*KSF*' refers to knowledge sharing factor. '*SbF*' refers to the sub-factor of the associated knowledge sharing factor and it ranges from 1 to k, '*k*' refers to the total number of sub-factors for associated knowledge sharing factors.

Table III represents the level of consensus used in this study. A standard deviation between '0' and '1' shows that the experts scoring is very close to each other, whereas a higher standard deviation showed that the experts' scoring was spread out over a large range [35].

H. Data Collection and Analysis Methods

This section presents the data collected from Delphi experts and the analysis of the data collected depending on the analysis procedure defined in sub-section 'G'. The performed Delphi study involved two rounds. The details concerning data collection are discussed in the following sub-sections.

TABLE III. DECISION CRITERIA FOR THE LEVEL OF CONSENSUS

Standard Deviation (SD=X)	Level of Consensus
0 ≤ X <1	High
1 ≤ X <1.5	Fair Level
1.5 ≤ X <2	Low Level
2 < X	No Consensus

I. Delphi Round 1

In the Delphi Round 1, the questionnaire was given to the experts. They were given one week to complete the questionnaire. The phone calls were made to make sure that all experts were aware of the feedback submission date and time for Round 1. Round 1 of the Delphi survey was completed in two weeks. Round 1 aimed to collect demographic information from the experts. It also aimed to evaluate the list of provided knowledge sharing factors, related sub-factors, and categories for their naming convention, grouping, and sub-grouping which was generated as a result of our previous study based on SLR and expert review [24], [25], [42]. Round 1 involves the evaluation of the knowledge sharing factors for their practicality for the complete MCR process as well as their influence level for each MCR activity. In Round 1, the experts were also enquired to state any new industry-based knowledge sharing factors, related sub-factors, and categories that should be included in the list. The scale used to score the practicality and influence level is given in sub-section 'G'. The details about the Round 1 questionnaire is provided in sub-section 'E'. In Delphi Round 1 some recommendations were suggested by the expert so we need to conduct another Delphi round to have consensus on the suggested changes among the experts.

J. Delphi Round 2

In Round 2, the experts were given the summary of results obtained in Round 1. In Delphi Round 2 the experts were enquired to evaluate the level of practicality as well as the level of influence of subsequent knowledge sharing factors finalized after Round 1 for each MCR activity. In Round 2, the analysis method and the scoring scale was similar as in the case of Round 1. The details about the Round 2 questionnaire is provided in sub-section 'E'. Round 2 also took 2 weeks to be completed. In Round 2 the consensus was obtained for all the knowledge sharing factors therefore we stopped at the Delphi Round 2.

IV. RESULTS

This section presents the results obtained in the two Rounds of Delphi study. The results were then analyzed, and composite mean values of knowledge sharing factors were calculated based upon the mean values of their associated sub-factors. Similarly, the mean influential values of knowledge sharing factors were calculated based upon the mean influential values of their associated sub-factors. The practicality level of each knowledge sharing factors along with the standard deviation for Delphi Round 1 and Delphi Round 2 are shown in Fig. 2 and Fig. 3. Fig. 2 shows that all the provided knowledge sharing factors in both rounds were perceived as practical by the experts as the composite mean value of all the factors lies between 3 and 5. Fig. 3 shows that the level of consensus was increased in Round 2 for the practicality of the identified knowledge sharing factors among the experts.

Table IV shows the ranking of knowledge sharing factors for their level of practicality.

Regarding the most influential knowledge sharing factors, the mean influential values of sub-factors of each knowledge sharing factor in final Delphi Round for; Source Code Preparation values were from 1.4 to 5.0, Source Code Submission values were from 1.4 to 5.0, Reviewer Selection and Notification values were from 1.1 to 5.0, Source Code Review ranges from 2.4 to 5.0, Source Code Approval values were from 2.0 to 5.0. The most influential factors were identified by calculating the composite mean influential value of their connected sub-factors. The factors with composite mean values equal to or above 4.00 were considered as the most influential factors for particular MCR activity. The most influential factors grounded on their composite mean values for each MCR activity after the final Delphi Round are shown in Tables V to IX along with the standard deviation.

Based on the Delphi survey results we formulated a knowledge sharing framework for MCR to diminish waiting waste. The developed framework constitutes knowledge sharing factors, sub-factors, and categories as well as the most influential knowledge sharing factors for each MCR activity. The developed knowledge sharing framework is attached in Appendix A.

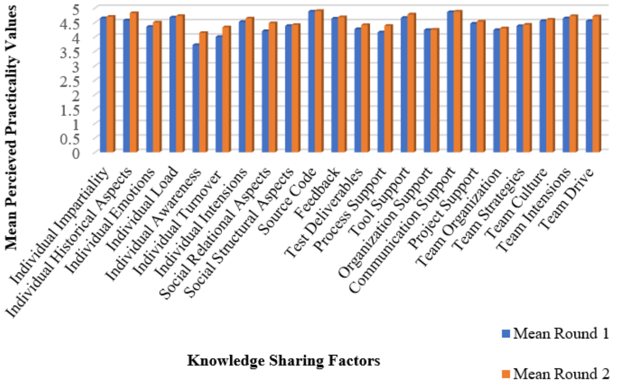


Fig. 2. Composite mean Perceived Value of Practicality of Knowledge Sharing Factors (Round 1 and Round 2).

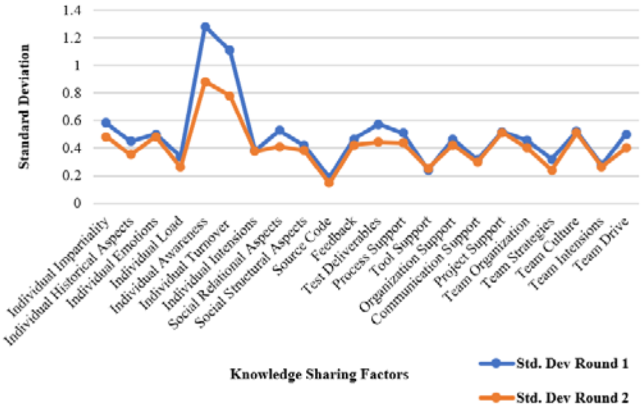


Fig. 3. Consensus Level among the Panelists for mean Perceived Values of Practicality of Knowledge Sharing Factors (Round 1 and Round 2).

TABLE IV. RANKING OF KNOWLEDGE SHARING FACTORS FOR PERCEIVED LEVEL OF PRACTICALITY

Knowledge Sharing Factors	**Composite Mean Practicality Values**	**Standard Deviation**	**Rank**
Source Code	4.9	0.146176337	1
Communication Support	4.88	0.298142397	2
Individual Historical Aspects	4.825	0.353553391	3
Tool Support	4.78	0.253859104	4
Individual Load	4.725	0.263523138	5
Team Intensions	4.722	0.265274142	6
Team Drive	4.714	0.402373908	7
Individual Impartiality	4.7	0.483045892	8
Feedback	4.68	0.423515147	9
Individual Intensions	4.64	0.377123617	10
Team Culture	4.6	0.510990324	11
Project Support	4.53	0.512799145	12
Individual Emotions	4.5	0.483045892	13
Social Relational Aspects	4.475	0.411636301	14
Team Strategies	4.425	0.241522946	15
Social Structural Aspects	4.4167	0.382486988	16
Test Deliverables	4.411	0.443053379	17
Process Support	4.383	0.436738756	18
Individual Turnover	4.333	0.779363463	19
Team Organization	4.3	0.402768199	20
Organization Support	4.25	0.421637021	21
Individual Awareness	4.14	0.880656321	22

TABLE V. INFLUENTIAL LEVEL OF KNOWLEDGE SHARING FACTORS FOR SOURCE CODE PREPARATION

Most influential Knowledge Sharing Factors	**Composite Mean Influential Value**	**Standard Deviation**	**Rank**
Source Code	4.92	0.170469437	1
Tool Support	4.63	0.357460176	2
Individual Historical Aspects	4.6	0.349602949	3
Team Strategies	4.57	0.437797518	4
Team Drive	4.44	0.311167795	5
Team Organization	4.44	0.359010987	6
Organization Support	4.4	0.357460176	7
Individual Load	4.37	0.337474279	8
Project Support	4.33	0.434613494	9
Feedback	4.1	0.333333333	10
Test Deliverables	4.08	0.293972368	11
Process Support	4.06	0.380058475	12
Individual Intensions	4.06	0.418993503	13
Individual Awareness	4	0.837987006	14

TABLE VI. INFLUENTIAL LEVEL OF KNOWLEDGE SHARING FACTORS FOR SOURCE CODE SUBMISSION

Most influential Knowledge Sharing Factors	Composite Mean Influential Value	Standard Deviation	Rank
Tool Support	4.51	0.202758751	1
Source Code	4.47	0.395487366	2
Test Deliverables	4.45	0.472712164	3
Team Strategies	4.4	0.45338235	4
Process Support	4.25	0.275546595	5
Project Support	4.18	0.215165741	6
Organization Support	4.1	0.298142397	7
Individual Historical Aspects	4	0.387298335	8

TABLE VII. INFLUENTIAL LEVEL OF KNOWLEDGE SHARING FACTORS FOR REVIEWER SELECTION AND NOTIFICATION

Most influential Knowledge Sharing Factors	Composite Mean Influential Value	Standard Deviation	Rank
Individual Historical Aspects	4.88	0.300462606	1
Social Structural Aspects	4.83	0.36004115	2
Social Relational Aspects	4.77	0.411636301	3
Individual Impartiality	4.7	0.471404521	4
Tool Support	4.52	0.194365063	5
Team Strategies	4.47	0.418330013	6
Team Culture	4.4	0.45338235	7
Organization Support	4.3	0.223606798	8
Project Support	4.15	0.129099445	9
Source Code	4.13	0.359248979	10
Process Support	4.1	0.403686714	11

TABLE VIII. INFLUENTIAL LEVEL OF KNOWLEDGE SHARING FACTORS FOR SOURCE CODE REVIEW

Most influential Knowledge Sharing Factors	Composite Mean Influential Value	Standard Deviation	Rank
Source Code	4.961	0.191708468	1
Individual Load	4.925	0.263523138	2
Test Deliverables	4.822	0.3022549	3
Communication Support	4.76	0.4163332	4
Individual Intensions	4.7	0.27080128	5
Tool Support	4.7	0.266666667	6
Feedback	4.68	0.191070501	7
Individual Impartiality	4.65	0.5	8
Team Intensions	4.53	0.210818511	9
Process Support	4.43	0.370185139	10
Individual Emotions	4.35	0.341565026	11
Individual Historical Aspects	4.3	0.278886676	12
Project Support	4.2	0.344265186	13
Organization Support	4.175	0.383695481	14
Team Strategies	4.1	0.25819889	15
Team Drive	4	0.338061702	16

TABLE IX. INFLUENTIAL LEVEL OF KNOWLEDGE SHARING FACTORS FOR SOURCE CODE APPROVAL

Most influential Knowledge Sharing Factors	Composite Mean Influential Value	Standard Deviation	Rank
Source Code	4.884	0.269535847	1
Individual Historical Aspects	4.75	0.353553391	2
Team Strategies	4.65	0.5	3
Tool Support	4.6	0.274873708	4
Project Support	4.53	0.327730693	5
Process Support	4.5	0.36004115	6
Organization Support	4.4	0.45338235	7
Team Culture	4.2	0.414996653	8
Individual Impartiality	4	0.459468292	9

V. CONCLUSION

Knowledge sharing plays a significant role in the minimization of waiting waste. This study involves statistical analysis of knowledge sharing factors to identify the list of most influential knowledge sharing factors for MCR activities. The study results reported 22 knowledge sharing factors, 135 sub-factor, and 5 categories. The obtained results were expressed as a knowledge sharing framework for MCR to diminish software engineering waste. This framework will guide the software engineers involved in MCR activities to effectively share knowledge and reduce the production of waiting waste.

VI. FUTURE WORK DIRECTIONS

This developed knowledge sharing framework is specific to the MCR activity of Software Engineering to diminish waiting waste. The study can be further extended to other software engineering activities to minimize waiting waste in other software engineering activities for instance requirement engineering, modeling, and testing. This research study delivers a list of factors influencing knowledge sharing in MCR to diminish waiting waste. Our ongoing research activities are 1) to validate the developed knowledge sharing framework regarding minimization of waiting waste through experiment, 2) to develop a web-based knowledge sharing framework for MCR to have an electronic knowledge sharing guideline for software engineers involved in MCR activities to minimize waiting waste.

VII. CONTRIBUTION

The investigation contributed to software engineering body of knowledge (SWEBOK), knowledge base software engineering (KBSE), and green software engineering (GREEN SE) by stressing the significance of knowledge sharing, most influencing knowledge sharing factors, and by providing the knowledge sharing framework for MCR to diminish waiting waste. The work can guide software engineers to effectively share knowledge by managing the undesirable facets of identified factors.

REFERENCES

[1] DeFranco and P. A. Laplante, “Review and Analysis of Software Development... - Google Scholar,” IEEE Trans. Prof. Commun., vol. 00, no. 00, pp. 1–18, 2017.

[2] H. Alahyari, T. Gorschek, and R. Berntsson Svensson, "An exploratory study of waste in software development organizations using agile or lean approaches: A multiple case study at 14 organizations," Inf. Softw. Technol., vol. 105, no. August 2018, pp. 78–94, 2019.

[3] T. Sedano and P. Ralph, "Software Development Waste," in Proc. IEEE/ACM 39th International Conference on Software Engineering, 2017.

[4] N. Fatima, S. Nazir, and S. Chuprat, "Software engineering wastes-A perspective of modern code review," ACM Int. Conf. Proceeding Ser., pp. 93–99, 2020.

[5] A. Bacchelli and C. Bird, "Expectations, outcomes, and challenges of modern code review," in Proc. International Conference on Software Engineering, 2013, pp. 712–721.

[6] S. Nazir, N. Fatima, and S. Chuprat, "Situational factors affecting Software Engineers Sustainability: A Vision of Modern Code Review," in 6th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS) , in press, 2019.

[7] S. Nazir, N. Fatima, and S. Chuprat, "Does Project Associated Situational Factors have Impact on Sustainability of Modern Code Review Workforce?," in 6th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS) , in press, 2019, pp. 1–5.

[8] M. E. Fagan, "Design and code inspections to reduce errors in program development," IBM Syst. J., vol. 38, no. 2.3, pp. 258–287, 1999.

[9] L. MacLeod, M. Greiler, M. A. Storey, C. Bird, and J. Czerwonka, "Code Reviewing in the Trenches: Challenges and Best Practices," IEEE Softw., vol. 35, no. 4, pp. 34–42, 2018.

[10] A. Bosu, J. C. Carver, C. Bird, J. Orbeck, and C. Chockley, "Process Aspects and Social Dynamics of Contemporary Code Review: Insights from Open Source Development and Industrial Practice at Microsoft," IEEE Trans. Softw. Eng., vol. 43, no. 1, pp. 56–75, 2017.

[11] S. Nazir, N. Fatima, and S. Chuprat, "Modern code review benefits-primary findings of a systematic literature review," in ACM International Conference Proceeding Series, 2020, pp. 210–215.

[12] C. Sadowski, E. Söderberg, L. Church, M. Sipko, and A. Bacchelli, "Modern code review: : A Case Study at Google," in Proc. ACM/IEEE 40th International Conference on Software Engineering: Software Engineering in Practice, 2018, pp. 181–190.

[13] M. V. P. Pessôa, W. Seering, and E. Rebentisch, "Understanding the waste net: A method for waste elimination prioritization in product development," Proc. DETC '08, vol. 55, no. 21, pp. 1–9, 2008.

[14] S. Mujtaba, R. Feldt, and K. Petersen, "Waste and lead time reduction in a software product customization process with value stream maps," Proc. Aust. Softw. Eng. Conf. ASWEC, pp. 139–148, 2010.

[15] J. Urrego, R. Munoz, M. Mercado, and D. Correal, "Archinotes: A global agile architecture design approach," Lect. Notes Bus. Inf. Process., vol. 179 LNBIP, pp. 302–311, 2014.

[16] S. Nazir, N. Fatima, and S. Malik, "Effective hybrid review process (EHRP)," Proc. - Int. Conf. Comput. Sci. Softw. Eng. CSSE 2008, vol. 2, pp. 763–771, 2008.

[17] G. Gousios, M.-A. Storey, and A. Bacchelli, "Work practices and challenges in pull-based development," in Proc. 38th International Conference on Software Engineering, 2016, pp. 285–296.

[18] E. W. dos Santos and I. Nunes, "Investigating the Effectiveness of Peer Code Review in Distributed Software Development," in Proc. 31st Brazilian Symposium on Software Engineering, 2017, pp. 84–93.

[19] D. M. German, U. Rey, and J. Carlos, "' Was my contribution fairly reviewed ?' A Framework to Study the Perception of Fairness in Modern Code Reviews," in Proc. ACM/IEEE 40th International Conference on Software Engineering Synthesizing, 2018, no. 2, pp. 523–534.

[20] L. Novikova, "Poor knowledge sharing is the second biggest challenge for software development teams," 2019. [Online]. Available: https://blog.onebar.io/poor-knowledge-sharing-is-the-second-biggest-challenge-for-software-development-teams-a4843f9b9aa. [Accessed: 10-Aug-2019].

[21] R. Anwar, M. Rehman, K. S. Wang, A. Amin, and R. Akbar, "Conceptual framework for implementation of knowledge sharing in global software development organizations," ISCAIE 2017 - 2017 IEEE Symp. Comput. Appl. Ind. Electron., pp. 174–178, 2017.

[22] X. Chen, Y. Zhou, D. Probert, and J. Su, "Managing knowledge sharing in distributed innovation from the perspective of developers: empirical study of open source software projects in China," Technol. Anal. Strateg. Manag., vol. 29, no. 1, pp. 1–22, 2017.

[23] N. S. Safa and R. Von Solms, "An information security knowledge sharing model in organizations," Comput. Human Behav., vol. 57, pp. 442–451, 2016.

[24] N. Fatima, S. Nazir, and S. Chuprat, "Knowledge sharing, a key sustainable practice is on risk: An insight from Modern Code Review," in 2019 6th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS), 2019, pp. 1–6.

[25] N. Fatima, S. Nazir, and S. Chuprat, "Knowledge sharing factors for modern code review to minimize software engineering waste," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 1, pp. 490–497, 2020.

[26] P. Bourque and R. E. Fairley, Software Engineering - Body of Knowledge. 2014.

[27] S. Nazir, N. Fatima, and S. Chuprat, "Individual Sustainability Barriers and Mitigation Strategies: Systematic Literature Review Protocol," in 2019 IEEE Conference on Open System, ICOS 2019, 2019, pp. 1–5.

[28] M. Poppendieck and T. Poppendieck, Lean Software Development: An Agile Toolkit. 2003.

[29] N. Fatima, S. Nazir, and S. Chuprat, "Individual, Social and Personnel Factors Influencing Modern Code Review Process," 2019 IEEE Conf. Open Syst. ICOS 2019, pp. 40–45, 2019.

[30] S. Nazir, N. Fatima, and S. Chuprat, "Situational factors for modern code review to support software engineers' sustainability," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 1, pp. 498–504, 2020.

[31] N. Fatima, S. Chuprat, and S. Nazir, "Challenges and Benefits of Modern Code Review-Systematic Literature Review Protocol," in Proc. International Conference on Smart Computing and Electronic Enterprise, 2018, pp. 1–5.

[32] O. Kononenko, O. Baysal, and M. W. Godfrey, "Code Review Quality: How Developers See It," in Proc. International Conference on Software Engineering, 2016, pp. 1028–1038.

[33] A. Ram, Achyudh ; Sawant, Anand; Castelluccio, Marco; Bacchelli, "What Makes a Code Change Easier to Review? An Empirical Investigation on Code Change Reviewability," in Proc. ESEC/FSE, 2018.

[34] P. C. Rigby and C. Bird, "Convergent Contemporary Software Peer Review Practices Categories and Subject Descriptors," in Proc. ESEC/FSE, 2013, pp. 202–212.

[35] H. A. von der Gracht, "Consensus measurement in Delphi studies. Review and implications for future quality assurance," Technol. Forecast. Soc. Change, vol. 79, no. 8, pp. 1525–1536, 2012.

[36] G. J. Skulmoski, F. T. Hartman, and Jennifer Krahn, "The Delphi Method for Graduate Research," J. Inf. Technol. Educ., vol. 6, 2007.

[37] T. Hatcher and S. Colton, "Using the internet to improve HRD research: The case of the web-based Delphi research technique to achieve content validity of an HRD-oriented measurement," J. Eur. Ind. Train., vol. 31, no. 7, pp. 570–587, 2007.

[38] H. W. Lanford, Technological forecasting methodologies; a synthesis. New York: American Management Association, 1972.

[39] D. F. Polit and C. T. Beck, Nursing research: generating and assessing evidence for nursing practice. Lippincott Williams and Wilkins, Philadelphia, 9th ed. LWW, 2011.

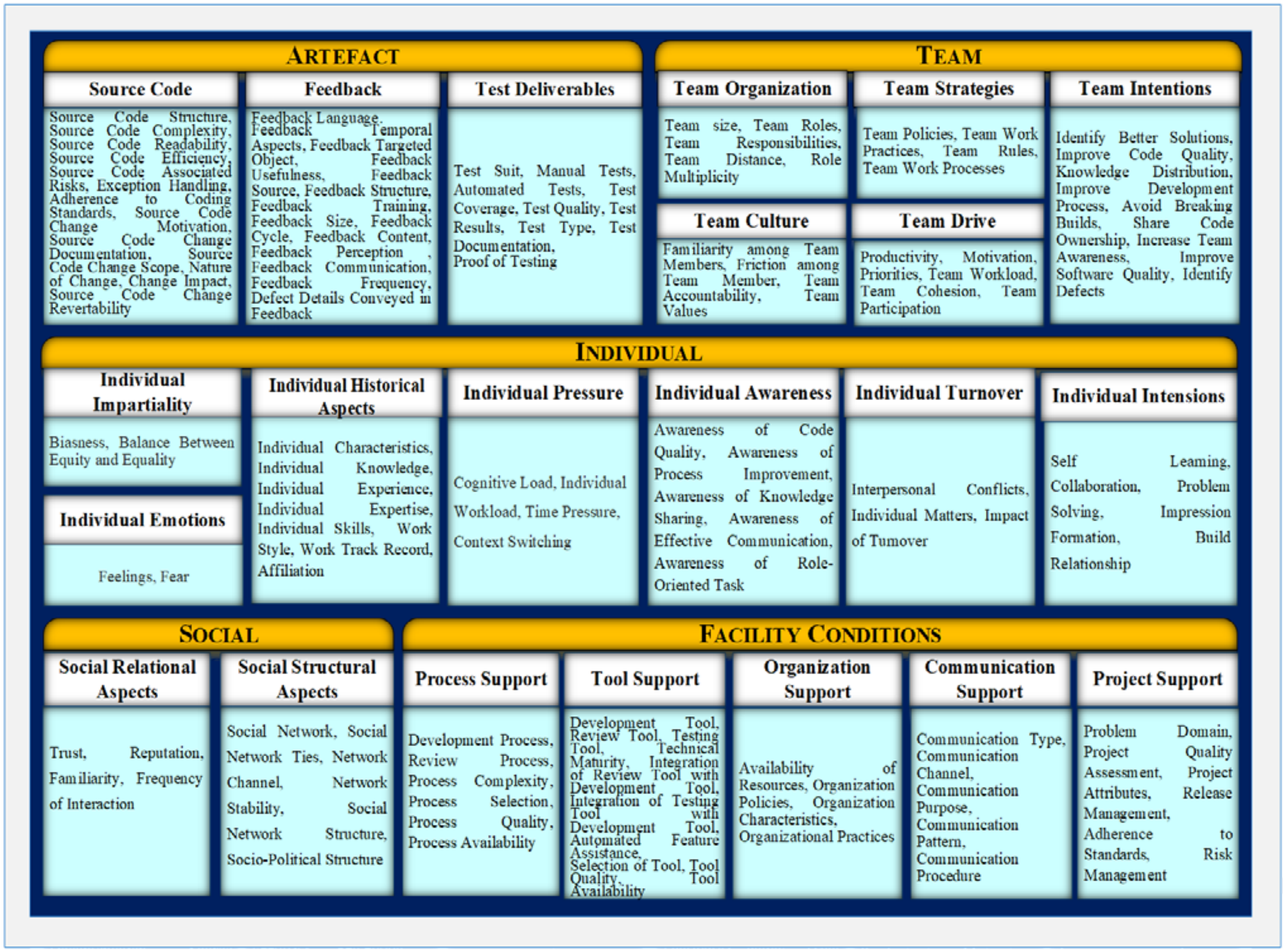
[40] S. G. Naoum, Dissertation research and writing for construction students, Second Edition, 2nd ed. Oxford : Butterworth-Heinemann, 2012.

[41] J. Cohen, Statistical power analysis for the behavioral science, 2nd ed. New: Lawrence Erlbaum Associates, 1988.

[42] N. Fatima, S. Nazir, and S. Chuprat, "Understanding the Impact of Feedback on Knowledge Sharing in Modern Code Review," in 6th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS) , In Press, 2019, pp. 1–5.

APPENDIX A

KNOWLEDGE SHARING FRAMEWORK FOR MODERN CODE REVIEW



ARTEFACT

Source Code

Source Code Structure, Source Code Complexity, Source Code Readability, Source Code Efficiency, Source Code Associated Risks, Exception Handling, Adherence to Coding Standards, Source Code Change Motivation, Source Code Change Documentation, Source Code Change Scope, Nature of Change, Change Impact, Source Code Change Revertability

Feedback

Feedback Language, Feedback Temporal Aspects, Feedback Targeted Object, Feedback Usefulness, Feedback Source, Feedback Structure, Feedback Training, Feedback Size, Feedback Cycle, Feedback Content, Feedback Perception , Feedback Communication, Feedback Frequency, Defect Details Conveyed in Feedback

Test Deliverables

Test Suit, Manual Tests, Automated Tests, Test Coverage, Test Quality, Test Results, Test Type, Test Documentation, Proof of Testing

TEAM

Team Organization

Team size, Team Roles, Team Responsibilities, Team Distance, Role Multiplicity

Team Strategies

Team Policies, Team Work Practices, Team Rules, Team Work Processes

Team Intentions

Identify Better Solutions, Improve Code Quality, Knowledge Distribution, Improve Development Process, Avoid Breaking Builds, Share Code Ownership, Increase Team Awareness, Improve Software Quality, Identify Defects

Team Culture

Familiarity among Team Members, Friction among Team Member, Team Accountability, Team Values

Team Drive

Productivity, Motivation, Priorities, Team Workload, Team Cohesion, Team Participation

INDIVIDUAL

Individual Impartiality

Biasness, Balance Between Equity and Equality

Individual Emotions

Feelings, Fear

Individual Historical Aspects

Individual Characteristics, Individual Knowledge, Individual Experience, Individual Expertise, Individual Skills, Work Style, Work Track Record, Affiliation

Individual Pressure

Cognitive Load, Individual Workload, Time Pressure, Context Switching

Individual Awareness

Awareness of Code Quality, Awareness of Process Improvement, Awareness of Knowledge Sharing, Awareness of Effective Communication, Awareness of Role-Oriented Task

Individual Turnover

Interpersonal Conflicts, Individual Matters, Impact of Turnover

Individual Intensions

Self Learning, Collaboration, Problem Solving, Impression Formation, Build Relationship

SOCIAL

Social Relational Aspects

Trust, Reputation, Familiarity, Frequency of Interaction

Social Structural Aspects

Social Network, Social Network Ties, Network Channel, Network Stability, Social Network Structure, Socio-Political Structure

FACILITY CONDITIONS

Process Support

Development Process, Review Process, Process Complexity, Process Selection, Process Quality, Process Availability

Tool Support

Development Tool, Review Tool, Testing Tool, Technical Maturity, Integration of Review Tool with Development Tool, Integration of Testing Tool with Development Tool, Automated Feature Assistance, Selection of Tool, Tool Quality, Tool Availability

Organization Support

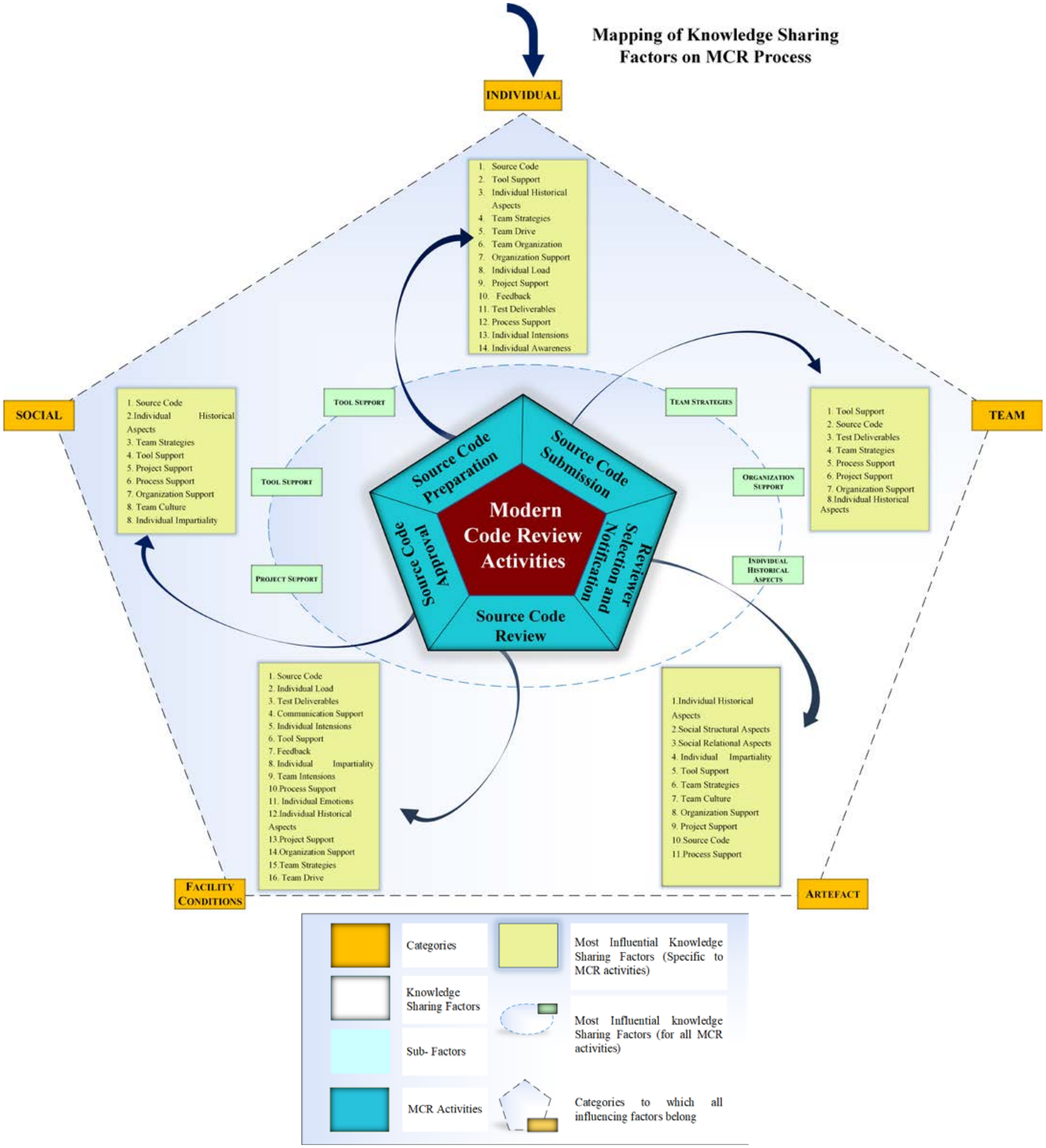
Availability of Resources, Organization Policies, Organization Characteristics, Organizational Practices

Communication Support

Communication Type, Communication Channel, Communication Purpose, Communication Pattern, Communication Procedure

Project Support

Problem Domain, Project Quality Assessment, Project Attributes, Release Management, Adherence to Standards, Risk Management


Mapping of Knowledge Sharing Factors on MCR Process
INDIVIDUAL
1. Source Code
2. Tool Support
3. Individual Historical Aspects
4. Team Strategies
5. Team Drive
6. Team Organization
7. Organization Support
8. Individual Load
9. Project Support
10. Feedback
11. Test Deliverables
12. Process Support
13. Individual Intensions
14. Individual Awareness
SOCIAL
1. Source Code
2.Individual Historical Aspects
3. Team Strategies
4. Tool Support
5. Project Support
6. Process Support
7. Organization Support
8. Team Culture
8. Individual Impartiality
TOOL SUPPORT
TOOL SUPPORT
PROJECT SUPPORT
TEAM STRATEGIES
ORGANIZATION SUPPORT
INDIVIDUAL HISTORICAL ASPECTS
TEAM
1. Tool Support
2. Source Code
3. Test Deliverables
4. Team Strategies
5. Process Support
6. Project Support
7. Organization Support
8.Individual Historical Aspects
Source Code Preparation
Source Code Submission
Reviewer Selection and Notification
Source Code Review
Source Code Approval
Modern Code Review Activities
1. Source Code
2. Individual Load
3. Test Deliverables
4. Communication Support
5. Individual Intensions
6. Tool Support
7. Feedback
8. Individual Impartiality
9. Team Intensions
10.Process Support
11. Individual Emotions
12.Individual Historical Aspects
13.Project Support
14.Organization Support
15.Team Strategies
16. Team Drive
1.Individual Historical Aspects
2.Social Structural Aspects
3.Social Relational Aspects
4. Individual Impartiality
5. Tool Support
6. Team Strategies
7. Team Culture
8. Organization Support
9. Project Support
10.Source Code
11.Process Support
FACILITY CONDITIONS
ARTEFACT
Categories
Most Influential Knowledge Sharing Factors (Specific to MCR activities)
Knowledge Sharing Factors
Most Influential knowledge Sharing Factors (for all MCR activities)
Sub- Factors
MCR Activities
Categories to which all influencing factors belong

A Framework for Semantic Text Clustering

Soukaina Fatimi[1], Chama EL Saili[2], Larbi Alaoui[3]
TIC Lab, International University of Rabat
Sala Al Jadida
Morocco

***Abstract*—Existing approaches for text clustering are either agglomerative, divisive or based on frequent itemsets. However, most of the suggested solutions do not take the semantic associations between words into account and documents are only regarded as bags of unrelated words. Indeed, traditional text clustering methods usually focus on the frequency of terms in documents to create connected homogenous clusters without considering associated semantic which will of course lead to inaccurate clustering results. Accordingly, this research aims to understand the meanings of text phrases in the process of clustering to make maximum usage and use of documents. The semantic web framework is filled with useful techniques enabling database use to be substantial. The goal is to exploit these techniques to the full usage of the Resource Description Framework (RDF) to represent textual data as triplets. To come up a more effective clustering method, we provide a semantic representation of the data in texts on which the clustering process would be based. On the other hand, this study opts to implement other techniques within the clustering process such as ontology representation to manipulate and extract meaningful information using RDF, RDF Schemas (RDFS), and Web Ontology Language (OWL). Since Text clustering is an indispensable task for better exploitation of documents, the use of documents may be more intelligently conducted while considering semantics in the process of text clustering to efficiently identify the more related groups in a document collection. To this end, the proposed framework combines multiple techniques to come up with an efficient approach combining machine learning tools with semantic web principles. The framework allows documents RDF representation, clustering, topic modeling, clusters summarizing, information retrieval based on RDF querying and Reasoning tools. It also highlights the advantages of using semantic web techniques in clustering, subject modeling and knowledge extraction based on processes of questioning, reasoning and inferencing.**

Keywords*—Text clustering; similarity measure; ontology; semantic web; RDF; RDFS; OWL; reasoning; inferencing rules; SPARQL; topic modeling; summarization

I. Introduction

It has been a while since the web has changed from the web of documents to the web of data. Before knowing this upgrade, the information on the web was designed to be human-understandable only. Therefore a device or a robot could not access information in the same manner as humans, and artificial intelligence cannot evolve under these circumstances. This particular issue is considered as the motivation behind the evolution of information representation and the launch of the Semantic web as a web of connected data. The concept base is to transform the web of unstructured data to a network of interconnected chunks of information. Hence, both humans and machines can navigate between bits of data to explore it and retrieve more information from it. This collection of interrelated data is referred to as Linked Data [2]. The Linked Data is guided with a set of principles to allow easy sharing of structured data planet-wide. To represent and enable the use of this linked data and to allow the navigation between pieces of information, special representation should be used. The Resource Description Framework is at the core of the linked data paradigm. The RDF model, in which the data is represented as triples of interconnected subject and object with the intermediary of a predicate, is the mainstay of the interconnection of the information in the semantic web. This model enables the navigation between pieces of information following RDF links. It is indeed true that unstructured representation of information is still used, and that studies have provided significantly valuable tools for the manipulation tasks of textual documents, such as text clustering, information retrieval topic identification, etc. Still, the advantages of the linked data are captivating. Therefore, providing semantic data manipulation based on a semantic web model needs to be explored and strongly highlighted.

Text clustering has been widely explored for textual document manipulation. Yet proposed methodologies have lacked in the use of semantic relationships between words. Generally, documents are considered a bag of unrelated words and semantics are not explored in the process of text clustering.

Nevertheless, this work aims to use the semantic web approach for a semantic text clustering using graph-based representation model RDF with the respect of the linked data principles. We propose a system that is an integrated set of techniques in which the textual documents are transformed into an RDF graphs representation and divided into homogenous clusters based on a semantic clustering approach. These documents are further explored using semantic web techniques such as querying and information retrieval using inferencing and reasoning tools.

Text clustering is an indispensable task for better exploitation of documents to retrieve information, identify topics in more efficient ways. The provided system is a holistic approach allowing better understanding and use of textual documents with the mean of a semantic framework based on the RDF model. The purpose of working with RDF is due to its countless advantages, the self-explanatory or semantic characteristics of RDF data and is very beneficial for better semantic similarity computing and more efficient clustering.

We present an overall framework, and show how to apply machine learning techniques to mine textual documents using

This work is within the framework of the research project "Big Data Analytics - Methods and Applications (BDA-MA)". Authors S. Fatimi and C. El Saili are financially supported by a PhD grant of International University of Rabat.

Linked Data principles and highlight the importance of text clustering and the use of semantics in text clustering based on the RDF model.

The rest of the paper is organized as fellows. The next section introduces a review and presents the general context of our work, such as text clustering, semantic web, semantic similarity measurement, and topic identification. In the third section, the overall framework is presented and the steps of the system are discussed in the subsections emphasizing the clustering process. Finally, a conclusion and perspective work are given in the last section.

II. Related Work

The semantic oriented clustering approach that we are presenting in this paper is a combination of interesting concepts and techniques, from text clustering to similarity measurement and also to semantic web concepts and frameworks. In this section, we will give an overview of all the above notions and mention some of the related studies and works on these fields.

A. Semantic Web

The Semantic Web concept was introduced by Tim Berners-Lee as a novel form of web content that is understandable by humans [2]. The main goal of this concept is to interconnect and structure data in the World Wide Web to create an environment where programs can ramble between different pages to understand, process, and question existing information. Semantic web has caught the attention of many researchers ([3], [4], [5], [6]). Tim Berners-Lee introduced several principles for semantic web concept, he defined Resource Description Framework (RDF) a graph model to present data on the web, RDF interconnects data as triplets of a subject, predicate and object where subject and object are nodes and property is an arc. These RDF elements may be a textual value, or a blank and may be represented as Universal Resources Identifiers (URI) to distinct notions and relations that can connect them [2]. The use of RDF allows machines to understand the meaning of these notions and their linkage. This type of data is stored in special repositories called Triple Stores [7]. One other fundamental component of Semantic Web is ontology creation. Researchers have intensely studied this concept and its application in many domains like biomedical network security [8], smart cities [9] and robotic application [10]. Ontology is defined as a collection of information that describes a concept and provides its vocabulary. Ontologies are understandable by both humans and machines and allow semantics and syntactic exchange. Definite web ontology languages have been unified due to research in the Semantic Web that allows to efficiently describe a domain with the use of the semantic web languages RDF Schemas (RDFS), and Web Ontology Language (OWL). Ontologies are engineered based on the domain concepts referred to by "classes" and the relationships between these concepts which can be hierarchical as subclass relationships or predefined as properties. The models can also include constraints on the expressed information.

B. Text Clustering

Document clustering is an unsupervised learning process that separates documents into significant groups. It’s one of the main techniques of text mining [11]. Document clustering is to designate a corpus of content documents into distinctive bunches so that documents within the same gather depict the same subject. Clustering of documents contains three categories: partitioning methods, agglomerative and divisive clustering. Researchers proposed several document clustering algorithms like K-Means, Hierarchical Agglomerative clustering and Frequent Itemsets based clustering and more algorithms having been utilized in this learning process.

Text clustering is a dynamic field that caught the researcher’s consideration. The enormous textual data shared on the net is considered as a bag of information and can be labeled as the crude fabric of information. Diverse methods are actualized to move forward the extraction of profitable data from this information. Text clustering consists of indexing, crawling and filtering the information. We distinguish four steps within the process of text clustering: the collection of data, the preprocessing at that point, the clustering and the post processing of the clusters. Initially, documents are collected and put away, and it is basic to preprocess all these documents to dispose of the commotion [11] before clustering these documents.

The Internet is nowadays advancing from a Web of documents to a Web of Information, employing a graph-based representation and a set of basic standards, known Linked Data Principles [12]. In any case, statistics on the LOD Cloud (April 2017) reports that over 149 billion realities are as of now put away as RDF triples in 9960 information sources. Hence, the number of RDF datasets distributed on the Net is continuously and rapidly expanding. A client willing to use these datasets will begin to have to investigate them in order to decide which data is pertinent to his particular needs. Therefore, to encourage this interaction, a topical see of an RDF dataset can be given by applying the clustering instrument which can be characterized as making a set of homogeneous clusters with expansive intra-cluster similarity and expansive inter-cluster disparity [12].

C. Text Documents to RDF Triples

In order to handle documents of unstructured text data with semantic web techniques, it is obvious that the conversion of text documents to RDF triples is the major step to be done. The objective of this transformation is to change plain text into data units understandable by machines. Authors in [33] proposed another approach that converts a given content into RDF triples based on the semantic and syntactic structure of sentences. Based on this approach they built a system called T2R that creates important triples with all fundamental linguistic relations and semantic parts of the text. This approach can be used for any plain text. T2R inputs a text document into a syntactic parser using the Stanford tool and semantic parser utilizing the Senna tool.

LODifier [34] is one of the inspiring approaches in the Knowledge graph construction process to provide a tool for the conversion of texts to RDF. It is based on both deep semantic analysis and named entity recognition systems. Based on

LODifier, authors in [13] proposed a conversion of tweets into RDF triple where tweets are assembled topic-wise, by utilizing topic identification methods and shaping homogeneous clusters using the K-Means algorithm. Only the tweets containing named substances in DBpedia datasets are used. Each topic corpus is summarized then transformed into an RDF chart utilizing the LODifier tool. In the work [14], another model that collects resumes data from the internet and classifies them based on the cosine similarity measure was proposed. In this model, the data is represented using the semantic web like RDF based on Protégé tool and SPARQL. Another methodology is proposed by authors of [15] as a combination of the techniques of similarity computing, visualization procedures and RDF query language (SPARQL) to manipulate academic contents. They utilized also the ontological model to form syllabus information justifiable by both people and computers. Another framework for automatic knowledge graph (KG) extraction from unstructured text was proposed by authors of [16] and extended in their second work [17]. They underlined two RDF extraction steps. Firstly, candidate generation by focusing on the importance of mapping predicates to a referential KG for more searchability increase, and secondly, candidate selection process using pre-defined ontologies. Following the same path, [18] proposed an open-source platform for KG construction that includes graph management and downstream application support and is based on tools such as Stanford CoreNLP, Neo4j and Apache Solr.

D. Similarity Measures

The literature has many methods for computing the semantic similarity between terms. Semantic similarity measures can be classified into four categories: Edge Counting Measures, Information Content Measures, Feature Based Measure and Hybrid Methods. In the followings we present the overall idea behind similarity measurement and highlight the antecedent works about semantic similarity measure and its uses.

1) Similarity measure concept: A similarity measure is a function that assigns a non-negative real number to each pair of patterns, defining a notion of resemblance and having the target range between [0,1]. Similarity measures form the basis for many patterns matching algorithms. Besides that, similarity measures compare vectors which should be symmetrical and assign a value to them becoming larger when they are similar and getting the largest value when they are identical. Usually measured as the cosine of the angle between vectors, that is, the so-called cosine similitude, the Cosine similarity is one of the foremost well-known closeness measures in various data recovery applications and clustering as well. A Jaccard degree was introduced in [19] and is in some cases alluded to as the Tanimoto coefficient measures closeness between limited test sets and is characterized as the estimate of the intersection isolated by the estimate of the union of the test sets. For this measure, the Jaccard coefficient compares individuals for two sets to see which individuals are shared and which are unmistakable. The foremost AHC strategies do a calculation on this similarity matrix and develop a progressive structure to indicated connections or proximities among the data.

2) Semantic similarity measure: Research on the semantic similarity measures based on RDF data has mainly been done for the similarity measurement of RDF graphs for the query matching. The goal is to extract the best matching result. A similarity measure (gSemSim) was proposed to progress ordinary similarity measures to decrease their impediments. The notable feature of this semantic similarity measure is its capacity to display more reasonable similarity between concepts in the viewpoint of space information. Reference [20] demonstrates pairwise word interactions and displays a new similarity center instrument to recognize vital correspondences for superior similitude estimation. These thoughts are executed in a neural network design that illustrates state-of-the-art precision on three SemEval assignments and two reply determination tasks.

E. Semantic Text Clustering

Document clustering is one of the main techniques of text mining that is considered as an unsupervised learning process that separates documents into significant groups. It is to designate a corpus of content documents into distinctive bunches so that documents within the same gather depict the same subject. Researchers proposed several document clustering algorithms like Hierarchical Agglomerative clustering and Frequent Itemsets based clustering and others that are used in this learning process [21]. Traditional text clustering methods usually focus on the frequency of terms in documents to create connected homogenous clusters, thus, documents can be semantically related so these approaches will conduct inaccurate clustering results. The complexity of natural language results in the complexity of having accurate and efficient text clustering. Researchers have made use of semantic web technologies such as ontologies to take advantage of the semantic relationship between words in clustering. Walaa K. Gad and Mohamed S. Kamel [34] proposed a semantic similarity-based model (SSBM) to handle the semantic in documents. They incorporated the use of ontology in their case WordNet to obtain the semantic similarities between words, such as synonyms and hypernyms, and the documents vector is constructed based on a refined terms weight that includes term frequency (TF) and Inverse document frequency (IDF) and the semantic weight based on terms semantic relationships. Following the same path, authors in [35] applied text clustering Based on the semantic body for Chinese spam mail Filtering, the proposed methodology is based on lexical chains and HowNet semantic similarity to handle the words' synonyms, this technique helps to overcome defiance related to synonyms and near-synonyms by merging them. Thus, the results of the experiment were good, but the use of HowNet resulted in some limitations since it doesn't cover all possible similarities between words. [22] also presents an approach using lexical chains combined with WordNet; A WordNet-based semantic similarity measure for solving the problem of Polysemy and synonymy, and lexical chains to extricate a little subset of the semantic features which not as it denoted the topic of documents but moreover are advantageous to clustering. In [15] a combination of the techniques, methods, and algorithms such as cosine similarity, visualization procedures have been used for semantic text clustering, moreover, the ontological model to form syllabus information intelligible for both people and computers. In [12], using Candidate Description (CD) as a set of predicates, a form of RDF clustering algorithm has been developed, it used similarity matrix which contains the pairwise similarities between CDs clusters and utilized Cosine Similarity, Jaccard similarity and Sorensen Dice To measure the similarity between CDs.

F. Topic Modeling

Topic models are unsupervised machine learning techniques used to thematically describe a set of documents, it intends to detect the group of words that characterize and describe the collection of documents. Topic modeling is among important techniques used for the measurement of document similarity for classification [23], the clustering and cluster labeling, summarizing documents, and more [24]. Topic modeling was firstly introduced for textual documents. Yet, its use for unstructured types of data such as images has been explored. In multiple researches, topic modeling has been combined with semantic web [25] to improve the topic modeling results. However, few are the techniques that have been provided in order to apply topic modeling over unstructured topics. [26] Proposed a framework for applying topic modeling to RDF graph data based on LDA, they highlighted some of the major challenges in using topic modeling over RDF data. These challenges are related to the sparseness and the unnatural language of the RDF graphs and gave some methods to tackle it. In [27] a method to profile RDF datasets on Knowledge-based modeling techniques is given with the goal to describe the content of the datasets. The extracted representative topics for the RDF dataset are annotated with Wikipedia categories. Knowledge-based topic modeling has been earlier used for entity summarization in [36] using a probabilistic model called ES-LDA that uses a modified version of the LDA algorithm was used to handle the challenges of working with the RDF model.. The model uses prior knowledge for statistical learning techniques to create representative summaries for the large semantic web documents in order to facilitate the use of semantic web entities.

The whole approaches presented have not provided significantly valuable tools for the manipulation tasks of textual documents, such as text clustering, information retrieval topic identification, etc. Still, the disadvantages of the linked data are captivating. Therefore, providing semantic data manipulation based on a semantic web model needs to be explored and strongly highlighted. This work aims to take advantage of most of the semantic web techniques' benefits and present an overall framework for semantic text clustering based on RDF data more efficient than these approaches.

III. Methodology

Text is considered the essential and mostly utilized representation of data, numerous investigations and strategies have been examined to move forward the information disclosure based on textual information. The aim behind transforming textual data into an RDF model is to make it understandable by both humans and machines. The transformation should take into consideration syntactic and semantic relations between terms. The goal is to analyze, summarize, and extract information from this data. All these errands require a profound understanding of the basic structures and semantics of the documents. Exploring large amounts of data in order to retrieve relevant information can be a frustrating task. Based on the RDF framework and using associated techniques such as SPARQL for querying the data, RDF Schema (RDFS) and Web Ontology Language (OWL) to apply reasoning and inference support on the data.

Furthermore, clustering methods result in improving these interactions in order to provide better results and is considered as the pillar for other knowledge discovering tasks such as summarization and visualization. Classic clustering methods ignore the semantics between the words, generally, documents are considered as a bag of words, and do not make use of the relations that may exist between the words. Words can have multiple meanings depending on the context there are used in. Therefore, separating words from their context can lead to a misunderstanding of the words. The use of RDF based models for textual documents clustering is a step toward preserving semantics in documents and providing efficient clustering with better accuracy.

In this sense, and as previously mentioned, this work aims to take advantage of most of the semantic web techniques' benefits. Therefore, it proposes a graph data model for clustering and mining text documents. The model is based on the use of semantic web technology RDF to represent the information, SPARQL, RDFS and OWL to use reasoning engine in order to retrieve information from it.

The proposed methodology starts with the extraction of RDF representation from textual documents based on the semantic and syntactic nature of sentences, and then several mining techniques are introduced. This methodology focuses on clustering based on a proposed similarity measure of the RDF graphs, in order to group related documents in homogenous clusters, and topic modeling process to extract the underlying topics presented in the documents. Finally, an inferencing model is introduced based on RDFS and OWL language in order to extract more facts from the data.

Fig. 1 summarizes the proposed framework whose main components are explained in the subsequent.

A. Extract RDF from Data

The first step toward our semantic-based clustering system and documents querying is to transform the textual unstructured documents into RDF triples representation. As previously discussed in the above section, there have been many studies tackling the transaction from text to RDF triples. The main goal of this transformation is to switch from textual sentences that are understandable by humans only to interconnected information and intelligible by both humans and machines.

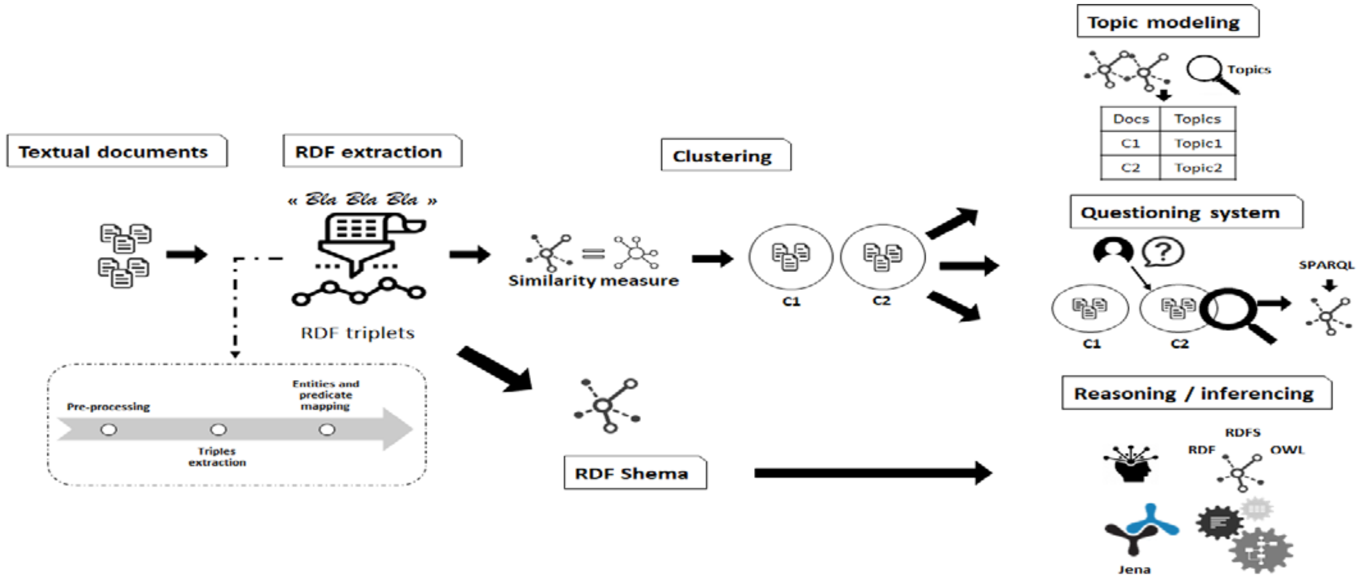


Fig. 1. Semantic-based Text Clustering Framework.

An RDF graph G is defined as a collection of statements. A statement is a triple (t) representing the relationship - named predicate (p) which is generally presented as a URI - between a subject (s) and an object (o) in the form of t=(s,p,o). A subject can be either a resource (URI or IRI) or a basic string (Literal), while an object can be denoted as a resource, a literal or an abstract identifier (Blank).

Overall, the task of text transformation to RDF is an iterative process that consists of converting each sentence into RDF triples under its semantic and syntactic form. This process can be represented by the schema given in Fig. 2.

The preprocessing phase is vital before addressing the triples extraction. Usually, sentence parsers that can be used for the triples extraction cannot handle some special case words, such as capital names, multiple word names which are considered as independent words and also the ambiguities in distinguishing named entities. These issues can be handled during the preprocessing phase to prepare the sentences for the RDF extraction. Another cause for concern is the multi-clause sentences, the parsing of these sentences will lead to shortage or false representation of the real meaning of the sentence. Isn this case, multi-clause sentences should be split into single-clause ones with the maintenance of the semantics in the original text.

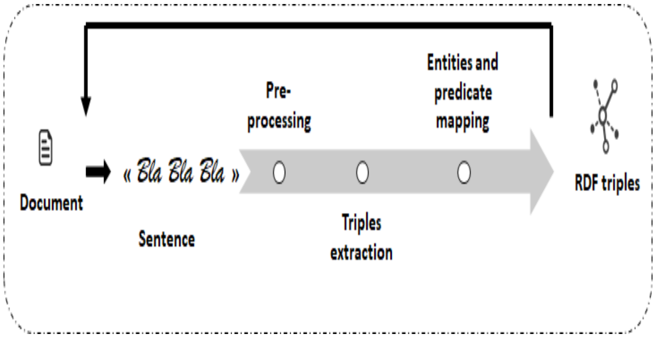


Fig. 2. RDF Extraction Process.

For each extracted single-clause sentence, the Stanford parser [28] can be used to analyze the grammatical structure of sentences, and in particular to identify the subject or object of a verb, in order to represent sentences in the form of the triple (subject, verb, object). Senna parser [29] can also be used to enrich the RDF extraction. Senna provides useful information, by allowing entity name recognition. It allows labeling of the named entities with given categories such as organizations, monetary value, person, etc. Its semantic role labeling can as well be used to enhance the discovery of the semantic sense of words in a given sentence.

Now that the triplets have been discovered, it is time to map to each extracted entity and predicate its Unified Resource Identifier (URI). DBpedia is a data set powered by Wikipedia articles that relates an entity to a Wikipedia article and provides a URI to identify it. In the mapping of RDF triples, using URIs provided by DBpedia. The use of DBpedia is due to the richness of the subject's details and the Multilanguage's description provided as well as the continuity of updates of the data sets. The identification of the most relevant meaning of words can be done based on the synsets provided by WordNet, which is a large-scale lexical database for English. After identifying the most relevant meaning of an entity based on its context and the syntactic and semantic role, it can be associated with its DBpedia URIs or WordNet URIs if it existed.

The named entity recognition Wikifier [30] can be used to obtain links to Wikipedia of the associated articles to the named entity. The following example illustrates the transformation from an unstructured text to an RDF graph.

Considering the sentence: "The WHO declared Covid-19 a pandemic". The Stanford parser will enable the tagging of this sentence (Table I).

Using WordNet to discover the appropriate sense of the words and select the named entity that corresponds to it in order to assign its DBpedia URIs/IRIs, Fig. 3 represents an RDF graphic representation of the sentence.

TABLE I. PART OF WORDS TAGGING USING STANFORD PARSER

	Word	POS		POS	Word
1	The	DT	4	COVID-19	VBN
2	WHO	NN	5	A	DT
3	declared	VBN	6	Pandemic	NN

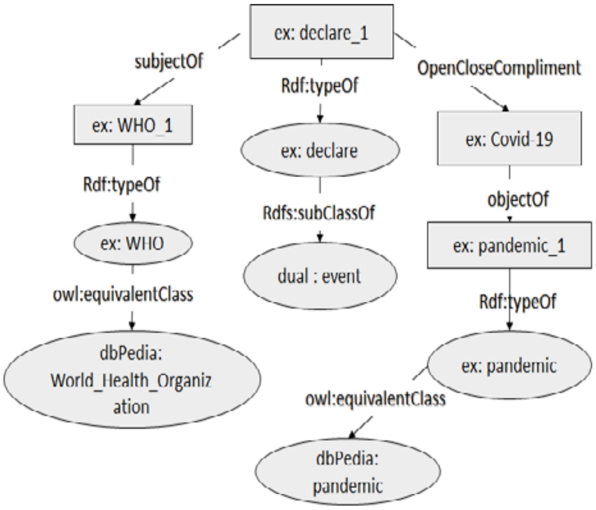


Fig. 3. Example of RDF Schema.

The extracted sets of triplets of each document not only allows the comprehension of documents for machines, but it will also be used in more sophisticated tasks such as document clustering, query answering and text summarization.

B. RDF based Clustering

After retrieving RDF triples for each text document we can proceed to the clustering of these documents. The clustering process consists of grouping documents in related clusters based on the similarity between them. In this case, the documents are represented using RDF graphs. The RDF triples in the graphs correspond to the document sentences, where each subject, object, or predicate are identified using a URI from the DBpedia datasets. In traditional clustering, the similarity between documents is calculated based on the text words considered as independent items. The use of RDF representation allows the incorporation of the semantic relationship of terms. However, the key to an efficient clustering is the use of a similarity measure that results in better matching between documents, not only based on relevant words with the highest strength or occurrence frequency in the documents as feature words for clustering but taking into consideration the semantic relationship between words and between documents. The extracted RDF graphs are loaded with semantic and syntactic information about the texts.

Our goal is to put forward a semantic similarity measure based on the RDF model with the exploitation of the semantic web tools. To tackle this issue, a similarity function based on RDF graph matching is going to be set up to compute the similarities between documents.

As earlier discussed, textual unstructured data is transformed into RDF graphs, the matching of two RDF graphs consists of the matching of their unitary elements which are the RDF triples and precisely it's about the calculation of the similarity between triples' subject, predicate, and object. Graph matching algorithms for RDF have been used for the matching of RDF queries and RDF graphs in order to implement searching processes or to put in place Linked Data recommendation processes. This study introduce the Graph matching algorithm to calculate the similarity between RDF graphs corresponding to a documents' dataset in order to perform clustering over these documents. In the following, we introduce and discuss the RDF graph distance computing in the clustering process.

1) Similarity computing between documents and clustering: We assume in the following that the previous step of RDF extraction is completed and that each document is represented as an RDF graph, where every RDF graph consists of a list of RDF statements.

The matching of two graphs can be translated to an assignment problem that would be solved using the Hungarian matching algorithm over a bipartite graph. The bipartite graph whose vertices are the set of triples of the two RDF graphs, each RDF's triples group is considered as an independent vertex of the bipartite graph, whereas the weight of the edges of the graph is the similarity measure between the triples.

Considering two documents D1 and D2, we represent these documents by two RDF graphs G1 and G2 respectively, and we define the bipartite graph BG as BG:=(U,V,E), U and V being the BG partition's parts such that U is the set of nodes related to the triples of the document D1, and V represents the triples of the second document D2. Moreover, E is stating the edges of the graph and the weight of these edges is the computed similarity measure between the nodes connecting the edges. The Hungarian matching algorithm is used over the BG to find the maximum similarity matching between the pairs of triples represented by U and V. Based on the matching result, the overall similarity measure can be computed between the two RDF graphs. This ability of measuring the similarity between a pair of documents yields to a similarity matrix based on the computed results. The computed similarity matrices can be used in multiple clustering techniques to provide homogenous clusters. This proposed framework allows therefore introducing an agglomerative hierarchical clustering to extract the clusters.

The following subsection discusses how we can obtain the similarity measure between a pair of triples.

2) Similarity computing between triples: As a result of the previous section, computing the similarity between triples is most crucial tasks in the clustering process that is to put in place since it can impact the clustering efficiency. As already mentioned, several methods consider the text as a dissociated bag of words, ignoring the semantics in texts. This is what we aim to tackle by handling unstructured textual data within the context of an RDF model in order to preserve the semantic relationships in the text, linking it to the important knowledge base in our case DBpedia and furthermore include a semantic similarity measure to compute the distance between RDF triples.

It is trivial that in order to compute the similarities between documents we need to measure the unitary similarity of the triples pairs. One of the major advantages of RDF representation of the unstructured textual documents is to be able to utilize the data values of resources to calculate the resources' similarity scores.

Considering two triples t1=(s1, p1, o1) and t2=(s2, p2 , o2) the similarity between t1 and t2, Sim_t (t1, t2), is related to the similarities between subjects Sim_s(s1,s2), between predicates Sim_p(p1,p2) and between objects Sim_o(o1,o2). Firstly, to compare words, in this case, it is essential to use a linguistic similarity measure based on a reliable source such as WordNet. Several researchers have tackled the use of WordNet in the similarity computing based on the provided synsets, and one of the most used formulae is Lin's similarity. Secondly, and in the case of a string value or not being able to find the word in WordNet database we can proceed to a string similarity measurement such as the normalized compression distance and the Levenshtein Distance [31]. Finally, for the URI form of data, if the corresponding value can be matched to a WordNet word then we could use the linguistic computing method, and if not the string similarity measurement could be used instead. The triple's object similarity will be handled in the same way as the subject, as for the predicate, we can consider the fact that is two triple's subjects (s1, s2) and objects (o1, o2) are similar then it is very likely that the predicates (p1, p2) are also similar, otherwise linguistic and string similarity computing methods can be used.

C. Topic Modeling

Clustering goal is to divide a collection of text documents into different category groups so that documents in the same category group describe the same topic. One of the major challenges related to clustering is cluster labeling and how to provide a clear description of the clusters. Therefore, we can note that for identifying and describing the constructed clusters Topic models can be used. In this case, clustering allows the inference of more coherent topics. A topic is a group of words that resume and refers to the content of the cluster. The identification of the cluster's topic allows a previous view over its content and eases the searching process. Topic models were firstly introduced for text documents and are easily adaptable to the case of RDF graphs. In [26] an approach to use topic modeling with RDF data was proposed using Latent Dirichlet Allocation (LDA) which is a commonly used model to identify the topics of documents. LDA aims to extract thematic information from documents' collections and it is based on the bag of words as vocabulary extracted from these documents. In our use case, the documents are the RDF Graph and the words are the extracted words from the graphs' triples. [26] introduced several limitations and challenges of using topic modeling for RDF graphs; Firstly, the sparseness of RDF data which means that even when having large datasets, the preprocessing of this data could result in a restricted set of words that could be used as a bag of words. Secondly, the lack of context is encountered since used words can have several meanings. In the case of RDF data, the context is hard to be determined due to unnatural language of data and/or to the sparseness of RDF graphs. The unnatural language is related to the graph representation, unlike sentence representation that enhances the understanding of the words, and finally, the short text problem which can be handled by either text supplementing or providing a modified version of the LDA algorithm. However, the strengths of our RDF graph representation process help overcoming these challenges. since textual documents were converted to RDF graphs, the use of semantic and syntactic parsing tools and the introduction of Dbpedia and WordNet synsets for efficient entity recognition based on the context of the sentence helps to tackle the issues related to unnatural language nature of RDF science the identified entities are based on the context of the documents, in order to identify the most relevant meaning of an entity. On the other hand, the RDF graph is enhanced with semantic role relations and Dbpedia classes allow overcoming the likely sparseness nature of RDF and text shortness problems.

D. Summarizing Clusters and Questioning System based on RDF Clustered Data

In order to enhance the exploration of RDF data and due to the big amount of RDF data and its complexity, RDF summarization was introduced to assist the understanding and use of this type of data. Summarization aim is to provide brief, concise, and significant information. Our framework goal is to make better use of the textual documents through a semantic text clustering system. Therefore the use of the RDF summarization techniques in the proposed framework is guided with the attention to improve the information extraction from the handle datasets. Hence, it is important to assist the queuing system based on RDF data since the extracted RDF graphs clusters can be significantly large resulting in a querying process that is extremely expansive with regards to resource and time.

Summarization can be used for various reasons or applications such as ontology extraction from RDF graphs, assisting users by providing graph visualization, and improving the querying process. In our case, we are basically interested in these applications related to the advancement of the querying task in many ways. In particular, indexing is when summary graphs are seen as an index for the larger RDF graph. In this case, a query is initially matched with the summary graph for finding equaling index nodes, and then the original graph is explored after detecting the matching nodes. Thus this process reduces the computation time and improves the querying task. To be noticed is that a summary will also help identify the best matching data partition to apply the querying when working with distributed systems.

There are multiple RDF summarization approaches, some include ontologies to handle the summarization of an RDF graph, and others can ignore their use and only work on the bare RDF graph. Based on some recent reviews such as [32], a RDF summary can either be compact information that contains the major meanings of the graph or can be a graph that is exploitable rather than the massive original graph. In [32] the existing summarization approaches can be classified based on multiple criteria. Among others we cite the input and output type, the purpose and the methods which can be structural, statistical, pattern-mining, or hybrid. In order to reach our goal structural quotient summaries for its wide applicability in the indexing and query answering tasks. Quotient summarization graphs are summary graphs where each summary node is connected to the IDs of the original graph nodes. These graphs can be obtained based on equivalence relations such as bi-similarity.

E. Reasoning using Jena Inference Engine

RDF schema (RDFS) allows defining and organizing RDF data vocabularies. In RDFS, the relationships between properties and resources are defined using RDF which offers a typing arrangement for RDF models. These relations are hierarchical like the notion of classes, subclasses, and properties, a huge amount of links between elements can be identified by specifying properties of classes and inheritance between classes, therefore RDF objects are considered as an instance of one or many classes and are specified with the class properties and parent class specifications. Many projects have incorporated the use of RDF(S) representation format such as Protégé, and Mozilla. Web Ontology Language enhances the describing properties and classes by providing an extended vocabulary. It allows for example expressing the cardinality of relations between classes, offers other assets such as equality and symmetry of properties, and so on [1]. These OWL characteristics result in more detailed ontologies allowing high performance in documents reasoning tasks. As we already mentioned semantic web concept is all about allowing both humans and machines to understand data, therefore data should be presented in a well-structured form and rules should be provided in a well-defined language in order to implement reasoning process and allows data to be shared onto the web [2]. Logically, notions are inferred from ontologies if they conform to their associated models; this process is referred to as reasoning. The clustering semantic-based framework proposed in this work can be enhanced by including a reasoning layer that allows deriving additional truth from the RDF graphs. Tools such as the Jena framework have been widely used to extract data from RDF graphs and OWL ontologies.

IV. Conclusion

This paper showed how semantic web techniques can be used for the textual documents clustering and exploration. It underlined some of the existing works of manipulating RDF data and got inspired from these to present a connected pipeline of semantic processes for the semantic text clustering based on RDF. The main contribution consists on presenting an overall framework for semantic text clustering based on RDF data modeling. This framework combines multiple techniques in order to get an efficient and accurate system, allowing exploring textual documents using machine learning techniques combined with semantic web principles. The system allows documents RDF representation, clustering, topic modeling and clusters summarizing as well as information retrieval using both RDF querying and reasoning tools. The aim is to take advantage of the semantic web in order to enhance the exploration of documents and enhance the use of semantics along the whole process. In future work we intend to validate our framework and improve it by choosing most relevant tools and techniques in each of the framework's steps. An experiment of the proposed approach on a real dataset will be further attacked.

References

[1] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, "OWL 2 web ontology language primer," https://www.w3.org/TR/owl-primer/, 2012.

[2] T. Heath and C. Bizer, Linked Data: Evolving the Web into a Global Data Space, 1st ed.; Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool, 2011.

[3] M. Noura, A. Gyrard, S. Heil, and M. Gaedke, "Automatic knowledge extraction to build semantic web of things applications," IEEE Internet Things J., vol. 6, no. 5, pp. 8447–8454, Oct. 2019.

[4] P. Ristoski and H. Paulheim, "Semantic Web in data mining and knowledge discovery: A comprehensive survey," J. of Web Semantics, vol. 36. Elsevier, pp. 1–22, Jan. 01, 2016.

[5] A. Gangemi, V. Presutti, D. Reforgiato Recupero, A. G. Nuzzolese, F. Draicchio, and M. Mongiovì, "Semantic web machine reading with FRED," Semant. Web, vol. 8, no. 6, pp. 873–893, Aug. 2017.

[6] P. Pauwels, S. Zhang, and Y. C. Lee, "Semantic web technologies in AEC industry: A literature overview," Automation in Construction, vol. 73. Elsevier B.V., pp. 145–165, Jan. 01, 2017.

[7] M. Trianes Torres, A. Sánchez Sánchez, M. Blanca Mena, and J. García, "Competencia social en alumnos con necesidades educativas especiales: nivel de inteligencia, edad y género," Rev. Psicol. Gen. y Apl. Rev. la Fed. Española Asoc. Psicol., vol. 56, no. 3, pp. 325–338, 2003.

[8] G. Xu, Y. Cao, Y. Ren, X. Li, and Z. Feng, "Network security situation awareness based on semantic ontology and user-defined rules for Internet of Things," IEEE Access, vol. 5, pp. 21046–21056, Aug. 2017.

[9] N. Komninos, C. Bratsas, C. Kakderi, and P. Tsarchopoulos, "Smart city ontologies: improving the effectiveness of smart city applications," J. Smart Cities, vol. 1, no. 1, pp. 31–46, Nov. 2016.

[10] G. Sarthou, R. Alami, and A. Clodic, "Semantic Spatial Representation: a unique representation of an environment based on an ontology for robotic applications," in Proc. Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics (SpLU-RoboNLP), pp. 50-60, Association for Computational Linguistics, 2019.

[11] G. S. Reddy, T. V. Rajinikanth, and A. A. Rao, "A frequent term based text clustering approach using novel similarity measure," Souvenir 2014 IEEE Int. Adv. Comput. Conf. IACC 2014, pp. 495–499, 2014.

[12] S. Eddamiri, E. M. Zemmouri, and A. Benghabrit, "An improved RDF data Clustering Algorithm," Procedia Comput. Sci., vol. 148, pp. 208–217, 2019.

[13] M. Achichi, Z. Bellahsene, D. Ienco, and K. Todorov, "Towards Linked Data Extraction From Tweets," https://hal.archives-ouvertes.fr/hal-01411403/document, 2016.

[14] A. M. Abirami, A. Askarunisa, R. Sangeetha, C. Padmavathi, and M. Priya, "Ontology based ranking of documents using Graph Databases: a Big Data Approach,": https://www.amrita.edu/icdcn/sangeetha-r.pdf, 2014.

[15] V. Saquicela, F. Baculima, G. Orellana, N. Piedra, M. Orellana, and M. Espinoza, "Similarity detection among academic contents through semantic technologies and text mining," in IWSW, pp. 1-12, 2018.

[16] N. Kertkeidkachorn and R. Ichise, "T2KG: An end-to-end system for creating knowledge graph from unstructured text," AAAI Work. - Tech. Rep., vol. WS-17-01-, pp. 743–749, 2017.

[17] N. Kertkeidkachorn and R. Ichise, "An automatic knowledge graph creation framework from natural language text," IEICE Trans. Inf. Syst., vol. E101D, no. 1, pp. 90–98, 2018.

[18] R. Clancy, I. F. Ilyas, J. Lin, and D. R. Cheriton, "Knowledge Graph Construction from Unstructured Text with Applications to Fact Verification and Beyond," in Proc. Second Workshop on Fact Extraction and VERification (FEVER), pages 39–46 Hong Kong, November 3, 2019. Association for Computational Linguistic.

[19] T. Tran, H. Wang, and P. Haase, "Hermes: Data Web search on a pay-as-you-go integration infrastructure," J. Web Semant., vol. 7, no. 3, pp. 189–203, 2009.

[20] H. He and J. Lin, "Pairwise word interaction modeling with deep neural networks for semantic similarity measurement," 2016 Conf. North Am.

Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. NAACL HLT 2016 - Proc. Conf., pp. 937–948, 2016.

[21] H. Patil and R. S. Thakur, "Frequent Term-Based Text Clustering Using Hidden Support," Proc. Int. Conf. on Recent Advancement on Computer and Communication, B. Tiwari et al. (eds.), Lecture Notes in Networks and Systems 34, vol. 34. Springer Singapore, 2018.

[22] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, “A semantic approach for text clustering using WordNet and lexical chains,” Expert Syst. Appl., vol. 42, no. 4, pp. 2264–2275, 2015.

[23] V. S. Anoop, S. Asharaf, and P. Deepak, “Topic modeling for unsupervised concept extraction and document ranking,” Adv. Intell. Syst. Comput., vol. 683, pp. 123–135, 2018.

[24] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, “An overview of topic modeling and its current applications in bioinformatics,” Springerplus, vol. 5, no. 1, 2016.

[25] L. Yao et al., “Incorporating knowledge graph embeddings into topic modeling,” 31st AAAI Conf. Artif. Intell. AAAI 2017, pp. 3119–3126, 2017.

[26] J. Sleeman, T. Finin, and A. Joshi, “Topic modeling for RDF graphs,” CEUR Workshop Proc., vol. 1467, pp. 48–62, 2015.

[27] S. Pouriyeh, M. Allahyaril, G. Cheng, H. R. Arabnia, K. Kochut, and M. Atzori, “R-LDA: Profiling RDF Datasets using knowledge-based topic modeling,” Proc. - 13th IEEE Int. Conf. Semant. Comput. ICSC 2019, pp. 146–149, 2019.

[28] “The Stanford Natural Language Processing Group.” https://nlp.stanford.edu/software/lex-parser.shtml.

[29] “SENNA,” https://ronan.collobert.com/senna/.

[30] J. Brank, G. Leban, and M. Grobelnik, “Annotating documents with relevant {Wikipedia} concepts,” Proc. Slov. Conf. Data Min. Data Wareh., p. 4 pages, 2017.

[31] K. Al-Khamaiseh, "A Survey of String Matching Algorithms," Int. J. of Engineering Research and Applications, vol. 4, Issue 7 (Version 2), pp.144-156, 2014.

[32] Š. Čebirić, F. Goasdoué, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou, and M. Zneika, “Summarizing semantic graphs: a survey,” The VLDB J., vol. 28, no. 3, pp. 295–327, 2019.

[33] K. Hassanzadeh, M. Reformat, W. Pedrycz, I. Jamal, and J. Berezowski, "T2R: System for Converting Textual Documents into RDF Triples;" in Proc. IEEE/WIC/ACM Int. Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013.

[34] Augenstein, I., S. Padó, and S. Rudolph, "Lodifier: Generating linked data from unstructured text," In ESWC, pp. 210–224, 2012.

[35] Q.-Y Zhang, P. Wang, and H.-J Yang, "Applications of Text Clustering Based on Semantic Body for Chinese Spam Filtering," J. of Computers, vol. 7, no. 11, pp. 2612-2616, 2012.

[36] Seyedamin Pouriyeh, Mehdi Allahyari, Krzysztof Kochut, Gong Cheng, and Hamid Reza Arabnia. Es-lda: Entity summarization using knowledge-based topic modeling. in Proc. of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), vol. 1, pages 316–325, 2017.

Comparitive Study of Time Series and Deep Learning Algorithms for Stock Price Prediction

Santosh Ambaprasad Sivapurapu
Department of Applied Mathematics
Liverpool John Moores University
London, Unites Kingdom

Abstract*—Stock Price Prediction has always been an intriguing research problem in financial domain. In the past decade, various methodologies based on classical time series, machine learning, deep learning and hybrid models which constitute the combinations of algorithms have been proposed with reasonable effectiveness in predicting the stock price. There is also considerable research work in comparing the performances of these models. However, from literature review, stems a concern, that is, lack of formal methodology that allows comparison of performances of the different models. For example, the lack of guidance on the generalizability of the time series models and optimised deep learning models is concerning. In addition, there is also a lack of guidance on general fitment of models, which can vary in accordance with forecasting requirement of stock price. This study is aimed at establishing a formal methodology of comparing different types of time series forecasting models based on like for like paradigm. The effectiveness of Deep Learning and Time-Series models have been evaluated by predicting the close prices of three banking stocks. The characteristics of the models in terms of generalizability are compared. The impact of the forecasting period on performance for various models are evaluated on a common metric. In most of the previous studies, the forecasting was done for the periods of 1 day, 5 days or 31 days. To keep the impact of volatility in the stock market due to various political and economic shocks both at international and domestic domains to the minimum, the forecasting periods of up 2 days for short term and 5 days for long term are considered. It has been evidenced that the deep learning models have outperformed time series models in terms of generalisability as well as short- and long-term forecasts.

Keywords—Time series; deep learning; ARIMA; VAR; LSTM; GRU; CNN 1D; genetic algorithm; Tree Structured Parzen Estimator (TPE)

I. Introduction

The stock markets and associated indexes are considered as one of the important economic indicators of the state. The movement of the stock price is considered as a representation of the confidence that businesses are entitled to. Likewise, a healthy stock market movement indicates a general positive confidence in the economy. A steady and upward increase of the stock price of a business indicates the progression of business in the right direction. And such circumstances provide businesses an opportunity to use stock market as a sustainable and economical source of raising capital for further investments and growth. This cycle of business investment and successive growth assuredly brings monetary benefits to investors.

The other category of market players who materialise the stock price movements into monetary benefits are the traders. Traders exploit the crests and troughs of the stock price movement by buying and selling the shares for profit, in addition to hedging their bets. Forecasting the stock prices allow traders to make an informed decision thus, optimising the trading strategies and in turn maximising the benefits.

The stock price movement is essentially a culmination of multitudinal events including human sentiments. The randomness in the sticker movement coupled with influence of numerous exogenous variables, which often represent global macro-economic events poses unique challenges in building reasonable predictive machine. However, the monetary benefits from the returns of stock market investments and the abundance of data availability, in addition to the technical advances in hardware acceleration motivates constant research in this domain. From the perspective of implementing the predictive model, there are problems associated with selection of the algorithms for effective accuracy with the available data, how far the forecasting period can be extended while the accuracy of prediction is within operational requirements and what kind of algorithms are generalizable. This research aims to solve this problem. The overall study is structured into following broad objectives.

Study the effectiveness in forecasting: Both classical time series and deep learning models will be trained using the same stock data and the measure of accuracy in forecasting the stock prices is compared. The feature sets used for training constitute the daily prices of the stock, the technical features of the respective stock and macro-economic indicators termed as exogenous variables henceforth.

Study the generalizability of the models: Both time series and deep learning models will be evaluated on different banking stocks and the effectiveness of the prediction will be compared to understand the generalizability of the models.

Study the impact of forecasting periods on accuracy: Both the time series and deep learning models will be trained and used to forecast the stock prices for varying periods. The variation of accuracy with the forecasting period will be studied for different models.

The rest of the document has been structured into Literature review, Methods and Techniques employed for this study and a discussion on results and scope of further research.

II. Literature Review

Traditionally fundamental analysis played very important role in predicting the long-term trend of stock prices. The financial ratios produced from the fundamental analysis, gives intuition of the stock movement[1]. The process of using fundamental variables to make stock trading decisions begins with Benjamin Graham, as early as 1928. There has been numerous research works investigating and shaping the modern fundamental analysis. The fundamental analysis is essentially based on the key attributes of the enterprise for example earnings-to-price yield, Profit /Earnings ratios, Total debt, book price ratios etc.

While the fundamental analysis is helpful for long-term investors, the requirements of traders are not satisfied or often appear overlooked by these studies. Technical analysis is alternate discipline of financial markets study that fills the gap left by fundamental analysis. The temporal fluctuations in the stock price gives earning opportunities for the traders. Technical analysis is the field of science that deals with predicting the temporal fluctuations to support the traders make informed statistical decisions. The ability of predicting the prices measures the success of traders' portfolio. Technical analysis involves study of the stock prices based on pattern matching, measures of various technical indicators which are mathematical derivates. All the approaches of the technical analysis are dependent on three basic principles of technical analysis, namely: a) Prices move in trends b) Volume goes with the trend c) A trend, once established tends to persist. [1], [2]. The mathematical intuition behind the approaches of technical analysis makes the field an ideal use case to apply various statistical and machine learning based predictive models.

In a well-respected study by Neftci et al. [3] technical analysis was done on closing prices of gold and trade-bills. The results of the technical analysis were statistically evaluated. It was concluded that there is a significant relationship between the moving average and the close price of the stock.

Brick et al. in [4], by two technical trading rules ,moving averages and trend line are validated. It was concluded, that these technical rules result in statistically significant earnings.

Regardless of the random walk theory which states that the successive price changes in stock prices are independent and identically distributed random variables, numerous studies have been published with reasonable success in predicting the stock prices, using various machine learning techniques like classical time series, contemporary machine learning and deep learning methods. Broadly there are two type of systems exploited in predicting the stock prices - a) Statistical methods b) Machine learning and AI based deep learning methods.

Classical time series models are statistical methods employing linear processes as predicting techniques, such as the autoregressive integrated moving average model (ARIMA), the autoregressive conditional heteroscedasticity (ARCH) models.

In the work "Stock Price Prediction Using the ARIMA Model", [5] Adebiyi et al. presented extensive process of building stock price predictive model using the ARIMA. The study used one and half of decade data of two stocks from two different stock exchanges across the world New York Stock Exchange and Nigerian Stock exchange. The data was used to build ARIMA models. Results obtained revealed that the ARIMA model has a strong potential for short-term prediction and it was envisaged that the model can compete favourably with existing techniques for stock price prediction. Although the ARIMA model built is able to forecast the prices reasonably well, there is no comparative evaluation of the models. This creates a gap in statistically validating the performance of the models.

In the study," An Effective time series analysis for stock trend prediction Using ARIMA Model for Nifty Midcap-50" [6] Uma B et al. have analysed the 5 years' worth of data of the top four stocks from National Stock Exchange, India. ARIMA model has been proposed as a favourable model reasonably identifying the trends and able to forecast the prices. However, there is significant mean absolute percentage error in the forecasts and there is a very good scope of improving the model by employing one step forecasting.

Mondal et al. in their work [7] studied the generalizability of the ARIMA model by training the model using twenty-three months of data related to fifty-six different stocks from range of sectors from National Stock Exchange, India. The performance of the model based on size of data and generalizability of the models to different sectors has been studied. It was concluded that, ARIMA model performed well for the stocks which have seen relatively lower standard deviation. In addition, it was evidenced that the changes in accuracy for different sizes of data is not significant. There are some key gaps that have been identified in this literature. The auto ARIMA is not always guaranteed to converge to a perfect order of differentiation, regression and differencing of the time series. Investigation of root node of the time series, auto correlations and respective correlated residuals optimise the convergence of the model. This methodology will be implemented in the current study and manual Arima will be compared with Auto Arima.

In their very recent work [8], Kumar et al. compared the performance of contemporary machine learning classifiers in recommending 'Buy' and 'Sell' for range of stocks. Support Vector Machine (SVM), Random Forest, K-Nearest Neighbours (K-NN) and Naïve Bayes classifies have been compared. It was concluded that Random Forest algorithm outperforms the other algorithms. However, with minimum training data, Naïve Bayes classifier outperformed Random forest classifier. It was also identified that performance of the models increased with addition of technical indicators as features.

However, prediction systems based on statistical methods have their own limitations as they require more historical data to meet statistical assumptions for example identifying the cyclic trends in the data, other statistical attributes like stationarity and causality. In addition to this, most of statistical models are univariate in nature and therefore, additional features that can impact the stock price will remain transparent to the model. These characteristics inevitably impacts the short to long terms predictions. In addition to this problem, the models will fail to generalise on stocks belonging to a similar sector.

With the advancements in the hardware technology and the advent of deep learning, there have been increasing attempts to apply deep learning techniques to stock price prediction. Neural Networks (NNs) are inherently data-driven, adaptive and non -linear methods with few prior assumptions than the linear models. In [9] Abdul et al. have compared the performance of statistical model ARIMA and Artificial Neural Network (ANN). Although the error in the forecast was within acceptable levels for both AIRMA and ANN, it is understood that, there is a tremendous scope of improvement in this study. There are various variations of ANN model, which can be applied and optimised for enhancing the performance of the model. Intuitively the performance of ANN model is expected to be superior to that of ARIMA.

In the study, [10] Kara et al. compared two non-linear classification techniques viz SVM and ANN to predict the direction of stock movement in Istanbul Stock Exchange using a decade worth of trading data. It was observed that both the SVM and ANN have performed well in predicting the direction, with ANN performing marginally better at 75.74% accuracy as against 71.52% of SVM. One of the major factors that has boosted the model performance while controlling the variance, is adding various technical indicators in the data. These technical indicators played key role in predicting the movement of price.

In the recent work [11] Ugur et al. trained a 2 Dimensional Convoluted Neural Network (2D CNN) network using Exchange Traded Fund (ETF) data of New York stock exchange. The 14 years' worth of daily historic stock prices and respective technical features have been used to create a sliding window tensor to train 2D CNN. It has been concluded that the 2D CNN model has outperformed other classical predictive models with Mean Absolute Percentage Error (MAPE) of 72.9%.

In a notable comparative study [12] three neural network models - time delay, recurrent, and probabilistic neural networks are compared. It was concluded that the Recurrent Neural Network (RNN) showed the best performance among other models due to its inherent capability of incorporating temporal dimensions of the data as a result of internal recurrence. One of the key improvements that can be made in this study are scoping in family of RNN networks and optimising these networks.

In the study, [13] Lee et al. compared the performance of the Seasonal ARIMA (SARIMA) model with Back Propagation Neural Network(BPNN), while predicting the weekly and monthly averages of Korean Stock price index (KOSPI). The SARIMA model provided more accurate forecasts for the KOSPI than the BPNN model does. This relative superiority of the SARIMA model over the BPNN model is pronounced for the mid-range forecasting horizons. However, the difference in forecasting accuracies of the two models was not found to be statistically significant. In addition, it appears inappropriate to model the stock price as a seasonal component. The seasonality could be attributed to the impact of confounding variables for example, periodical announcements by the enterprise.

The recent work [14] Torres D G et al. employed the state-of-the-art Recurrent Neural Networks in multivariate time series forecasting scenario , to predict the future sequence. Two variants of RNN – Long Short-term memory (LSTM) and Gated Recurrent Unit (GRU) have been applied to predict the next day close price of the Bitcoin data. The performance of LSTM and GRU are compared with ARIMA and ARIMA Dynamic regression. It has been concluded that there is no considerable difference in terms of prediction accuracy between LSTM and GRU. However, both deep learning models outperformed ARIMA models, as one step forward forecasting technique is not used. The results of the model indicate symptoms of overfitting in case of LSTM and GRU. Regularising the models and optimising them can lead to better performing models. There have been recent studies employing evolutionary algorithms in optimizing the deep learning network parameters and this has reasonably enhanced the predicting capability [15], [16].

In the very recent work [17] Chou J et al. employed a non-linear modelling technique - Least Squares Support Vector Regression (LSSVR) to predict the next day close prices of stocks of Taiwanese construction companies. This methodology has achieved far greater accuracy in predicting the close price of the stocks compared with contemporary models. In the current study, the sliding window technique will be used for deep learning models. The time series models will be modelled on one step ahead methodology.

Motivated by the success of deep learning algorithms, considerable work has been done in exploring the hybrid models for predicting the stock price and associated price movements.

The study [18] Pai et al. proposed hybrid models in predicting the stock price using ARIMA and SVMs. 50-day historic data of ten different stocks have been used to predict the close price of respective stock using one step ahead forecasting technique. In this hybrid model, ARIMA model was used to estimate the close price of stock price. The residuals of the ARIMA model were then calculated and passed to SVMs model to predict the errors. These predicted errors were used to correct the ARIMA predictions. The hybrid model has dominated the single ARIMA and single SVMs performance.

There have also been some approaches to integrate qualitative information with deep learning techniques for stock market forecasting. Yoshihara et al. in the work [19] exploited the textual information as input variable and predicted market trends based on RNN model combined with restricted Boltzmann machine (RBM) to investigate the temporal effects of past events.

III. Methods and Techniques

To make reasonable comparison, the context of prediction can be divided into short-term and long-term prediction. The short-term prediction is about forecasting the stock price up to the next 2 days and long-term forecasting is to predict the price up to 5 days.

A. Data

The data comprises daily historic prise of Lloyds Bank Group share for 10 years, from 01 January 2009 until 31 December 2019.The daily prices include the Open , High , Low and Close price. The close price is by definition bound with in the range of the three prices. There are a total of 2857 data points. In addition to the Lloyds Banking Group share, the data comprising daily historic price for the last 10 years, for two additional shares is selected – Barclays Bank plc and Royal Bank of Scotland plc. This data is selected to evaluate the generalizability of the models.

Technical indicators which statistically describe the movement of stocks are found to be very useful features in predicting the future price [10]. These technical indicators are calculated based on the Lloyds stock price features viz Open, Close, High and Low. Table I shows the mathematical formulae of these indicators.

Deep learning and Time series models are supervised learning algorithms and as such, these models will need training data in the form of learning features (Independent variables) and ground truth label or regression outcome (Dependent variable). The data sourced for this study is a time series data and so, the data is converted in to supervised learning format using custom data generators.

TABLE I. TECHNICAL INDICATORS OF STOCK

Simple 10-day moving average	$\frac{(C_t + C_{t-1} + C_{t-2} + \cdots + C_{t-10})}{10}$
Weighted 10-day moving average	$\frac{(n * C_t + (n-1) * C_{t-1} + (n-2) * C_{t-2} + \cdots + 1 * C_{t-10})}{\frac{n * (n-1)}{2}}$
Momentum	$C_t - C_{t-n}$
Stochastic K%	$\left\{\frac{C_t - LL_{t-n}}{HH_{t-n} - LL_{t-n}}\right\} * 100$
Stochastic D%	$\frac{\sum_{i=0}^{n-1} K_{t-i}\%}{n}$
RSI(Relative Strength Index)	$100 - \frac{100}{1 + \frac{\frac{\sum_{i=0}^{n-1} Up_{t-i}}{n}}{\frac{\sum_{i=0}^{n-1} Dw_{t-i}}{n}}}$
MACD (Moving Average Convergence Divergence)	$MACD(n)_{t-1} + \frac{2}{n} + 1 * (DIFF_t - MACD(n)_{t-1})$
A/D Oscillator	$\frac{H_n - C_{t-1}}{H_n - L_n} * 100$
CCI (Commodity. Channel Index)	$\frac{M_t - SM_t}{0.015 D_t} * 100$
Larry William's R%	$\frac{H_n - C_t}{H_n - L_n} * 100$

C_t is the closing price , L_t the low price, H_t the high price , U_{pt} the upward price change, D_{wt} the downward price change at time t.

DIFF: $EMA(12)_t - EMA(26)_t$, EMA is exponential moving average, $EMA(K)_t$: $EMA(K)_{t-1}$+$\propto *$ $(C_t - EMA(k)_{t-1}$; $\propto$ is the smoothing factor : 2/1+k where k is the time period of k day exponential moving average. $M_t = H_t + L_t + C_t/3$; $SM_t = (\sum_{i=1}^{n} M_{t-i+1})/n$;$D_t = (\sum_{i=1}^{n} |M_{t-i+1} - SM_t|)$.

B. Data Transformation

Each of the features used for training the models have different range leading to very wide feature space. Thus, the training of deep learning models is prone to swinging gradients. As part of the Data transformation step, the attributes of a dataset are normalized by scaling its values using the Min-Max normalisation technique. Mathematically the Min-max transformation can be represented as (1).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

C. Statistical Tests

The time series models like ARIMA have prior assumptions about the data, to ensure convergence to global minima, for example, the models assume that the time series should have no unit root is therefore stationary. Two types of statistical tests will be performed to satisfy these prior assumptions.

Augmented Dicky Fuller Test (ADFT) is used to identify the existence of the unit root for a time series component. If the time series has a unit root, then the time series under consideration can be categorised as a non-stationary series and the series will need to be converted to stationary series before applying statistical methods. If there is a unit root, then the series can be classified as non-stationary. However, lack of unit root doesn't make the series stationary [20].

Assume, a time series at any given time 't' can be represented as linear combination of value of the series at previous time step 't-1' and some error 'e' at the time t, then mathematically it can be represented as shown in (2).

$$y_{t\,=}\ \phi * y_{t-1} + e_t \tag{2}$$

if $\boldsymbol{\phi} = \boldsymbol{1}$, then the series has a unit root and will not be stationary. ADFT establishes a null hypothesis that the time series that is being tested will have unit root and is therefore non-stationary. Equation (3) represents mathematically the hypothesis test.

$$\Delta y_t = y_t - y_{t-1} = (\phi - 1) * y_{t-1} + e_t \tag{3}$$

Simple t-test will be done to check the probability of ϕ being equal to 1.

Granger Causality Test is used to identify the causal relationships between the feature vectors of a time series data. Assume y_t is the value of time series variable at time t and assuming, the time series is auto correlated, it can be represented as shown in (4).

$$y_t = \alpha_0 + \mu_t + \sum_{i=1}^{m} \alpha_i y_{t-i} \tag{3}$$

α_0 *is the time series constant*, α_i *is the linear* equation coefficients and μ_t *is the noise* in the time series.

Another time series x_t is said to be having granger causal relationship with time series y_t, if it can be linearly expressed as shown in (5).

$$y_t = \alpha_0 + \mu_t + \sum_{i=1}^{m} \alpha_i y_{t-1} + \sum_{j=1}^{m} \beta_j x_{t-j} \tag{5}$$

The null hypothesis of Granger causality test establishes that, two stationary time series components are statistically not causally related. At 95% confidence, if the p-value of the test is less than the significance level 0.05, the null hypothesis can be rejected. The pre-requisite of the Granger causality test is that the two series are stationary or co-integrated; otherwise the problem of 'spurious regression' might occur [21]. The Fig. 1 shows the summary results of ADF and Granger Causality test.

D. Modelling

The following time series and deep learning models are trained and evaluated-

- Time Series Models – Manual and Auto ARIMA, Vector Auto Regression.
- Deep learning Models – LSTM, GRU, CNN2D-LSTM and CNN1D.

1) ARIMA Models: ARIMA models forecasts the time series by modelling the predictor variable as a regressor. The time series is assumed to be a composition of three main components. The first component is Auto regressor (AR) which measures the correlation between an instance of the time series variable and the variable itself with a time lag. Equation (6) shows the AR component of the time series.

$$yt=c+\phi 1\ yt-1+\phi 2\ yt-2 +\cdots+\phi p\ y\ t-p+\varepsilon t \qquad (6)$$

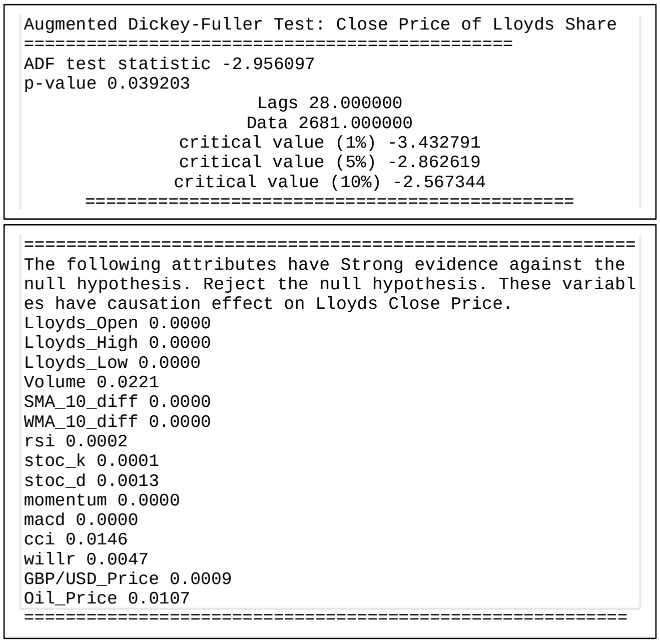
The above time series is also called the Auto regressive representation of the time series of order 'p'. $\varepsilon_{t.}$ is the error component in the time series. Just as linear regression, the algorithm learns the weights of the variables, while minimising the error.

The second component of the time series is the differencing which involves converting non-stationary series to a stationary series. A stationary series is a series which has constant mean and variance. The properties of the series are independent of the time. On the other hand, time series whose properties vary with time are called non-stationary series. Statistical tests, for example ADFT can be used to identify the stationarity of the time series. The process of differencing is finding the difference of the consecutive instance of variables. The differencing of the time series will stabilise the mean of the time series and therefore, the successive differencing will yield a stationary time series. The properties of a stationary time series viz. mean, variance and correlation help to accurately forecast the time series, as these properties do not vary with time. Thus, the stationarity is an important property of the time series for applying ARIMA model.

The third component of the time series is moving average (MA) component. The MA component represents the linear relationship of the predictor variable using the residuals of the forecast. Equation (7) shows the MA component of the time series.

$$Y_t=c+\varepsilon_t+\theta_1\ \varepsilon_{t-1}+\theta_2\ \varepsilon_{t-2}+\cdots+\theta_q\varepsilon_{t-q} \qquad (4)$$

ε is the residual of the forecast, technically called as 'white noise'. The above equation is called the moving average component of order 'q' of given time series. Just like the auto regression parameters, the model learns the weights 'θ' for optimising the errors in prediction.



```
Augmented Dickey-Fuller Test: Close Price of Lloyds Share
===============================================
ADF test statistic -2.956097
p-value 0.039203
                      Lags 28.000000
                     Data 2681.000000
               critical value (1%) -3.432791
               critical value (5%) -2.862619
              critical value (10%) -2.567344
      ===============================================
```

```
===========================================================
The following attributes have Strong evidence against the
null hypothesis. Reject the null hypothesis. These variabl
es have causation effect on Lloyds Close Price.
Lloyds_Open 0.0000
Lloyds_High 0.0000
Lloyds_Low 0.0000
Volume 0.0221
SMA_10_diff 0.0000
WMA_10_diff 0.0000
rsi 0.0002
stoc_k 0.0001
stoc_d 0.0013
momentum 0.0000
macd 0.0000
cci 0.0146
willr 0.0047
GBP/USD_Price 0.0009
Oil_Price 0.0107
==========================================================
```

Fig. 1. Summary of Statistical Tests.

For manual ARIMA, the auto regression and moving average components of the time series are calculated manually using the auto correlation and partial auto correlation plots. There is an extensive research and a standard methodology proposed for calculating these orders [22]. These guidelines are followed to manually derive the AR and MA orders for forecasting the Lloyds Share close price. The concept of the auto ARIMA is much similar to the manual ARIMA. In Auto ARIMA, the order of the linear equation and the associated weights (Regressor, Integrated and Moving Average) are determined programmatically using the Hyndman-Khandakar algorithm [23].

The goodness of the models in Auto ARIMA is measured by Akaike Information Criteria (AIC) and Bayesian Information Criteria(BIC). AIC and BIC measures are used to estimate the likelihood of a model forecasting the future variable. These indicators are the measure of a good fit of a model. Equations (8) and (9) shows the mathematical formulae of these measures.

$$AIC = - 2 \ln(L) + 2\ k \qquad (5)$$

$$BIC = - 2 \ln(L) + 2 \ln(N)\ k \qquad (6)$$

L is the value of likelihood; N is the number of observations and k is the number of estimates parameters.

2) Vector Auto Regression: Vector Auto Regression (VAR) is a statistical model used to capture the linear interdependencies among multiple time series. VAR models generalize the univariate autoregressive models (ARIMA models) by allowing more than one evolving variable [24].

VAR models require the features of time series variables to affect each other temporally. For Lloyds share close price prediction, there is a clear relation between the open, high and low price of the share. Mathematically, each variable is expressed as the linear combination of the variable itself at a time lag and the past values of the other variables at similar lags, as shown in (10) and (11).

$$y_{1,t}=c_1+\phi_{11,1}y_{1,t-1}+\phi_{12,1}y_{2,t-1}+e_{1,t} \quad (7)$$

$$y_{2,t}=c_2+\phi_{21,1}y_{1,t-1}+\phi_{22,1}y_{2,t-1}+e_{2,t} \quad (8)$$

The VAR model for ‘n’th cointegrated variable can be represented as a linear combination of past values of ‘n’th variable itself and the past values of all the other co-integrated variables. Algorithm learns the weights of the variables, while minimizing the error. Equation (12) shows the mathematical formulation of VAR.

$$Y_{n,t}=c_n+\phi_{n1,1}y_{1,t-1}+\phi_{n2,1}y_{2,t-1}+ \phi_{n2,1}y_{2,t-1} +\ldots\ldots\ldots..+\phi_{nn,1}y_{n,t-1} + e_{n,t} \quad (12)$$

3) Recurrent Neural Networks (LSTM and GRU): Long short-term memory (LSTM) and Gated recurrent Unit (GRU) belong to class of artificial recurrent neural network (RNN) architectures.

Fig. 2 shows the basic architecture of the RNN [15]. In addition to the dense connections between the layers of the network, the RNN networks have additional feedback connections within neurons. Thus, each neuron of the network receives the input vector which is the output of activation function of previous neuron and the output of the time step (feedback connection), thus allowing the network to learn temporal changes in the data. The output of a particular neuron ‘Y_t’ at any time ‘t’ can be represented mathematically as shown in (13).

$$Y_t = \Phi\ (W_x x_t + W_y y_{t-1} + b) \quad (9)$$

W_x are connection weights at time ‘t’ with respect to feedforward input, W_y is the connection weights at time ‘t’ with respect to feedback connection, b is the bias term. x_t is the feed forward input and y_{t-1} is the feedback input of the neurons. Φ is the activation function. The activation function will be normalised as part of optimisation of the performance of the network. One of the main disadvantages of the RNN is high susceptibility to vanishing and exploding gradients. Due to the number of weights involved and temporal connections within the neurons, the output of back propagation reduces drastically as the depth of network increases. Thus, the effect of back propagation fades away.

LSTM Networks are enhanced RNNs which works exactly in the same way as RNN, except that the neurons in the network are actually composite cells. Fig. 3 shows a common LSTM neuron, which is composed of an input gate, an output gate and a forget gate.

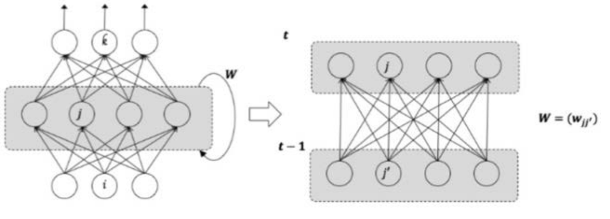


Fig. 2. Basic Architecture of RNN.

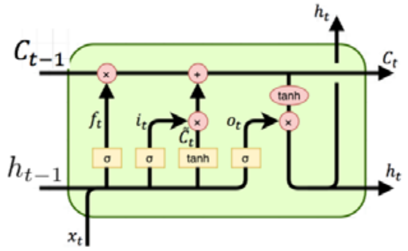


Fig. 3. LSTM Cell.

The input to each cell of LSTM at any time ‘t’ comprises of three signals, a long-term memory C_{t-1},a short-term memory h_{t-1} and x_t [25] . $f_{t.}$ represents the forget gate, $i_{i.}$ represents the input gate and o_t represents the output gate.

The forget gate controls, what part of the long-term memory is erased. The forget gate does a multiplicate operation on the long-term state and logistic activation function of short-term signal at previous time step and input to the cell at current time step. Equation (14) represents the backpropagation of the forget gate.

$$f_{t\,=}\,\sigma(W_{xf}^T x_t + W_{hf}^T h_{t-1} + b_f) \quad (10)$$

The input gate controls, what values of the current time state can be added to the long-term memory. This gate is the main gate which takes the current time state(x_t) and short memory state from previous time step(h_{t-1}), applies the respective activation functions and adds the result to the long-term state. There are two activation functions one being the logistic function and the other one is a ‘tanh’ function. These functions are used to feed in the temporal state of the input to long term memory. Equation (15) and (16) shows the inputs of the input gate.

$$\overline{C}_{t\,=}\,tanh(W_{xc}^T x_t + W_{hc}^T h_{t-1} + b_c) \quad (11)$$

$$i_{t\,=}\,\sigma(W_{xi}^T x_t + W_{hi}^T h_{t-1} + b_i \quad (12)$$

The output gate controls what part of the long-term memory should be added to the output of the cell (h_t = $y_{t.}$). Equation (17) represents the net outcome of the output gate.

$$o_{t\,=}\,\sigma(W_{xo}^T x_t + W_{ho}^T h_{t-1} + b_o) \quad (13)$$

Summarizing the output of each gates and applying the additive and multiplicative operations, the final output of the LSTM cell is shown using equations.

$$C_t = (C_{t-1} \otimes f_t) \oplus (i_t \otimes \overline{C_t}) \quad (14)$$

$$h_t = o_t \otimes \tanh(C_t) \quad (15)$$

$W_{xf}^T, W_{xo}^T, W_{xi}^T$ *and* W_{xc}^T are the weights of the four gates within the cell for the connections to the given input *x*.

$W_{hf}^T, W_{ho}^T, W_{hi}^T$ *and* W_{hc}^T are the weights of the four gates within the cell for the connections to the given short-term state.

b_f, b_i, b_c, b_o are the bias terms for the four gates.

C_t and h_t are long- and short-term memories respectively.

As evident from the mathematical equations of the LSTM, the number of trainable parameters increases exponentially as the depth of layer increases, slowing down the training time. LSTMs are also prone to overfitting due to the activations at forget gate.

GRU networks on the other hand are the simplified version of the LSTM networks.

Fig. 4 shows the GRU network cell. The cell consists of input gate and a forget gate. Both the states, the long -term and short-term memory, at any time *'t'* is merged into a single state h_t.

There is no output gate control and a single controller controls both input and forget gate. At any given time-step, the full state vector is output without any activation. Mathematically, the control equations are on similar lines to that of LSTM and represented as shown in (20), (21), (22) and (23).

$$z_t = \sigma(W_{xz}^T x_t + W_{hz}^T h_{t-1} + b_z) \quad (16)$$

$$r_t = \sigma(W_{xr}^T x_t + W_{hr}^T h_{t-1} + b_r) \quad (17)$$

$$\overline{h_t} = tanh(W_{xg}^T x_t + W_{hg}^T (r_t \otimes h_{t-1}) + b_g) \quad (18)$$

$$h_t = z_t \otimes h_{t-1} + (1 - z_t) \otimes \overline{h_t}_{-1} \quad (19)$$

$W_{xz,}^T W_{xr}^T$ *and* W_{xg}^T are the weights of the gates within the cell for the connections to the given input *x*.

$W_{hz,}^T W_{hr}^T$ *and* W_{hg}^T are the weights of the gates within the cell for the connections to the given short-term state (temporal feedback).

b_z, b_r *and* b_g are the bias terms for the gates.

$\overline{h_t}$, h_t , z_t, *and* r_t are the output states at respective gates.

GRU has a smaller number of trainable parameters compared to the LSTM. Due to these reasons, the training time of GRU is much smaller and the network is less susceptible to vanishing gradients and overfitting, compared with LSTM. In the recent work, [26] it was observed that the GRU and LSTM performance is reasonably same. In this study, python Keras deep learning APIs for LSTM and GRU will be used for model training and optimisation.

4) Convolution Networks: Convolution Neural Network (CNN) is class of deep learning neural network which can accept the data in a multi dimension matrix form called tensor and learns the intrinsic dependencies of the data.

The most important building block of CNN is a convolution layer. The convolution layer works on the principle of mathematical function called 'convolution'. Mathematically, convolution of two functions produces the third function, which explains how the shape of one function is modified by the second function.

$$f(t) * g(t) = \int_{-\infty}^{\infty} f(\tau)\, g(t-\tau) d\tau. \quad (20)$$

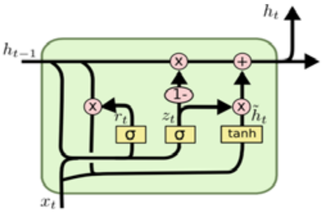


Fig. 4. GRU Cell.

Equation (24) shows two functions $f(t), g(t)$ which are convoluted. The convolution involves bounded integral of the product of two functions represented as a function of a temporal change τ.The resultant product is the convolution of the two functions. CNN applies these convolutions on set of pixels rather than each pixel to extract the abstract features from given data. The phenomenon of extracting the abstract features is called pooling.

The features extracted from each layer of the CNN is represented as feature map [27]. Multiple CNN layers extract the feature maps successively and the intricate dependencies within the features is learned by the model. Fig. 5 shows the architecture of CNN.

1D CNN is the extension of the CNN architecture, where the network can accept one dimensional sequence of the data and feature extraction is done using the convolutions. The extracted abstract features are used for forecasting the output. Therefore, this architecture of CNN can be used for time series forecasting problems. Several 1d convolution and pooling layers, stacked each other, makes the network powerful to extract the hidden dependencies within the temporal data. A filter of size *(1 x m)* is convoluted on a time step of size *(1 x n)*. Each stride of the filter extracts abstract feature from the time step, creating a feature map. Such feature maps are pooled through successive convoluted layers for forecasting the time series.

1D CNN stacked LSTM network is a hybrid network used to forecast the temporal data. Fig. 6 shows 1D CNN+ LSTM network. The network architecture consists of a convolution layer, which accepts the input as a tensor and performs convolutions using appropriate strides. The convoluted feature map is fed to LSTM. Deep LSTM network learns the convoluted feature maps and forecasts the time series. Therefore, 1D CNN is used as feature extractor and LSTM is used as sequence predictor.

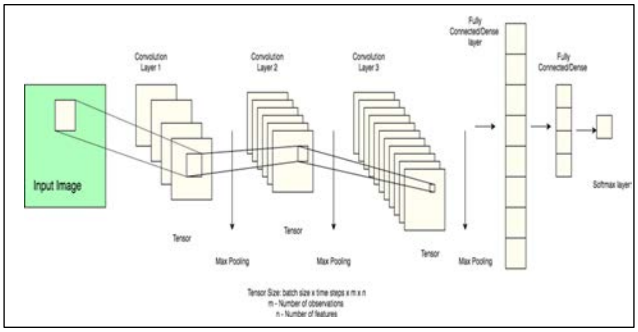


Fig. 5. General CNN Architecture.

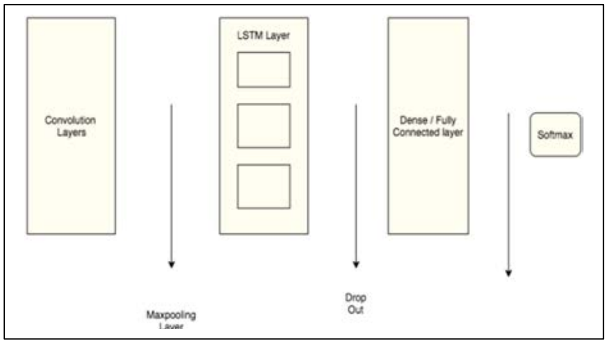


Fig. 6. CNN and LSTM Network Architecture.

5) Performance optimisation: All the deep learning networks mentioned above have respective hyper parameters. These hyper parameters will be optimised using Genetic algorithm [15], [16] and Tree-structured Parzen Estimator Approach (TPE Approach) [28].

Genetic algorithm will be used to identify the optimal number of time steps that algorithms use for learning the sequence. This is the temporal dimension that the model should consider for forecasting. TPE optimization is used for optimising the rest of hyper parameters viz. batch size of the inputs, which is the number of observations that network will learn for adjusting the weights following feedforward and back propagation cycles, number of neurons with in each layer of the network, the activation functions within each dense layers, the size of drop out, learning parameter and learning optimizer function.

Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE) will be used to evaluate and select the models. Percentage errors have the advantage of being scale-independent, and so are frequently used to compare forecast performance across different data sets. These errors put a heavier penalty on positive errors than on negative errors.

Assume, y_{pred} is the predicted value of a time series and y_{actual} is the actual value. Then, MAPE and MAE are mathematically represented as

$$\text{MAPE} = \text{mean}\left\{\frac{100 * ||(y_{actual} - y_{pred})||}{y_{actual}}\right\} \tag{21}$$

$$\text{MAE} = \text{mean}\left(\left|\left|y_{actual} - y_{pred}\right|\right|\right) \tag{22}$$

The best performing models from time series and deep learning classes will be compared based on following characteristics:

- The measure of error (MAPE) in forecasting the short-term prices and long-term errors.
- The measure of error (MAE) in forecasting the short-term prices and long-term errors.
- The change in the measure of error when exogenous variables are added to the model.
- The ability to generalize and perform on peer stocks.

IV. Results

As part of this study, the performance of the models while forecasting close price of Lloyds share is measured. The length of forecasting period is limited to 5 days, to avoid the impact of volatility due to sudden political or economic shocks within the share market. However, to study the general model performance on the test data and to support residual analysis, the size of test data is set to higher value than maximum forecasting period.

The total observations are split in 90%:10% ratio as training and test data sets respectively. This ratio yields 2439 as training observations and 271 as test observations.

The test observations are further split into two equal sets. These sets are used for cross validation and out of sample tests. This methodology allows having equal test size data sets for time series and deep learning models. Table II shows the data set sizes.

For each of the time series and deep learning models, metrics are recorded against each characteristic, and the conclusions are drawn based on these. Table III shows the performance of the time series models.

In terms of execution time, all the models performed roughly on same scale, with VAR marginally running for longer duration. While ARIMA based models builds the linear relationship between stock price and itself at different lags, the VAR model builds linear equation based on intrinsic relationship between each of the explanatory features in addition to the dependencies within the features at different time lags. This results in higher execution time compared to other linear models.

Based on the MAE and MAPE metrics, ARIMA models stands out. The manual ARIMA has performed marginally better when compared with auto ARIMA. In case of auto ARIMA, the orders of three components of ARIMA model viz AR, I and MA are calculated programmatically as (2,1,0). However, based on the heuristic methodology, as part of manual ARIMA modelling, the order (1,1,0) has been used to train the model. It is evident that the over differencing of the time series resulted in loss of performance, in case of auto Arima.

Table IV shows the performance of the deep learning models. The training time plays a crucial role in predicting the close price of the stock. In commercial context, a very high execution time, meant that there is loss of opportunity and the model predictions might not be relevant to the day and time of trade. The training time of CNN1D+LSTM is exceptionally high nearing 22hrs on an Nvidia Tesla K80, 4 core CPU. Therefore, the prediction of the model loses business justification and is practically not feasible. Same is the case with LSTM model with training time close to 20hrs.

TABLE II. Technical Indicators of Stock

Complete Dataset	2710
Training Dataset	2439
Cross Validation Dataset	136
Test Data Set	135

TABLE III. METRICS OF TIME SERIES MODEL

Parameter	**Manual ARIMA**	**Auto ARIMA**	**VAR**
Training Time	0:00:03	0:00:47	0:01:23
Training MAE	21.029	1.029	1.032
Training MAPE	31.708	1.718	1.718
Test MAE (2 days)	1.684	0.491	0.565
Test MAPE (2 days)	2.863	0.884	1.016
Test MAE (5 days)	1.194	0.590	0.657
Test MAPE (5 days)	2.039	1.058	1.177

TABLE IV. METRICS OF DEEP LEARNING MODEL

Parameter	**LSTM**	**CNN1D + LSTM**	**GRU**	**CNN1D**
Training parameters	103,937	634,049	385,409	18,465
Training Time[*]	19:36:42	21:52:13	11:45:47	7:34:37
Training MAE[τ]	0.0155	0.0007	0.0011	0.0005
Test MAE (2 days)	0.008	0.007	0.0040	0.0110
Test MAPE (2 days)	2.110	1.901	0.9850	2.907
Test MAE (5 days)	0.005	0.004	0.0040	0.007
Test MAPE (5 days)	1.413	1.021	1.1760	1.807

[*] The training time includes the time taken to optimise the network using Tree Parzen Estimator random search and genetic algorithm-based optimisation techniques. The training time depends on the total number of parameters used for training.

[τ] The networks are trained using multiple epochs to reduce the bias. Therefore, the training MAE is the average value of all the epochs.

Considering the training and test errors, CNN1D has recorded higher test error compared with mean training error. This is an overfitting characteristic and the model is highly prone to bias and variance. In terms of MAE and MAPE the GRU has performed marginally better compared with CNN1.

Based on the out of sample evaluation of the models, Manual ARIMA and GRU models are selected as favourable forecasting models.

The RNN – GRU and manual ARIMA models are trained and evaluated on RBS and Barclay's stock data. Table V shows the evaluation metrics of RNN—GRU and ARIMA models. Deep learning models outperform the ARIMA model in predicting the short-term prices. ARIMA model, on the other hand is able to generalise the stock price towards the tail end of the forecasting period with no significant betterment over the deep learning models.

TABLE V. DEEP LEARNING VS TIME SERIES MODEL

Stock	Model	MAE_1_Day	MAE_5_Day
LBG	RNN GRU	0.004	0.004
LBG	Manual ARIMA	1.402	1.004
Barclays	RNN GRU	0.027	0.014
Barclays	Manual ARIMA	3.66	2.783
RBS	RNN GRU	0.038	0.028
RBS	Manual ARIMA	5.736	5.552

V. CONCLUSION

Table VI shows the comparison of residuals of the optimised GRU with that of contemporary research papers. The optimised deep learning model has shown superior performance in predicting the close price and is highly viable model from commercial context. Deep learning models outperformed time series models for both short- and long-term forecasts. Fig. 7 shows the performance of the GRU and Manual ARIMA models in forecasting the long term and short-term stock price.

TABLE VI. COMPARISON OF OPTIMISED MODEL AND PREVIOUSLY PUBLISHED MODELS

Authors	**Algorithm/Models**	**Metric and Error**	**Optimised RNN – GRU trained as part of this study**
Haider Khan Z, Alin A S, Hussain T	ANN	MAE (for 1-day forecast)- **0.0174**	MAE (for 1-day forecast)- **0.002**
Torres Douglas, Qiu Hongliang	RNN - LSTM and GRU	RMSE (for 1-day forecast) ARIMA - **147.6** LSTM - **518.6** GRU - **396.4**	RMSE for 1-day forecast – **0.004**
Chung Hyejung, Shin Kyung Shik	GA Optimised LSTM	MAPE (1-day forecast) – **0.91**	MAPE for 1-day forecast – **0.592**
Patel Jigar, Shah Sahil, Thakkar Priyank, Kotecha K	Hybrid Models - Fusion of SVR, ANN and RF	MAPE (5-day forecast) SVR–ANN - **11.2** SVR–SVR - **2.41** SVR–RF - **11.31**	MAPE for 5-day forecast – **1.176**
Lee Kyungjoo, Jin John Jongdae	SARIMA and ANN	MAE (7 – day forecast) ANN - **1.110** SARIMA - **1.927** ANN-SARIMA – **0.742**	MAE for 7-day forecast – **0.004**

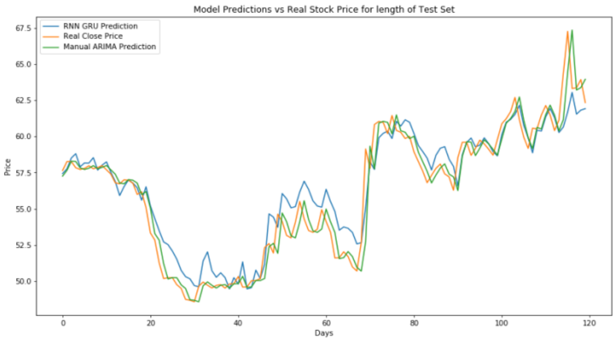


Fig. 7. Performance of ARIMA and GRU Models.

The approach and the experiments adopted in this study, reasonably answered the objectives of the research viz. comparing the performance of Time Series and Deep learning models, study the generalizability of the models and behaviour of the models for predicting long term forecasts. The generalizability of the deep learning models is superior to that of the time series models. This is particularly true for banking sector, in which the stocks appear to be correlated with confounded variables. With the increase in the length of forecast, the deep learning models can generalise the models well compared with time series models. However, there is a definitive length of the forecast during which this behaviour is observed. Beyond this forecast length, the volatility in the market outweighs the model behaviour.

However, there are some areas where, further research can be applied. Vector Auto Regression which can regress the time series data by modelling the series as a linear combination of the series itself at different lags and the respective cointegrating time series, has performed poorly compared with univariate time series models. This can be attributed to variance in the unit root of the cointegrating time series viz. technical and economic indicator. Further research can be done to understand the nuances of VAR and explore the limitations and scope of improvements. In addition, the sliding window technique used for training the deep learning models can be applied to linear time series models viz. ARIMA and VAR to construct the temporal dimension to enhance the performance. There has been appreciable research in the area of reinforcement learning for forecasting the movement of stock price. Research in measuring and characterising the reinforcement algorithms by comparing them with deep learning algorithms will greatly help the knowledge base and commercial applications.

ACKNOWLEDGMENT

I would like to thank my mentor Dr.Manoj Jayabalan, depart of Applied Mathematics, Liverpool John Moors University, London, United Kingdom, M.Jayabalan@ljmu.ac.uk for the constant guidance, encouragement, valuable feedback and evaluation of this study.

REFERENCES

[1] B. Vanstone and G. Finnie, “An empirical methodology for developing stockmarket trading systems using artificial neural networks,” Expert Syst. Appl., vol. 36, no. 3 PART 2, pp. 6668–6680, 2009, doi: 10.1016/j.eswa.2008.08.019.

[2] M. J. Pring, Technical Analysis Explained. 2014.

[3] S. N. Neftci, “Naive Trading Rules in Financial Markets and Wiener-Kolmogorov Prediction Theory: A Study of ‘Technical Analysis,’” J. Bus., vol. 64, no. 4, p. 549, 1991, doi: 10.1086/296551.

[4] W. Brock and J. Lakonishok, “American Finance Association Simple Technical Trading Rules and the Stochastic Properties of Stock Returns Author (s): William Brock , Josef Lakonishok and Blake LeBaron Source : The Journal of Finance , Vol . 47 , No . 5 (Dec ., 1992), pp . 1731-1764,” J. Finance, vol. 47, no. 5, pp. 1731–1764, 1992.

[5] A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo, “Stock price prediction using the ARIMA model,” Proc. - UKSim-AMSS 16th Int. Conf. Comput. Model. Simulation, UKSim 2014, no. June, pp. 106–112, 2014, doi: 10.1109/UKSim.2014.67.

[6] Uma B, S. D, and A. P, “An Effective Time Series Analysis for Stock Trend Prediction Using ARIMA Model for Nifty Midcap-50,” Int. J. Data Min. Knowl. Manag. Process, vol. 3, no. 1, pp. 65–78, 2013, doi: 10.5121/ijdkp.2013.3106.

[7] P. Mondal, L. Shit, and S. Goswami, “Study of Effectiveness of Time Series Modeling (Arima) in Forecasting Stock Prices,” Int. J. Comput. Sci. Eng. Appl., vol. 4, no. 2, pp. 13–29, 2014, doi: 10.5121/ijcsea.2014.4202.

[8] I. Kumar, K. Dogra, C. Utreja, and P. Yadav, “A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction,” Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2018, no. Icicct, pp. 1003–1007, 2018, doi: 10.1109/ICICCT.2018.8473214.

[9] S. A. Abdul-Wahab and S. M. Al-Alawi, Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction Ayodele, vol. 17, no. 3. 2014, pp. 219–228.

[10] Y. Kara, M. Acar Boyacioglu, and Ö. K. Baykan, “Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange,” Expert Syst. Appl., vol. 38, no. 5, pp. 5311–5319, 2011, doi: 10.1016/j.eswa.2010.10.027.

[11] M. Ugur Gudelek, S. Arda Boluk, and A. Murat Ozbayoglu, “A deep learning based stock trading model with 2-D CNN trend detection,” 2017 IEEE Symp. Ser. Comput. Intell. SSCI 2017 - Proc., vol. 2018-Janua, no. November, pp. 1–8, 2018, doi: 10.1109/SSCI.2017.8285188.

[12] E. W. Saad, D. V. Prokhorov, and D. C. Wunsch, “Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks,” IEEE Trans. Neural Networks, vol. 9, no. 6, pp. 1456–1470, 1998, doi: 10.1109/72.728395.

[13] K. Lee and J. J. Jin, “Neural Network Model Vs. Sarima Model in Forecasting Korean Stock Price Index (Kospi),” Issues Inf. Syst., vol. 8, no. 2, pp. 372–378, 2007.

[14] D. G. Torres and H. Qiu, “Applying Recurrent Neural Networks for Multivariate Time Series Forecasting of Volatile Financial Data,” pp. 1–10, 2018.

[15] H. Chung and K. S. Shin, “Genetic algorithm-optimized long short-term memory network for stock market prediction,” Sustain., vol. 10, no. 10, 2018, doi: 10.3390/su10103765.

[16] G. G. Szpiro, “Forecasting chaotic time series with genetic algorithms,” Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top., vol. 55, no. 3 SUPPL. A, pp. 2557–2568, 1997, doi: 10.1103/physreve.55.2557.

[17] J. Chou, “Forward Forecast of Stock Price Using,” IEEE Trans. Ind. Informatics, vol. 14, no. 7, pp. 3132–3142, 2018, doi: 10.1109/TII.2018.2794389.

[18] P. F. Pai and C. S. Lin, “A hybrid ARIMA and support vector machines model in stock price forecasting,” Omega, vol. 33, no. 6, pp. 497–505, 2005, doi: 10.1016/j.omega.2004.07.024.

[19] A. Yoshihara, K. Fujikawa, K. Seki, and K. Uehara, “Predicting the Trend of the Stock Market by Recurrent Deep Neural Networks,” Pacific Rim Int. Conf. Artif. Intell., pp. 1–11, 2014.

[20] A. Pal and P. Prakash, Practical Time Series Analysis. 2017.

[21] X. Wang, “A Granger causality test of the causal relationship between the number of editorial board members and the scientific output of

universities in the field of chemistry," Curr. Sci., vol. 116, no. 1, pp. 35–39, 2019, doi: 10.18520/cs/v116/i1/35-39.

[22] R. Nau, "Summary of rules for identifying ARIMA Models," Duke University, 2019. [Online]. Available: http://people.duke.edu/~rnau/ arimrule.htm.

[23] Rob J. Hyndman and Yeasmin Khandakar, "Automatic Time Series Forecasting: The forecast Package for R," J. Stat. Softw., vol. 27, no. 3, p. 22, 2008.

[24] Rob J Hyndman and George Athanasopoulos, Forecasting: Principles and Practice, 2nd ed. Sydney, Australia: OTEXTS, 2018.

[25] Aurelien Geron, Machine -Learning with Scikit-Learn and TensorFlow, First Edit. California: O'REILLY, 2017.

[26] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," IEEE Trans. Neural Networks Learn. Syst., vol. 28, no. 10, pp. 2222–2232, 2017, doi: 10.1109/TNNLS.2016.2582924.

[27] A. Chernodub and D. Nowicki, "Orthogonal permutation linear unit activation function (OPLU)," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9887 LNCS, no. July 2018, pp. 533–534, 2016, doi: 10.1007/978-3-319-44781-0.

[28] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," Adv. Neural Inf. Process. Syst. 24 25th Annu. Conf. Neural Inf. Process. Syst. 2011, NIPS 2011, pp. 1–9, 2011.

Enhanced Insertion Sort by Threshold Swapping

Basima Elshqeirat[1], Muhyidean Altarawneh[2], Ahmad Aloqaily[3]
Department of Computer Science, the University of Jordan, P.O Box 11942, Amman, Jordan[1, 2]
Computer Science Department, Maharishi International University, Lowa Fairfield, IA 52557, U.S.A[2]
Department of Computer Science and its applications, the Hashemite University, P.O. Box 150459, Zarqa 13115, Jordan[3]

***Abstract*—Sorting is an essential operation that takes place in arranging data in a specific order, such as ascending or descending with numeric and alphabetic data. There are various sorting algorithms for each situation. For applications that have incremental data and require e an adaptive sorting algorithm, the insertion sort algorithm is the most suitable choice, because it can deal with each element without the need to sort the whole dataset. Moreover, the Insertion sort algorithm can be the most popular sorting algorithm because of its simple and straightforward steps. Hence, the insertion sort algorithm performance decreases when it comes to large datasets. In this paper, an algorithm is designed to empirically improve the performance of the insertion sort algorithm, especially for large datasets. The new proposed approach is stable, adaptive and very simple to translate into programming code. Moreover, this proposed solution can be easily modified to obtain in-place variations of such an algorithm by maintaining their main features. From our experimental results, it turns out that the proposed algorithm is very competitive with the classic insertion sort algorithm. After applying the proposed algorithm and comparing it with the classic insertion sort, the time taken to sort a specific dataset was reduced by 23%, regardless of the dataset's size. Furthermore, the performance of the enhanced algorithm will increase along with the size of the dataset. This algorithm does not require additional resources nor the need to sort the whole dataset every time a new element is added.**

Keywords—Sorting; design of algorithm; insertion sort; enhanced insertion sort; threshold swapping

I. Introduction

Sorting is considered as one of the fundamental operations and extensively studied problems in computer science. It is one of the most frequent tasks needed mainly due to its direct applications in almost all areas of computing. The various applications of sorting will never be obsolete, even with the rapid development of technology, sorting is still very relevant and significant [1]. Formally any sorting algorithm will basically consist of finding a permutation or swapping of elements of a dataset (typically as an array) such that they are organized in an ascending (or descending) or lexicographical order (alphabetical value like addressee key). A large number of efficient sorting algorithms have been proposed over the last ten years with different features [2].

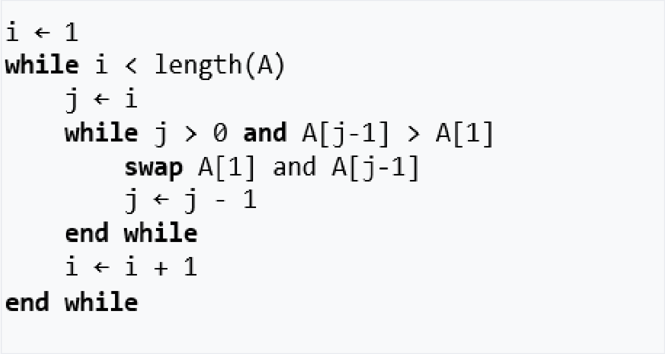
In this paper, we will consider the insertion sort (IS) algorithm, which is one of the popular and well-known sorting algorithms. It is simply building a sorted array or list by sorting elements one by one. The IS algorithm begins at the first element of the array and inserts each element encountered into its correct position (index), after determining and locating a suitable position. This process is repeated for the next element until it reaches the last element in the dataset. Fig. 1 illustrates a classical procedure of the insertion sort algorithm where A is an array of elements. The main side effect of the sorting procedure is overwriting the value stored immediately after the sorted sequence in the array.

The complexity of the insertion sort algorithm depends on the initial array. If the array is already sorted by examining each element, then the best case would be $O(n)$ where n is the array's size. However, the worst case would be $O(n^2)$, as each value has to be swapped through the whole dataset, which makes the complexity increase exponentially as the dataset size increases. The average case would be under $O(n^2)$, since most values will be sorted to the beginning of the dataset, which is highly expected in large datasets.

Note that the insertion sort algorithm is less efficient when it comes to huge datasets than advanced algorithms, such as heap sort, quick sort, or merge sort. The main insertion sort procedure has an iterative operation, which takes one element with each repetition and compares it with the other elements to find its correct place in the array. Sorting is typically done in-place, by iterating through the array and increasing the sorted array behind it [1].

The Insertion sort algorithm is the optimal algorithm when it comes to incremental, instantly, and dynamically initiated data, which is due to its adaptive behavior. This paper proposes an enhanced algorithm to reduce the execution time of the insertion sort algorithm by changing the behavior of the algorithm, more specifically on large datasets. This proposed algorithm called Enhanced Insertion Sort algorithm (EIS), which aims to enhance how the elements are relocated from the first part of the dataset, rather than waiting to find its correct position by comparing and swapping. Instead, a simple question is asked during the algorithm's execution; is the particular element less than the determined threshold? If yes, then the algorithm applies by traversing the elements that are under the threshold to find the correct position of this particular element. This algorithm will be explained in detail in Section III.

The structure of the paper is as follows. Section II presents a brief of related works that are proposed to handle and improve the insertion sort algorithm. Section III describes the proposed EIS algorithm with an explanation of its complexity cost, Pseudo-code, implementation code and finally simple comparisons between EIS and other IS algorithms. Section IV shows the experimental results of our proposed EIS algorithm. Finally, the conclusion of the paper presented in Section V.



```
i ← 1
while i < length(A)
    j ← i
    while j > 0 and A[j-1] > A[1]
        swap A[1] and A[j-1]
        j ← j - 1
    end while
    i ← i + 1
end while
```

Fig. 1. Pseudo-Code of a Typical IS Algorithm.

II. Related Work

The IS algorithm is considered as one of the best and most flexible sorting methods despite its quadratic worst-case time complexity, mostly due to its stability, good performances, simplicity, in-place and online nature. It is a simple iterative sorting procedure that incrementally builds the final sorted array [2]. There were several suggestions to improve insertion sort and some of them were even implemented, as seen in [2-4]. Insertion sort can be simplified by using an external element, known as a sentinel value [5]. Bidirectional approaches were proposed in [6], where it consists of two steps, the first step compares both the first and last elements, and then swaps them if the first element is larger. The second step takes two adjacent elements from the beginning of the array and then compares them as well. Abbasi and Dahiya [7] proposed a bidirectional approach to minimize the shifting process, which supposes that there are two sorted parts on the left and right. This approach reduces the shifting process, rather than an element that may shift through the whole array. Patel, et al. [8] presented an approach of inserting elements from the middle of a dataset and applying a bidirectional sorting, using arrays as structured data. Paira, et al. [9] proposed an approach that applies a dual scan from both directions, which locates the position from both sides. Sodhi, et al. [10] presents a binary insertion sort that achieves a time complexity of O($n^{1.585}$) for some average cases by reducing the number of comparisons. This approach starts with the middle element, which reduces the number of swaps needed. Then it determines the position of the suitable location for each element. Afterward, it chooses one direction, either left or right, then adds or appends it to another array. Khairullah [11] presented an approach that keeps track of both directions. It also starts from the middle element's location and compares it according to the element in the middle. Some approaches were not bidirectional, such as [12], which implements an algorithm to simply arrange a worst-case insertion sort that reverses the values.

However, bidirectional methods are efficient when compared with other classical insertion sort algorithms, but these kinds of approaches require the complete dataset to be sorted before knowing its size. In this case, these approaches cannot be implemented in applications that have incoming incremental data.

III. Methodology

The methodology section demonstrates a detailed explanation of the proposed EIS algorithm method in subsection III.A, then it presents an analysis of the algorithm's complexity in subsection III.B. Further, subsections III.C and III.D handling the details of Pseudo-code and the implementation code of the EIS algorithm. Finally, subsection III.E illustrates a simple comparison between the EIS and IS algorithms.

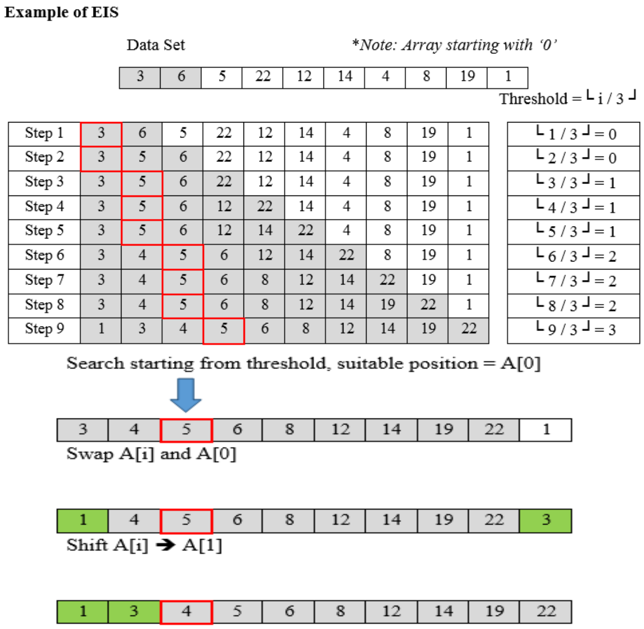
A. EIS Algorithm Method

In the EIS algorithm, the enhancement occurs when the algorithm behaves differently, more specifically when a value of the selected element is lower than a given threshold. The threshold is defined as the index of selected elements from the sorted part of the array, A, in the particular step during the EIS algorithm. Note that the Threshold= $\lfloor i/3 \rfloor$, where *i* is defined as the index of the particular element which is select to be sorted and $0<i\leq n$. Please note that if *i* element is selected now to be sort, then this means that all elements in the array A from A[0] to A[*i*-1], are fully sorted according to the original insertion sort algorithm behavior. Moreover, the threshold is determined by ⅓ of the dataset size, which changes dynamically as the algorithm sorts the elements one by one and *i* will increase by 1 in each iteration. The ratio, ⅓, of the dataset was chosen after executing several experiments, which concluded that it is the best case in terms of time complexity. It is clear that there is no specific way to determine the optimal ratio for the threshold unless you try out a bunch of different values and test the performance. This will be further discussed in section VI.

The functionality of the proposed algorithm is the same as the insertion sort process, but it asks a question before it starts comparing and swapping the selected *i* element in A[*i*] where the index=*i*; is the value of the *i* element being compared less than the value of the element in threshold index? If yes, then.

The EIS algorithm searches for the correct index to move the *i* element and place it. Note that EIS will start searching for the correct index for element *i* from the segment part (block) that contains elements that have values that are less than the value of the threshold index. When the suitable index is spotted, it swaps the selected element to the specific index and then shifts all other elements to the right. This operation reduces the number of comparisons and swapping of elements and this reduction will increase if the size of the array is also increased. In case that the value of the element is higher than the value of the threshold, then EIS behaves like the original insertion sort process and the original IS procedure will be done for this particular selected element in index *i*.

Fig 2. demonstrates an illustration example of the procedure of the proposed Enhanced Insertion Sort (EIS). The threshold is dynamically changing based on the size of the traversed elements while the algorithm sorts the elements on by one. In each step, the algorithm examines whether the value of the selected element *i* is lower than the value of the threshold index or not. In steps 6 and 9, the algorithm begins to search beyond the threshold, and then it replaces the elements to a suitable index, and then shifted all the elements to the correct index. To illustrate that, if you look at step 9, rather than comparing the element in A[*i*]=1, where *i*=9, with all the other elements from A[0] to A[*i*-1], it only performs three comparisons with elements A[2] = 5, A[1] = 4 and A[0] = 3 wherein this particular step the threshold = 3. Note that the number of the comparisons in step 9 will be 9 if we use the original LS, but when using our proposed EIS, the number of the comparisons will reduce to 3, as shown in Table I.



Step 1	3	6	5	22	12	14	4	8	19	1	⌊ 1 / 3 ⌋ = 0
Step 2	3	5	6	22	12	14	4	8	19	1	⌊ 2 / 3 ⌋ = 0
Step 3	3	5	6	22	12	14	4	8	19	1	⌊ 3 / 3 ⌋ = 1
Step 4	3	5	6	12	22	14	4	8	19	1	⌊ 4 / 3 ⌋ = 1
Step 5	3	5	6	12	14	22	4	8	19	1	⌊ 5 / 3 ⌋ = 1
Step 6	3	4	5	6	12	14	22	8	19	1	⌊ 6 / 3 ⌋ = 2
Step 7	3	4	5	6	8	12	14	22	19	1	⌊ 7 / 3 ⌋ = 2
Step 8	3	4	5	6	8	12	14	19	22	1	⌊ 8 / 3 ⌋ = 2
Step 9	1	3	4	5	6	8	12	14	19	22	⌊ 9 / 3 ⌋ = 3

Fig. 2. The Procedure of the Enhanced Insertion Sort (EIS).

B. The Complexity of the EIS Algorithm

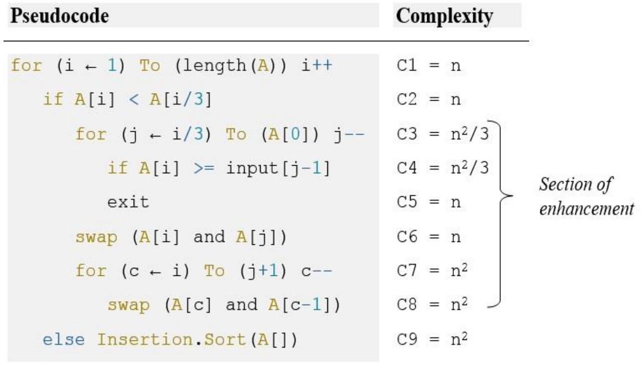
As shown in Fig. 3, the cost of the execution is reduced to $n^{2/3}$, as the algorithm relies heavily on the threshold procedure to reduce the search space. Given the following ➔ $2(n^{2/3}) + 3n^2 + 4n$. It will remain as $O(n^2)$. However, the results empirically show more efficient behavior. The Pseudo-code of the proposed Enhanced insertion sort algorithm is shown in Fig. 3.

C. Source Code of EIS Algorithm

Fig. 4 shows the java source code of the proposed Enhanced insertion sort algorithm. Furthermore, the full implementation of the EIS algorithm can be downloaded and run from the Github website:https://github.com/muhyidean/EnhancedInsertionSort-ThresholdSwapping.

D. Comparisons between the EIS and other IS Algorithms

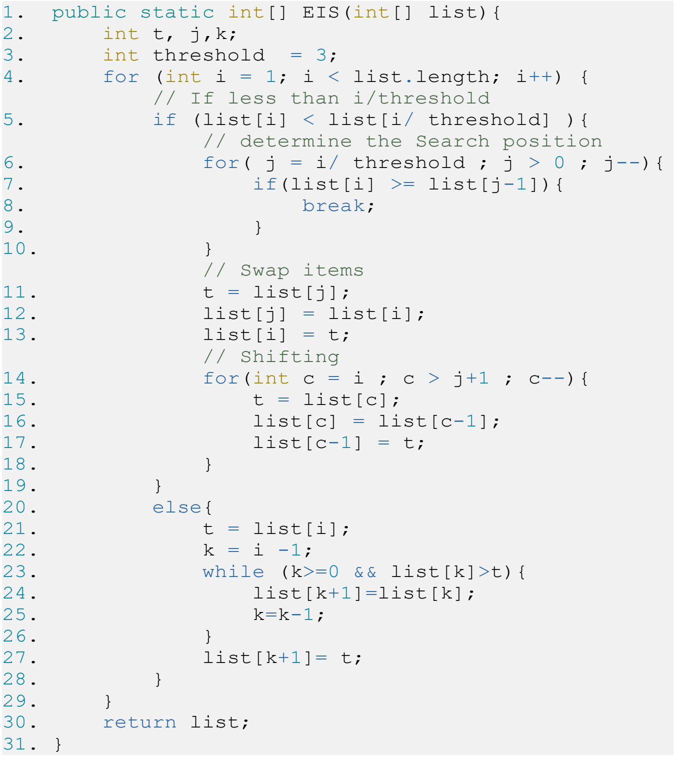
Table I shows a detailed comparison between the insertion sort (IS) and our proposed algorithm (EIS) using the same dataset in Fig. 2 as an example. As shown in Table I, it is clear that the number of comparisons for the dataset example using the proposed EIS is better than using IS, where the number of comparisons for the IS algorithm is 45; while the number of comparisons for EIS is 34. As a particular example for specific iteration in step 9, when *i* = 9, it is clear that the number of the comparisons will be 9 if we use the original IS, but when using our proposed EIS, the number of the comparisons will reduce to 3, which is the one third (⅓) threshold value as shown in Table I.



```
Pseudocode                                         Complexity

for (i ← 1) To (length(A)) i++                     C1 = n
   if A[i] < A[i/3]                                C2 = n
      for (j ← i/3) To (A[0]) j--                  C3 = n²/3
         if A[i] >= input[j-1]                     C4 = n²/3
         exit                                      C5 = n
      swap (A[i] and A[j])                         C6 = n
      for (c ← i) To (j+1) c--                     C7 = n²
         swap (A[c] and A[c-1])                    C8 = n²
   else Insertion.Sort(A[])                        C9 = n²
```

Fig. 3. Pseudo-Code of the EIS Algorithm.



```java
public static int[] EIS(int[] list){
    int t, j,k;
    int threshold  = 3;
    for (int i = 1; i < list.length; i++) {
        // If less than i/threshold
        if (list[i] < list[i/ threshold] ){
            // determine the Search position
            for( j = i/ threshold ; j > 0 ; j--){
                if(list[i] >= list[j-1]){
                    break;
                }
            }
            // Swap items
            t = list[j];
            list[j] = list[i];
            list[i] = t;
            // Shifting
            for(int c = i ; c > j+1 ; c--){
                t = list[c];
                list[c] = list[c-1];
                list[c-1] = t;
            }
        }
        else{
            t = list[i];
            k = i -1;
            while (k>=0 && list[k]>t){
                list[k+1]=list[k];
                k=k-1;
            }
            list[k+1]= t;
        }
    }
    return list;
}
```

Fig. 4. Java Code of the EIS Algorithm.

TABLE I. COMPARISONS BETWEEN THE EIS AND IS ALGORITHMS

Dataset ➔	3	6	5	22	12	14	4	8	19	1

	Number of Comparisons based on the IS algorithm	**Number of Comparisons based on the EIS algorithm**	**LIST AFTER SORTING**	**ETI Element To Insert**	**Number of Sorted items**	**VALUE OF (i)**
	1	1	3,6	6	0	**i = 1**
	2	2	3,5,6	5	2	**i =2**
	3	3	3,5,6,22	22	3	**i =3**
	4	4	3,5,6,12,22	12	4	**i =4**
	5	5	3,5,6,12,14,22	14	5	**i =5**
	1	6	3,4,5,6,12,14,22	4	6	**i =6**
	7	7	3,4,5,6,8,12,14,22	8	7	**i =7**
	8	8	3,4,5,6,8,12,14,19,22	19	8	**i =8**
	3	9	1,3,4,5,6,8,12,14,19,22	1	9	**i =9**
Total	**34**	**45**	** The number of comparisons was reduced using EIS*			

IV. Results

A. The Implementation Procedure

The proposed EIS algorithm was implemented using the Java programming language. In order to evaluate the performance of the proposed EIS algorithm. Three algorithms, the classical IS algorithm, the proposed EIS algorithm using ⅓ threshold, and ¼ threshold are utilized and the results are compared. The performance evaluations were performed on a computer machine with a 2.8 GHz Intel Core i5 processor with 4 GB 1600 MHz DDR3 memory on a windows platform. An experimental test has been done on empirical data (integer numbers) that are generated randomly using Java.

To verify that the same data is examined for each execution, a random dataset is generated and copied to three different arrays. Then the data is used to apply each algorithm and the time it takes (in milliseconds) to complete the sorting process was recorded. This is to assure that the algorithm's performance works for all types of data organizations and sizes. Twenty random datasets were utilized and the average execution times are reported. The sizes of the datasets that were utilized were 10000, 50000, 100000, and 500000.

B. Experimental Results and Discussion

Table II shows the overall performance results of the employed algorithms on different utilized datasets. As shown in Table II, the performance results, exposed by the classical IS algorithm, are stable on all utilized datasets. These results are expected since the computational complexity of the IS algorithm is practically the same. Further, the reported results in Table II emphasize that the proposed EIS algorithm using thresholds of ¼ and ⅓ reported enhanced performance results as compared to the classical IS algorithm. In fact, when the threshold equals ⅓ the results are superior. Consequently, a threshold of ⅓ states an appropriate threshold for employing the EIS algorithm.

The results are also showing, as the size of the dataset increases, the deviations of the performance results are also improved. Fig. 5 to 7 illustrate exceptional performance results of the proposed EIS especially when the threshold equals ⅓ and when the dataset's size is larger.

As Fig. 5 Shows, the performance results of the EIS algorithm are much improved in terms of computational complexity time. Although, the complexity of the EIS algorithm, as reported in section III.B, states that the EIS has an $O(n^2)$ computational time, but the results empirically demonstrate better performance.

Fig. 6 and 7 demonstrate also similar performance results. When the size of a dataset is larger, the EIS algorithm performs better. Overall, the EIS algorithm, with varying datasets sizes and with a range of threshold values, performs empirically better than the classical IS algorithm.

To further compare the performance of the employed algorithms, the average, maximum and minimum execution times for each dataset's size are reported and compared. As Table III shows the classical IS reported results are the lowest in terms of average execution time. When the threshold equals ¼, the EIS algorithm has a higher maximum value and a lower minimum value than using ⅓. This indicates that when the threshold equals ⅓ the reported results are more efficient. The lowest average execution times are highlighted to determine the most efficient performance. To conclude, The EIS algorithm, with a threshold of ⅓, demonstrates the best average performance results as highlighted in Table III.

The improvements of the proposed EIS over the IS algorithms are also compared in terms of the average execution time and reported. Table IV shows that the proposed EIS algorithm outperforms the classical IS algorithm. The performance improvements of the proposed EIS algorithm in terms of execution time on average was 23%. The main reason for the significant improvement in performance is that the threshold procedure of the EIS algorithm reduces the number of comparisons and swapping needed to complete the sorting procedure.

TABLE II. THE PERFORMANCE OF THE EMPLOYED ALGORITHMS

dataset size / algorithms	Execution time (in milliseconds) on 20 Datasets that are randomly generated																			
10^4	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Ins	36	45	38	34	71	48	60	39	57	69	67	36	41	44	74	42	39	112	42	38
EIS ⅓	34	32	54	43	49	34	41	42	33	33	48	35	35	36	44	42	37	52	31	34
EIS ¼	35	48	43	30	53	36	31	34	38	42	49	37	33	32	35	42	53	55	33	35
(10^4)*5	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Ins	769	827	770	774	792	813	766	865	809	833	780	774	766	794	763	809	811	773	784	782
EIS ⅓	773	625	565	601	597	609	556	567	632	559	574	582	606	597	763	571	565	578	568	603
EIS ¼	516	748	742	749	532	534	512	756	576	600	742	529	541	742	783	775	506	740	781	725
10^5	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Ins	3035	3074	3032	3051	3031	3034	3021	3041	3062	3050	3026	3037	3038	3063	3033	3055	3040	3023	3041	3047
EIS ⅓	2139	2134	2131	2150	3014	2146	2125	2161	2141	2134	2959	2151	2966	2149	2127	2158	2136	2133	2150	2963
EIS ¼	2972	2033	2973	2033	2019	2964	2010	3000	2023	2094	2967	2997	2204	3037	2969	2037	2969	2015	2962	2963
(10^5)*5	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Ins	75242	77496	75306	75761	76568	76665	76140	76173	75406	75483	75114	75408	75092	75518	75378	75821	76094	75198	75792	76444
EIS ⅓	53348	54684	53315	53697	54092	75198	55537	59393	55480	53693	53173	53633	73721	55514	53463	58709	53628	73674	73531	59116
EIS ¼	73333	52147	50568	50678	55212	73765	73352	51036	73816	73268	73196	73345	73087	50763	54505	73449	73583	73273	50997	50684

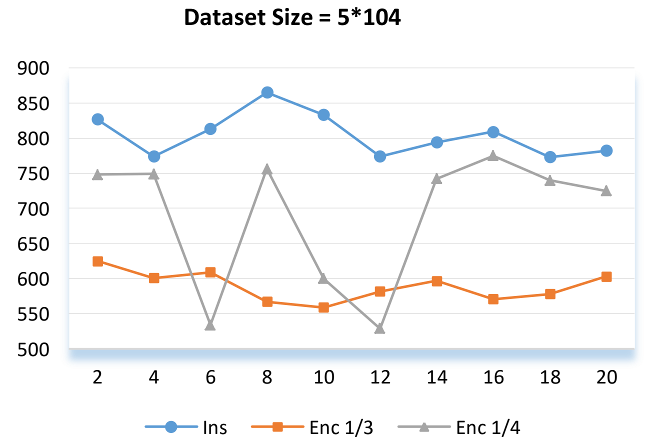


Fig. 5. The Performance of the Employed Algorithms on utilized Datasets of Size 5*104.

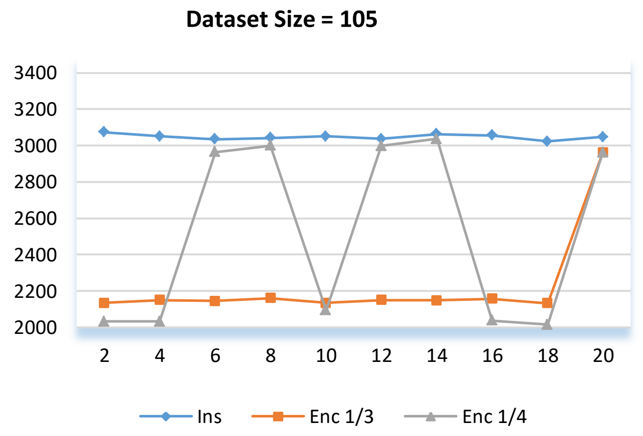


Fig. 6. The Performance of the Employed Algorithms on utilized Datasets of Size 105.

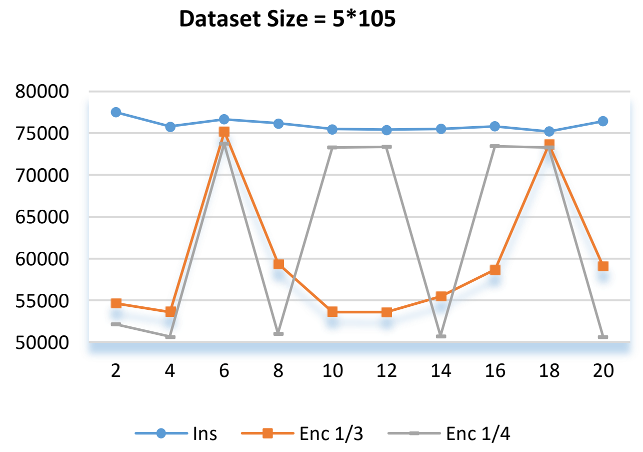


Fig. 7. The Performance of the Employed Algorithms on utilized Datasets of Size 5*105.

TABLE III. AVERAGE, MAX AND MIN EXECUTION TIME OF THE IS AND EIS ALGORITHMS

dataset size algorithms	Overall Performance		
10^4	**Average**	**Max**	**Min**
IS	51.6	112	34
EIS ⅓	39.45	54	31
EIS ¼	39.7	55	30
$5*10^4$	**Average**	**Max**	**Min**
IS	792.7	865	763
EIS ⅓	604.55	763	556
EIS ¼	656.45	783	506
10^5	**Average**	**Max**	**Min**
IS	3041.7	3074	3021
EIS ⅓	2308.35	3014	2125
EIS ¼	2562.05	3037	2010
$5*10^5$	**Average**	**Max**	**Min**
IS	75804.95	77496	75092
EIS ⅓	58829.95	75198	53173
EIS ¼	63702.85	73816	50568

TABLE IV. THE OVERALL PERFORMANCE RESULTS OF THE EIS ALGORITHMS

Dataset size Algorithms	10^4	$5*10^4$	10^5	$5*10^5$
IS	51.6	792.7	3041.7	75804.95
EIS ⅓	39.45	604.55	2308.35	58829.95
EIS ¼	39.7	656.45	2562.05	63702.85
Improvement of EIS ⅓ over the IS algorithm				
Improvement	**24%**	24%	24%	22%

C. *Source Code of Implementation*

Due to the space limitation, the source code of the proposed EIS algorithm is uploaded online to the GitHub website (https://github.com/muhyidean/EnhancedInsertionSort-ThresholdSwapping).

V. CONCLUSION

Insertion sort is the suitable sorting algorithm when it comes to incremental, instantly and dynamically initiated data. Yet, its complexity increases exponentially when the data size increases, making it inefficient. In this paper, an enhancement of the IS algorithm was proposed, named enchanted insertion sort, to improve the computational complexity of the IS algorithm by changing its behavior. The proposed algorithm reduces the number of comparisons and swapping needed to complete the sorting procedure. After executing the algorithm and comparing it with the classical insertion sort algorithm, there was an improvement of 23% in terms of the execution time taken to complete the sorting process. The worst case of the complexity remains $O(n^2)$, but the reported results are empirically promising. The efficiency of the proposed EIS algorithm is attributable to the reduction of the number of comparisons during the sorting process.

REFERENCES

[1] M. Aliyu and P. Zirra, "A Comparative Analysis of Sorting Algorithms on Integer and Character Arrays," The International Jornal of Engineering and Science, pp. 25-30, 2013.

[2] A. Chadha, R. Misal, T. Mokashi, and A. Chadha, "Arc sort: Enhanced and time-efficient sorting algorithm," International Journal of Applied Information Systems vol. 7, pp. 31-36, 2014.

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introduction to algorithms: MIT press, 2009.

[4] S. Samee, A. Chitte, and Y. Tangde, "Analysis of Insertion Sort in Java," Advances in Computational Research, vol. 7, p. 182, 2015.

[5] O. Nevalainen, T. Raita, and H. Thimbleby, "An improved insert sort algorithm," Software: Practice and Experience, vol. 33, pp. 999-1001, 2003.

[6] M. P. K. Chhatwani, "Insertion sort with its enhancement," International Journal of Computer Science and Mobile Computing, vol. 3, pp. 801-806, 2014.

[7] H. Abbasi and M. Dahiya, "Services Marketing: Challenges and Strategies," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 4, pp. 345-349, 2016.

[8] S. Patel, M. D. Singh, and C. Sharma, "Increasing Time Efficiency of Insertion Sort for the Worst Case Scenario," International Journal of Computer Applications, vol. 975, p. 8887.

[9] S. Paira, A. Agarwal, S. S. Alam, and S. Chandra, "Doubly Inserted Sort: A Partially Insertion Based Dual Scanned Sorting Algorithm," in Emerging Research in Computing, Information, Communication and Applications, ed: Springer, 2015, pp. 11-19.

[10] T. S. Sodhi, S. Kaur, and S. Kaur, "Enhanced insertion sort algorithm," International journal of Computer Applications, vol. 64, 2013.

[11] M. Khairullah, "Enhancing Worst Sorting Algorithms," International Journal of Advanced Science and Technology, vol. 56, pp. 13-26.

[12] P. S. Dutta, "An approach to improve the performance of insertion sort algorithm," International Journal of Computer Science & Engineering Technology (IJCSET), vol. 4, pp. 503-505, 2013.

A Comparison of Data Sampling Techniques for Credit Card Fraud Detection

Abdulla Muaz[1], Manoj Jayabalan[2*], Vinesh Thiruchelvam[3]
School of Computing, Asia Pacific University of Technology and Innovation, Kuala Lumpur, Malaysia[1, 3]
Faculty of Engineering & Technology, Liverpool John Moores University, Liverpool, UK[2]

***Abstract*—Credit Card fraud is a tough reality that continues to constrain the financial sector and its detrimental effects are felt across the entire financial market. Criminals are continuously on the lookout for ingenious methods for such fraudulent activities and are a real threat to security. Therefore, there is a need for early detection of fraudulent activity to preserve customer trust and safeguard their business. A major challenge faced in designing fraud detection systems is dealing with the class imbalance issue in the data since genuine transactions outnumber the fraudulent transactions typically account less than 1% of the total transactions. This is an important area of study as the positive case (fraudulent case) is hard to distinguish and becomes even harder with the inflow of data where the representation of such cases even decreases further. This study trained four predictive models, Artificial Neural Network (ANN), Gradient Boosting Machine (GBM) and Random Forest (RF) on different sampling methods. Random Under Sampling (RUS), Synthetic Minority Over-sampling Technique (SMOTE), Density-Based Synthetic Minority Over-Sampling Technique (DBSMOTE) and SMOTE combined with Edited Nearest Neighbour (SMOTEENN) was used for all models. The findings of this study indicate promising results with SMOTE based sampling techniques. The best recall score obtained was with SMOTE sampling strategy by DRF classifier at 0.81. The precision score for this classifier was observed to be 0.86. Stacked Ensemble was trained for all the sampled datasets and found to have the best average performance at 0.78. The Stacked Ensemble model has shown promise in the detection of fraudulent transactions across most of the sampling strategies.**

Keywords*—Data imbalance; credit card fraud; sampling techniques

I. Introduction

Transactions using credit cards have become an important aspect of our daily lives. Purchase of goods and services are no longer a chore that requires physical activity, rather it is initiated with a touch of a button on our smartphone or personal computers. The authorization of transactions is rigorous and secure although such conveniences are brought about by compromising the proof of identity checks which require personal identification documents, authorized signature and physical presence. The basis of the identity proof in such transactions is the information on the card along with digital identification tied to the cardholder.

The conveniences brought about by digital transactions makes it a target for fraudsters who employ elegant tactics for theft and illicit use. Credit card fraud is generally an unauthorized movement by an individual who is not authorized to perform the said account operation. It can be also classified where a person transacts with a card, without the explicit permission of the owner of the cardholder or card issuer [1]. The most common form of credit card frauds are stolen or lost cards, fraudulent applications, counterfeit card fraud, non-receipt fraud, card not present (CNP) and account takeovers [2]. According to the European Central Bank [3], on the composition of fraudulent credit card transactions for the year 2016, 73% of the fraudulent transactions were a result of CNP, where payments are made via the internet or telephone.

Billions worth of transactions is lost worldwide every year due to fraudulent credit card transactions. According to the Nelson Report on global payment systems, the amount of losses due to credit card fraud is $22.8 billion, and this indicated a 4.4 percent increase from the year 2015. It was also highlighted that 38.6% of this global credit card frauds are accounted for from frauds in the United States. The Nelson Report also projects that the credit card fraud losses are expected to grow by over $10 billion over the next three years [4].

With the increasing amounts of loss due to such illicit activities costing institutions and individual’s huge amounts of money, tackling this issue has become a priority over the past decade and various studies have been conducted to address this problem. Financial institutions are constantly on the verge of upgrading their fraud detection systems. Association of Certified Fraud Examiners (ACFE), suggests that proactive data analysis and continuous monitoring of real-time activity as the key for minimizing and preventing fraudulent credit card transactions [2].

Financial institutions and credit card issuers collect and store a vast amount of transaction data. Every credit card transaction composes key attributes such as the card identifier, transaction date, recipient and amount of transaction, which are stored in the databases. Fraud Detection Systems (FDS) implements various layers of validation to flag potential frauds using such datasets.

Although, machine learning and predictive analytics might not answer the question of exactly the type of fraud that may occur, it has the potential to flag suspicious activities and identify potential frauds with the help of a trained model, on historical data combined with expert analysis. Such systems can equip institutions with proactive insights into the future, to enable them to better cope and mitigate fraudulent transactions.

*Corresponding Author

Real-world implementation of FDS cannot reliably check all transactions as it is constrained by the human labour required to validate the sheer number of alerts raised the by system. It mainly relies on fraud investigators who are used as a confirmatory layer whereby flagged fraudulent activities (alerts) are verified and validated by the designated investigators.

Transactions which are then reported by the customer during this window are flagged or labelled as fraudulent and the unreported transactions labelled genuine transactions. To summarize, there are two ways FDS samples the data; immediate feedback samples (transactions with investigator feedback) and delayed samples (transactions whose fate are known only after a set reaction-time period). This is a crucial distinction to be considered when implementing an accurate FDS as every transaction is not immediately labelled either fraud or genuine [5].

Fraudulent labels in the dataset can, therefore, be safely assumed to be verified and validated by the investigators. However, there are other challenges in designing accurate machine learning techniques for such data. Firstly, the non-stationary data distribution (fraudulent and genuine transactions share similar profile). Often at times fraudsters mimic the cardholders spending behaviours, which makes the profiles of the fraudster and cardholder very similar in such cases and different in the other cases. This changing dynamics between genuine and fraudster profiles also known as concept drift, makes it particularly challenging for machine learning algorithms to accurately predict fraudulent transactions [5]–[9].

Secondly, the skewness or the class imbalance in these datasets poses a considerable challenge in building accurate machine learning models. This is the case for a variety of real-world applications where the true class or the interested observations tend to be a fraction of the total cases. Credit card fraud detection has this distinctive characteristic as majority of the transactions are genuine while the concerned cases (fraud activity) has very few transactions. This is known as the class imbalance and it is significant because the positive class is often the rare class and predicting this class becomes harder as the number of false class keeps on increasing. Machine learning models typically work on the assumption of an equal class balance and equal cost of misclassification, therefore adequate measures have to be taken in order to address this issue of class imbalance [5], [6], [10]–[12].

Detection of credit card fraud is classified as a cost-sensitive problem, where there is an associated cost incurred for incorrectly classifying a genuine transaction as fraudulent and incorrectly classifying fraudulent transaction as genuine. In the absence or no occurrence of fraud, there is no associated administrative costs incurred by the financial institution. However, failure to detect the fraud is a loss of the particular transaction amount. It is thus, an important proposition to incorporate in to the FDS, particularly in the development of models on class imbalanced datasets [9].

A. Contributions

The research contributes both theoretically and practically. The significance in terms of both means are summarized as follows.

This paper provides an overview of the most recent literature on credit card fraud detection strategies which focused on the newest Machine Learning techniques while addressing the major challenges faced by the traditional FDS. The research offers an up to date perspective on trends in the credit card fraud detection domain, model evaluation metrics that offer the best results and outlines limitations of existing FDS. Researchers can find this paper helpful as it is a good starting point, to kickstart a research on implementing machine learning techniques for credit card fraud detection.

The practical contributions of this research are to provide a sound and realistic model that articulates the classification problem pertaining to the domain of credit card fraud detection. Sampling strategies proposed to be implemented in this paper shall enable researchers to promptly use and adopt this technique which best serves their research goal. Various sampling techniques shall be implemented to generate and train different machine learning models, and conclusively summarize experimental results of the built models using a multitude of relevant model evaluation metrics.

The paper addresses key challenges faced in building machine learning models for FDS, and experimentally prove strategies to mitigate or minimize such challenges. Therefore, it is an invaluable contribution to the financial sector, with the contribution of a predictive model able to accurately predict fraudulent credit card transactions.

II. Related Works

This section synthesizes the contents and ideas in the existing studies and encompasses key subject matters regarding the domain of credit card fraud detection. These subjects include, machine learning techniques, sampling techniques, visual data analytics, feature engineering and model evaluation metrics.

A. Machine Learning Techniques

Credit card fraud detection studies on the use of predictive analytics have shown that researchers adopted various methods such as Artificial Neural Networks, k-Nearest Neighbour (kNN), Logistic Regression (LR), AdaBoost, Naïve Bayes (NB) and many more [6], [13]–[17].

In [6], used NB, kNN and LR on the European card holders dataset. This dataset contains anonymized transaction data of European credit card holders which were collected for a period of two days and contains 284,807 samples. The results of this study conclude that kNN produced the best results for accuracy, sensitivity and specificity. Although the authors argue that this potentially could be caused by the generation of synthetic samples using Synthetic Minority Over Sampling which uses KNN.

One study proposed an improved method of sampling to produce a better performance, which referred to as Moving to Adaptive Samples in Imbalanced (MASI) dataset. The study implemented Random Forest (RF), Support Vector Machines (SVM) and C 5.0 Decision Tree algorithm to conclude that SVM produced the best results [18].

In another study using the same dataset implemented LR, KNN, Linear SVM, RBFSVM decision trees, RF, and NB algorithms [17]. Although, both [18] and [6] implemented the same models, the sample size was 350, which was a result of random under sampling. The highest sensitivity score achieved for the study was SVM with a score of 94%.

Random Forest is implemented by majority of the researchers [6], [18], [19] with varying degree of results. In [20] experimented on a weight assignment approach to the RF, using out-of-bag error to compute the weights while other researchers typically opted for using various sampling techniques.

Deep Learning techniques such as ANN, Recurrent Neural Networks (RNN), Long Short-term Memory (LSTM) and Gated Recurrent Units (GRU) was implemented [21]. The LSTM and GRU outperformed the traditional ANNs, however the shortcomings are that the training was not conducted to achieve optimal model stability. It was cited that "performance improved whenever network size was increased", and future recommendation was made to identify an optimal stopping point.

Restricted Boltzmann Machines (RBM) was another topology of Deep Learning which was implemented by past researchers [22]. The researcher used a novel approach with the use of unsupervised machine learning techniques (Stacked Auto Encoder) to identify optimum weights, which was then applied to a supervised machine learning model RBM achieved an accuracy of 91.5%.

B. Issues of Class Imbalance

Numerous studies have shown different approaches to deal with this issue in the context of implementing accurate prediction model which are aimed at improving the detection rate of fraudulent transactions [6], [18], [23], [24].

The most common method implemented in the existing studies to handle the problem at data level, where the data is subjected to various sampling techniques. Random under sampling (RUS) is implemented where the majority class instances are removed [10], [17] or random oversampling (ROS) is used where minority class instances are added by replicating training samples with the same class representation. Some advanced methods were also used to oversample with techniques such as Synthetic Minority Oversampling Technique (SMOTE) which creates new synthetic instances of the minority class using kNN. Synthetic instances which are created using this technique have been shown to perform better, than simply using random oversampling or replication of instances [1], [25], [26]. Alternative methods to the SMOTE, was implemented by [23] and [18]. The drawbacks of the SMOTE sampling technique such as loss of potential information and potential for model overfitting for the synthetic samples.

In [18] proposed an improved method of sampling using an approach which the author refers to as Moving to Adaptive Samples (MASI) in Imbalanced dataset and obtained comparatively better performance against other sampling techniques such RUS, ROS and SMOTE. While SMOTE, resampling generates new instances and increase the data size prior to the implementation of the classifier, MASI adaptively creates synthetic samples which are created based on the density distribution of original data and up-samples the minority class by changing class labels. The researchers indicate this reduces the bias of the classifier as it moves the samples in minor class closer to the decision boundary.

Alternative to tackling the imbalance issue on the dataset, ensemble learning handles class imbalance issue at the algorithmic level. Ensemble methods typically include bagging and boosting that primarily aims to lower the variance in the data by using multiple classifiers. In bagging method, multiple weak classifiers are trained on different subsets of the majority class and minority class before combined final classifier is built using all the weak classifiers. AdaBoost employs similar strategy and can be implemented for many classification problems and it eliminates the need for exploring an optimum class balance ratio while alleviating the information loss which can be caused by RUS, and overfitting issue caused by ROS and SMOTE methods [25], [27].

One study implemented a new oversampling strategy which combined k-means clustering with genetic algorithm to oversample the minority class. The researchers propose this solution as opposed to SMOTE and other sampling strategies highlighting the potential for information loss and overfitting [23].

C. Feature Engineering

Fraudsters constantly change their behaviours and implement new ways to commit frauds, which renders traditional expert rules. Machine learning methods are also prone to this type of problems, however adoption of new strategies can assist to counter. Feature engineering is a method which can be used extensively to counter this effect, whereby new features are created based on the card holder's behaviour over time. These new features aids the machine learning models to distinguish patterns from the normal card holder behaviour [9], [28].

Feature engineering is proven to be an important aspect of predictive analytics for detection of credit card frauds. Financial institutions obtain and store large amounts of data related to transactions such as transaction amount, account holder details, time of transaction and more. While these collected data serve as good predictors in a classifier setting, it has the potential to be enriched with new information such as card holder spending habits in a set time frame, average amount spent in different geographical areas or product and service types. For example, a card holder can be profiled by his spending habit at home, but this may differ completely with his spending habit on a vacation in India. Such features could potentially be able to discover patterns and solve the concept-drift problem where card holder and fraudster behaviour is distinguished with the help of new data dimensions [9], [23].

It is also noteworthy, that single transaction information is typically insufficient for the purpose, rather aggregate measures which combines to form new features are ideal [9].

D. Evaluation Metrics

Evaluation metrics are an important aspect to understand the performance of the machine learning models. Detection of credit card fraud is classified as a cost-sensitive problem, where there is an associated cost incurred for incorrectly classifying a genuine transaction as fraudulent and incorrectly classifying fraudulent transaction as genuine [9]. As such, the choice of evaluation metric must be carefully chosen and shall be relevant in terms of the objective of the study and available data.

The existing studies have, adopted various evaluation metrics for binary classifiers such as Area Under the Curve (AUC), Sensitivity (also referred to as Recall), Precision and F1 score [10], [18], [21], [22] [14], [16], [21].

Machine learning models work on the assumption of equal class distribution and equal cost of misclassification. Using accuracy metric for evaluating a model is not suitable for datasets with class imbalance issue as it would bias the model towards majority class since the accuracy metric calculates the total of correct predictions [20], [10].

Area Under the Curve (AUC), is a measure of the probability that the model or classifier will choose a random positive instance higher than a random negative instance. AUC is a metric; many researchers have adopted [22], [26], [29] and gives a good indication of the overall predictive performance of the model across various probability threshold settings and is very well suited for the class imbalanced modelling.

Precision is the percentage of true positives among all positive predictions, while recall indicates the total correctly predicted positive classes over the total predictions for both correctly predicted positive class and falsely predicted positive class. F1 measure is the mean of sensitivity and precision. Out of all these metrics used in this study, the most useful metric which was able to give a clear indication of the best classifier was sensitivity or recall metric.

III. Methods and Techniques

This section briefs the research methodology that will be adopted to achieve the objectives of this research. The section includes an overview and key processes involved in the methodology, such as dataset summary, sampling techniques, and machine learning algorithms.

The dataset collected for this study is secondary data consist of transaction data of European credit card holders which were collected for a period of two days and contains 284,807 observations with 31 variables out of which 28 variables are anonymized using principal component [30].

The three non-anonymised variables are transaction time, amount and the class label (fraudulent or not fraudulent transaction). The class label indicates ‘0’ for non-fraudulent transaction and ‘1’ for fraudulent transaction. The dataset is highly imbalanced as the percentage of fraud instances accounts to 0.172%. The dataset does not contain any missing values and outliers, therefore pre-processing techniques on the dataset shall not be required. Table I describe the features in dataset. Features V1 to V28 are aggregated to single description for ease of reading.

TABLE I. Dataset Description

Features	**Description**
Time	Number of seconds elapsed between this transaction and the first transaction in the dataset
V1,V2,V3,V4,V5,V6,V7,V8,V9,V10,V11,V12, V13,V14,V15,V16,V17, V18,V19,V20,V21, V22,V23,V24,V25,V26, V27, V28	Result of a PCA Dimensionality reduction to protect user identities and sensitive features
Amount	Transaction amount
Class	1 for fraudulent transactions, 0 otherwise

A. Sampling Techniques

A reliable FDS with detecting all frauds is vital as well as reducing false flags where genuine transactions are misclassified as fraudulent. The associated costs are much higher, when a fraudulent transaction pass through the system undetected (False Negative). However, it is also an important issue when false flags are raised for non-fraud transactions (False Positive), which hurts the customer sentiment as well as an added cost of allocating investigative resources needlessly. Maximizing recall score, is thus significantly important as high recall scores indicate a higher ability for the classifier to detect True Positives (Frauds). Precision scores is also important as the FDS shall avoid or minimize misclassifying genuine transactions as frauds. Therefore, various sampling strategies were adopted, and four different classifiers implemented to conclusively deduce the best and most effective sampling strategies and classifiers best suited for the dataset.

1) Random Under Sampling (RUS): Random Under Sampling is one of the most commonly used sampling techniques, where the majority class is down sampled or reduced to the same number of minority class by randomly removing instances of the majority class. The major problem with RUS is that it is randomly removing data which leads to potential loss of important information which may have been captured.

2) Synthetic Minority Oversampling Technique (SMOTE): SMOTE create synthetic instances of the minority class. These data points are created by assessing the nearest neighbours for each of the minority sample and creating new synthetic instances in the feature space until the minority class is balanced to the given ratio.

3) Density-Based Synthetic Minority Oversampling Technique (DBSMOTE): DBSMOTE algorithm relies on a clustering algorithm called Density-Based Spatial Clustering of Applications with Noise (DBSCAN), which is widely used clustering algorithm used for data mining and machine learning applications. DBSCAN works by grouping together a set of data points based on how close together the points are packed in terms of a distance measurement such as the

Euclidean distance and a given number of minimum points to operate on the bi-dimensional space. DBSMOTE essentially implements the DBSCAN clustering algorithm to form a cluster of the minority class, which is then used to up-sample the minority class. The minimum samples specify the number of data points required to form the dense region.

4) Synthetic Minority Oversampling Technique with Edited Nearest Neighbor (SMOTEENN): SMOTEEN is another variant of SMOTE which is basically a combination of SMOTE and Edited Nearest Neighbour (ENN). The ENN is an effective method which is used to remove noise from the dataset. For any given data point of either class, ENN removes the data point which differs by at least half of the given k-Nearest Neighbour.

B. Machine Learning Techniques

This section, details and justifies the different benchmark machine learning models which are proposed in FDS, highlights the strengths of the machine learning models used, and lists out the evaluation metrics to be used focusing on the class imbalance nature of the dataset.

The machine learning algorithms are Gradient Boosting [25], [27], Stacked Ensemble [31], Artificial Neural Network – Multilayer Perceptron [21], [32] and Random Forest [13], [20].

The model shall primarily be evaluated with Recall score as it is more important to the FDS to accurately detecting the fraudulent transactions (increasing TPR). The precision although not as significant as the recall score, still has associated costs for an FDS and thus the second metric to consider shall be the precision score.

1) Stacked ensemble: Stacked Ensemble model have shown promising improvements in terms of classification accuracy when combined with diverse set of classifiers. In a study by [31], Stacked Ensemble was used for an imbalanced dataset and proved to have gained maximum performance among the other models. Modern applications of machine learning quite often must deal with imbalanced classification as is the case with this study. The current ensemble techniques offer a modification to the traditional ensemble models to allow for maximum performance on imbalanced learning. The Stacked Ensemble model allows for customization of parameters that are designed specifically to handle class imbalance issues [33]. The SE is a combined model of chosen base models of and uses General Linear Model (GLM) as a default meta learner to enhance the model performance.

2) Gradient boosting machine: Gradient Boosting Machine can be used for either regression or classification models. It is an ensemble learning method which operates on the concept of Boosting where weak learners are built gradually to allow for maximum prediction accuracy with each iteration. Unlike Random Forests which use Bagging, and trees are built independent of one another, Boosting aims to build trees which are built based on the results of previously built trees. Boosting although improves accuracy it is slower and has reduced interpretability than other traditional models.

This study shall use gradient boosting model to allow for a diverse set of classifiers where four different categories of learning is considered, namely, Bagging, Boosting, Deep Learning and Super Learning and gradient Boosting Machine [33].

3) Random forest: Random Forest is essentially and ensemble model consisting of many decision trees all of which are made from the same input dataset. The high prediction accuracy of random forests is due to the fact that a combined output is obtained in random forest by comparing outputs from all decision trees. Essentially multiple training subsets are built from the dataset and a decision tree is constructed for each of these training subsets. With each tree contributing towards voting and eventually majority of the votes determine the final class. This technique is known as random split and the trees are known as random trees.

For the purpose of this study, Random Forest shall be chosen to build a predictive classifier model, as this model gives the best classification accuracy and also due to the high speed of classification, interpretation ability of the knowledge or classifications, and model parameter handlining as indicated by [34].

4) Artificial neural network: Multilayer Perceptron (MLP) is a technique which is trained by the backpropagation algorithm. Essentially a MLP neural network composed of three layers namely; input layer, output layer and many hidden layers. The architecture is a densely connected network where every neuron in a layer is connected to neurons in prior and next layers. The other feature of this network is that there is no activation function in the input layer, but every neuron in the hidden and output layer has an activation function.

The initialization of weights is a random process in the MLP, however the network trains by working out the difference between the computed output and the actual output and adjusts the weights iteratively to this cause of minimizing the residual.

IV. Experimental Results and Discussions

This section briefs the model development stage and details the techniques implemented for this study. The sampling methods including RUS, SMOTE, DBSMOTE and SMOTEENN are implemented, since the dataset is highly imbalanced with 0.17% of positive instances. These methods are based on previous research conducted on the domain and is selected to offer diversity in terms of adopted sampling algorithm and attempts to find out which sampling strategy works best for the given dataset. The aim is also to understand in terms of the strengths of various classifiers and their ability to tackle each sampling strategy.

The class distribution after the dataset was split in to 70% for training set and 30% for holdout set using stratified random sampling. The holdout set contains 148 samples of the fraudulent transactions and will be used to evaluate the performance of all models to maintain consistency in scoring and model benchmarking. A separate hold out set is also the best strategy to adopt to avoid data leaks, which can be a problematic and frequently occurs, while using cross validation along with oversampling. The training set contains 199,020 non-fraud and 344 fraudulent transactions. This set will be used for both data over sampling using SMOTE based techniques, as well as under sampling using RUS.

The Table II shows a summary of class counts after implementing sampling techniques on the original training dataset. In all the cases the final class counts are equal other than the SMOTEEN technique with unequal class counts. This is due to removal of noise using ENN.

TABLE II. SUMMARY OF CLASSES AFTER SAMPLING

Sampling Strategy	Fraudulent	Non-fraud	Total
RUS	344	344	688
SMOTE	199,020	199,020	398,040
DBSMOTE	199,020	199,020	398,040
SMOTEENN	195,374	190,186	385,560

A. *Comparision of Sampling Techniques over Different Classifers*

The training dataset was used to produce four different sampled datasets which were used to train each classifier. Unsampled dataset was used as a baseline for each of the classifier. The evaluation metrics used are F1 score, Precision and Recall.

Distributed Random Forest (DRF), Artificial Neural Network (ANN), Gradient Boosting Machine (GBM), and Stacked Ensemble (SE) are the four classifiers which have been trained on the four different sampling strategies (RUS, SMOTE, DBSMOTE, SMOTEENN). The results provided in this section includes classifier specific results for all the employed sampling strategies. Each classifier is concluded with an overall summary of key evaluation metrics F1 score, Precision and Recall score. These metrics, along with the confusion matrix facilitate to understand how the classifier performed with each of the sampling strategies.

1) Artifical Neural Network (ANN): The ANN architecture in this case was set at 30 in the input layer and 200 neurons in a single hidden layer followed by two layers in the output layer. The activation function used was Rectified Linear Unit (ReLU). The model reached optimal performance at 61 epochs with each epoch iterating over the training dataset. A drop out of 40% was used in the hidden layer so that the model automatically drops the neurons in the hidden layer. The learning rate used for the model was set at 0.005.

The highest F1 score is at 0.8116 on unsampled dataset, which is ideal in the case where precision and recall are of equal importance or significance as the F1 measure is a harmonic mean between the two metrics. However, in the case of FDS, recall is of much more importance then precision. The sampling method, RUS had the highest recall of 0.8311, and the lowest precision score among all the sampled datasets. ANN with unsampled data produced the highest f1 score of 0.8116, although its recall is lower than the second highest recall for ANN with SMOTE at 0.7635 and significantly better precision of 0.8370. Therefore, a better model for ANN is with SMOTE sampling.

2) Distributed Random Forest (DRF): Number of trees was set to 43 with a maximum tree depth of 20. The low tree depth helps lower model complexity while avoiding overfitting. The min rows parameter set as 5 specifies that a minimum of 5 observations is used for each leaf. The sample rate specifies the rate for row sampling which was set at 63%. Column sample rate was set to 0.8, which takes in 80% of columns to construct an individual tree. Lowering the column sample rate will aid in producing diverse trees, which are able to regularize well.

The highest F1 score was produced by the unsampled dataset for the DRF, which was influenced by the highest precision provided at 0.9286. Recall as we consider as the more important and significant metric is at the highest for SMOTE sampling at 0.8176 with a reasonable precision of 0.86. SMOTE sampling is, therefore, the best model for DRF considering the high recall score. It can also be noted that SMOTEENN sampling produced the second-best recall score for DRF classifier. SMOTEENN technique performs data reduction or noise removal using Edited Nearest Neighbour technique which removes any sample which is misclassified by its three nearest neighbours. It is proven with this result, that the noise removal is not very effective as it produced a lower recall score than the original SMOTE sampling.

3) Gradient Boosting Machine (GBM): The number of trees was set to 116 with a maximum tree depth of 15. This allows for reduced model complexity and prevents model from overfitting. Minimum rows to sample for the creation of each tree was set to 100 and column sampling rate set at 0.8, which means that 80% of the columns will be used for each tree.

It is observed that the highest recall score was produced by two sampling methods SMOTE, and SMOTEENN at 0.81. In this case, where two classifiers produce similar recall score, F1 score could be used as a deciding factor since it reflects the model with the best precision. Reducing false flags (False Positive) is an important aspect of an FDS, and thus the model with the highest recall and precision is preferred. Therefore, for the GBM classifier the best model is using SMOTE sampling which resulted in 0.81 recall score and 0.90 precision score. The model with the highest F1 score (DBSMOTE) at 0.86 cannot be considered the best model as it has lower recall score at 0.79 compared to the previously mentioned models, although precision is at the highest at 0.94.

4) Stacked Ensemble (SE): The Stacked Ensemble has very little parameters to define. The SE is a combined model of all trained models (30 models), using a General Linear Model (GLM) as a meta learner to enhance the model performance. The meta learner folds was set to 5 to create a 5-fold cross validated model training with stratified sampling.

Stacked Ensemble model is a Super Learner based on the combinations of ANN, GBM and DRF. The Random Under sampling (RUS) method scores the lowest for the key metric at

0.68 as well as offered the lowest precision score. Highest observed recall score was for SMOTEENN with a combination of SMOTE oversampling and noise removal is using the Edited Nearest Neighbour (ENN) technique. Since this model also offers a reasonable precision score of 0.85 it can be considered as the best sampling strategy for the Stacked Ensemble. Unsampled dataset offered the highest precision score of 0.94 as a result of less noise since it is based on 100% of original data and no synthetic samples were introduced.

B. Summary

The results from all classifiers for each of the sampling methods employed were consolidated based on the performance metrics. The key metric for the domain of FDS are recall which is of the highest priority while also addressing minimal False Positives (FP); i.e.; higher precision. To this end, the primary metric which will be considered is the recall as it is the key metric which is indicative of the total True Positives (fraud cases) detected while minimizing False Negatives (fraudulent transactions classified as non-fraudulent).

The evaluation results were assessed from two perspectives; i) Optimal sampling strategy, ii) Optimal classifier for the domain. The Table III summarizes how various sampling strategies performed comparatively.

TABLE III. PERFORMANCE COMPARISON ACROSS SAMPLING TECHNIQUES

Sampling	**Model**	**F1-score**	**Recall**	**Precision**
UNSAMPLED	ANN	0.8116	0.7568	0.8750
	DRF	**0.8540**	**0.7905**	0.9286
	GBM	0.8284	0.7500	0.9250
	SE	0.8433	0.7635	**0.9417**
RUS	ANN	0.4184	**0.8311**	0.2795
	DRF	**0.7653**	0.7162	0.8217
	GBM	0.7352	0.6284	**0.8857**
	SE	0.7566	0.6824	0.8487
SMOTE	ANN	0.7986	0.7635	0.8370
	DRF	0.8403	**0.8176**	0.8643
	GBM	**0.8541**	0.8108	**0.9023**
	SE	0.8459	0.7973	0.9008
DBSMOTE	ANN	0.8029	0.7568	0.8550
	DRF	0.8467	0.7838	0.9206
	GBM	**0.8603**	**0.7905**	**0.9435**
	SE	0.8509	**0.7905**	0.9213
SMOTEENN	ANN	0.7985	0.6959	**0.9364**
	DRF	0.8351	0.8041	0.8686
	GBM	**0.8451**	**0.8108**	0.8824
	SE	0.8305	**0.8108**	0.8511

The key metrics recall score is considered as a first step for identifying the best sampling strategy. RUS has the highest observed recall score of 0.83 with ANN classifier. However, this was not chosen to be the best model since it offered very little precision of 0.27. This means that while most of the fraudulent transactions are detected by the system it also falsely flagged several genuine transactions as fraudulent. Fraud Detection System is mostly concerned with increasing True Positives it must also consider to be precise in this detection by reducing the number of False Positive.

The second highest recall score was then considered with SMOTE sampling strategy by DRF classifier at 0.81. Precision score for this classifier is observed to be 0.86, which is significantly better than the RUS by ANN. Therefore, SMOTE method can be considered a better sampling strategy to adopt. It is also observed that SMOTE with GBM classifier also offers a high recall which was the third highest recorded at 0.81 while offering even higher precision then the SMOTE with DRF at 0.90.

SMOTEENN sampling is another technique which offered promising results and performed consistently with all classifiers except for ANN. The recall scores for most of the models been at 0.81 while yielding a good precision score above 0.85 in all the cases.

Assessing the average performance of the sampling strategy across various classifiers gives an indication of the best overall sampling strategy to adopt. In a diverse classifier domain such as FDS the average performance of the sampling strategy is very much indicative of its generalizability in terms of adopting well for other datasets in the field. Adopting no sampling strategy resulted in the worst average recall scores while SMOTEENN sampling strategy offered the best average recall score at 0.79. SMOTE and DBSMOTE have the same average score of 0.78, although SMOTE produced the best classifier. The average score considerably dropped for SMOTE due to a very low recall of 0.76 with ANN.

V. CONCLUSION

Detection of credit card fraud is classified as a cost-sensitive problem, where there is an associated cost incurred for incorrectly classifying a genuine transaction as fraudulent and incorrectly classifying fraudulent transaction as genuine. In the absence or no occurrence of fraud, there is no associated administrative costs incurred by the financial institution. However, failure to detect the fraud is a loss of the particular transaction amount. It is thus, an important proposition to incorporate in to the FDS, particularly in the development of models on class imbalanced datasets. There is an associated cost with False Positives, where genuine transactions are flagged as fraud. However, the cost associated with the inability to identify a fraudulent transaction can be immense in contrast. Therefore, recall score was used as key metric as the target of the FDS is to maximize the True Positive Rate.

A base model was implemented using an unsampled dataset, followed by the implementation of four different sampling strategies. Four different classifiers including a Super learner (Stacked Ensemble) was used for each of the sampled datasets to train the models. Distributed Random Forest (DRF), Artificial Neural Network (ANN), Gradient Boosting Machine (GBM) and Stacked Ensemble (SE) are the four classifiers which have been trained on the four different sampling strategies (RUS, SMOTE, DBSMOTE, SMOTEENN). Each classifier is evaluated based on the overall summary of key evaluation metrics F1 score, Precision and Recall score.

The findings of this study indicate promising results with SMOTE based sampling techniques. The best recall score obtained was with SMOTE sampling strategy by DRF classifier at 0.81. Precision score for this classifier was observed to be 0.86. Therefore, SMOTE method can be considered a better sampling strategy to adopt.

Stacked Ensemble was trained for all the sampled datasets and found to have the best average performance at 0.78 with the second-best average for GBM classifier. ANN suffered with the worst recall score, which may be due to the high level of noise generated by the synthetic samples. The Stacked Ensemble model has shown promise in the detection of fraudulent transactions across majority of the sampling strategies.

A. Future Recommendation

Although the study was conducted to address the major problems in the domain of predicting fraudulent transactions, the limitations of the study with respect to time and resources contributed to selection of limited number of sampling strategies. Several other sampling strategies may be considered as an avenue for further research to improve the classifier performance.

Although un-supervised machine learning was not covered within the scope of this study it is still a promising area to be explored. This study may further be improved with the implementation of semi-supervised or un-supervised learning techniques such as one-SVM, k-means clustering and Isolation Forests.

Research can also be further expanded in identifying optimum thresholds for identifying the cut-off points to maximize the Recall score while finding the right balance between Precision and Recall could also yield potentially good results.

REFERENCES

[1] S. Manlangit, S. Azam, B. Shanmugam, K. Kannoorpatti, M. Jonkman, and A. Balasubramaniam, “An Efficient Method for Detecting Fraudulent Transactions Using Classification Algorithms on an Anonymized Credit Card Data Set,” in Intelligent Systems Design and Applications, 2017.

[2] K. T. Hafiz, S. Aghili, and P. Zavarsky, “The use of predictive analytics technology to detect credit card fraud in Canada,” Iber. Conf. Inf. Syst. Technol. Cist., vol. 2016-July, 2016.

[3] European Central Bank, “Fifth report on card fraud, September 2018,” 2018.

[4] AP NEWS, “SmartMetric Reports Worldwide Payment Card Fraud Losses Reach a Staggering $24.26 Billion While the USA Accounts for 38.6% of Global Card Fraud Losses,” BusinessWire, New York, Jan-2019.

[5] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, “Credit card fraud detection and concept-drift adaptation with delayed supervised information,” Proc. Int. Jt. Conf. Neural Networks, vol. 2015-Septe, 2015.

[6] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, “Credit card fraud detection using machine learning techniques: A comparative analysis,” Proc. IEEE Int. Conf. Comput. Netw. Informatics, ICCNI 2017, vol. 2017-Janua, pp. 1–9, 2017.

[7] A. Dal Pozzolo, O. Caelen, Y. A. Le Borgne, S. Waterschoot, and G. Bontempi, “Learned lessons in credit card fraud detection from a practitioner perspective,” Expert Syst. Appl., vol. 41, no. 10, pp. 4915–4928, 2014.

[8] C. Jiang, J. Song, G. Liu, L. Zheng, and W. Luan, “Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism,” IEEE Internet Things J., vol. 5, no. 5, pp. 3637–3647, 2018.

[9] A. Correa Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, “Feature engineering strategies for credit card fraud detection,” Expert Syst. Appl., vol. 51, no. January, pp. 134–142, 2016.

[10] R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin, “An Empirical Study on Class Rarity in Big Data,” Proc. - 17th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2018, pp. 785–790, 2019.

[11] C. X. Zhang, S. Xu, and J. S. Zhang, “A novel variational Bayesian method for variable selection in logistic regression models,” Comput. Stat. Data Anal., vol. 133, pp. 1–19, 2019.

[12] S. H. Ebenuwa, M. S. Sharif, M. Alazab, and A. Al-Nemrat, “Variance Ranking Attributes Selection Techniques for Binary Classification Problem in Imbalance Data,” IEEE Access, vol. 7, pp. 24649–24666, 2019.

[13] G. E. Melo-Acosta, F. Duitama-Munoz, and J. D. Arias-Londono, “Fraud detection in big data using supervised and semi-supervised learning techniques,” 2017 IEEE Colomb. Conf. Commun. Comput. COLCOM 2017 - Proc., 2017.

[14] S. C. Satapathy, K. Srujan Raju, J. K. Mandal, and V. Bhateja, “Proceedings of the second international conference on computer and communication technologies: IC3T 2015, Volume 1,” Adv. Intell. Syst. Comput., vol. 379, pp. 681–689, 2016.

[15] C. K. Maurya, D. Toshniwal, and G. V. Venkoparao, “Online anomaly detection via class-imbalance learning,” 2015 8th Int. Conf. Contemp. Comput. IC3 2015, pp. 30–35, 2015.

[16] A. Charleonnan, “Credit card fraud detection using RUS and MRN algorithms,” 2016 Manag. Innov. Technol. Int. Conf. MITiCON 2016, pp. MIT73–MIT76, 2017.

[17] A. Kumar and G. Gupta, “Fraud Detection in Online Transactions Using Supervised Learning Techniques,” 2018.

[18] L. T. Nghiem, T. T. Thu, and T. T. Nghiem, “MASI: Moving to adaptive samples in imbalanced credit card dataset for classification,” 2018 IEEE Int. Conf. Innov. Res. Dev. ICIRD 2018, no. May, pp. 1–5, 2018.

[19] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang, “Random forest for credit card fraud detection,” ICNSC 2018 - 15th IEEE Int. Conf. Networking, Sens. Control, pp. 1–6, 2018.

[20] S. Xuan, G. L. B, and Z. Li, “Refined Weighted Random Forest and Its Application to Credit Card Fraud Detection,” vol. 11280, pp. 343–355, 2018.

[21] A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams, and P. Beling, “Deep learning detecting fraud in credit card transactions,” 2018 Syst. Inf. Eng. Des. Symp. SIEDS 2018, pp. 129–134, 2018.

[22] A. M. Mubalaike and E. Adali, “Deep Learning Approach for Intelligent Financial Fraud Detection System,” UBMK 2018 - 3rd Int. Conf. Comput. Sci. Eng., pp. 598–603, 2018.

[23] I. Benchaji, S. Douzi, and B. El Ouahidi, Using genetic algorithm to improve classification of imbalanced datasets for credit card fraud detection, vol. 66. Springer International Publishing, 2019.

[24] H. He and E. Garcia, “Learning from imbalanced data,” Ieee Trans. Knowl. Data Engin, vol. 21, no. 9, pp. 1263–1284, 2009.

[25] Di. S. Sisodia, N. K. Reddy, and S. Bhandari, “Performance evaluation of class balancing techniques for credit card fraud detection,” IEEE Int. Conf. Power, Control. Signals Instrum. Eng. ICPCSI 2017, pp. 2747–2752, 2018.

[26] P. Xenopoulos, “Introducing DeepBalance: Random deep belief network ensembles to address class imbalance,” Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, vol. 2018-Janua, pp. 3684–3689, 2018.

[27] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, “Credit Card Fraud Detection Using AdaBoost and Majority Voting,” IEEE Access, vol. 6, pp. 14277–14284, 2018.

[28] K. Fu, D. Cheng, Y. Tu, and L. Zhang, "Credit Card Fraud Detection Using Convolutional Neural Networks," J. Soc. Mech. Eng., vol. 90, no. 823, pp. 758–759, 2017.

[29] H. Liu and M. Zhou, "Decision tree rule-based feature selection for large-scale imbalanced data," 2017 26th Wirel. Opt. Commun. Conf. WOCC 2017, pp. 1–6, 2017.

[30] ULB Machine Learning Group, "Credit Card Fraud Detection," 2016.

[31] U. R. Salunkhe and S. N. Mali, "Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach," Procedia Comput. Sci., vol. 85, no. Cms, pp. 725–732, 2016.

[32] M. Zamini and G. Montazer, "Credit Card Fraud Detection using autoencoder based clustering," 2018 9th Int. Symp. Telecommun., pp. 486–491, 2019.

[33] H2O AI TEAM, "H2O AI," 2016.

[34] S. J. Omar, K. Fred, and K. K. Swaib, "A state-of-the-art review of machine learning techniques for fraud detection research," 2018 IEEE/ACM Symp. Softw. Eng. Africa, pp. 11–19, 2018.

Usability Evaluation of a Tangible User Interface and Serious Game for Identification of Cognitive Deficiencies in Preschool Children

Sánchez-Morales, A.[1], Durand-Rivera J.A.[2], Martínez-González C.L.*[3]
Programa de Posgrado en Ingeniería de Sistemas, SEPI ESIME-Z, Instituto Politécnico Nacional, México[1,3]
División de Neurociencias, Laboratorio de Neuroprotección, Instituto Nacional de Rehabilitación, México[2]

***Abstract*—Detecting deficits in reading and writing literacy skills has been of great interest in the scientific community to correlate executive functions with future academic skills. In the present study, a prototype of a serious multimedia runner-type game was developed, *Play with SID*, designed to detect deficiencies in cognitive abilities in preschool children (sustained attention, memory, working memory, visuospatial abilities, and reaction time), before learning to read and write. Usability tests are used in Human-Computer Interaction to determine the feasibility of a system; it is the proof of concepts before the development of real systems. The aim of this paper was to evaluate the usability of the interface of the serious game, as well as the tangible user interface, a teddy bear with motion sensors. A usability study using the Wizard of Oz technique was conducted with 18 neurotypical preschool participants, ages 4 to 6. Concepts related to interactivity (interaction, the fulfillment of the activity objective, reaction to stimuli, and game time without distraction) were observed, as well as eye-tracking to assess attention and the Usability Scale System (SUS) to measure usability. According to the usability evaluation (confidence interval between 74.74% and 90.47%), the prototype has good to excellent usability, with no statistically significant differences between the age groups. The observed concept with the highest score was the game time without distraction. This characteristic will allow evaluating sustained attention. Also, we found out that the tangible interface use leads to the observation of laterality development, which will be added to the design of the serious game. The use of observation-based usability assessment techniques is useful for obtaining information from the participants when their communication skills are developing, and the expression of their perception in detail is limited.**

Keywords—User interface; wizard of Oz; usability; HCI; input device

I. Introduction

Cognitive skills related to reading and writing (reading and writing literacy), such as working memory, verbal comprehension, processing speed, and perceptual reasoning [1] have been identified as determining factors for the personal and social development of an individual [2]. Detecting deficits in these skills has been of great interest in the scientific community to correlate executive functions, processed in the prefrontal cerebral cortex, with future academic skills [3].

The term used to refer to a child who acquires literacy skills is *late-emerging poor reader* and was proposed by Chall, who determined that the deficit increased as the student progressed in his academic life [4]. The identification of these deficiencies is commonly carried out at a stage when students have already faced problems related to poor school performance [5].

There are studies on the use of multimedia technology for therapeutic purposes to detect cognitive deficiencies and improve them [6][7], which have shown encouraging results, specifically with the use of serious games. These games are characterized by having implicit objectives, in addition to the explicit ones of the game, such as learning or developing skills [8].

On the other side, tangible user interfaces (TUI) are used to improve existing learning tasks and an alternative to graphical user interfaces (GUI) to allow the user to control or navigate in a system with physical objects [9].

A. Similar Works

Among similar works, Valladares-Rodríguez et al. [10] developed games to detect cognitive deficits in older adults, specifically to link them to early detection of Alzheimer's Disease. Jung et al. [11] reported a remote assessment of cognitive disability with a mobile game. Tong and Chignell [12] proposed interesting recommendations for the development of serious games for cognitive assessment. To date, no serious game was found aimed at detecting cognitive deficiencies for literacy in children.

Shamilov et al. [13] developed a computer game with a tangible interface that verifies that the sense of touch increases the user's attention and participation in an activity. Schneider et al. [14] proposed a learning system based on a tangible user interface and complemented it with learning with traditional materials, they observed that the participants who first used the TUI and then studied the texts, performed better than those who first read and then they used the system.

B. Usability of Serious Games

As well as for any technological development, in serious games, usability is one of the most relevant aspects to determine if it can be used by specific users to achieve certain goals, with effectiveness, efficiency, and satisfaction, in a certain context of use [15]. This feature is intrinsically related to user-centered design and human-centered design processes and its most relevant activities are summarized in 1) the understanding and specification of the context of use, 2) the specification of the user and the organizational requirements,

3) the production of design solutions and 4) the evaluation of the designs concerning the requirements [2].

Assessing usability early in technology development is important, although this process is iterative [16]. Usability tests are used in Human-Computer Interaction to determine the feasibility of a system; it is the proof of concepts before the development of real systems [17]. These tests are considered as user research, although their main objective is not aimed at the user itself, but on learning about the participants, the use of the interface and the possible technologies that can be used [18].

Some specific cases of usability evaluation, according to the profile of the user, are children, elderly adults, and people with disabilities. In these cases, user-centered design is more than fundamental to the utility and usability of the application. Some of the conventional usability assessment methods have been adapted for these types of user profiles. Cano et al. [19] applied an evaluation method for the user experience of serious games of children with a cochlear implant, where not only is the user pediatric, but also has a type of disability. In that same sense, but considering the context of use, Sun et al. [20] evaluated the usability of a mobile application for pain management in children, in a hospital environment.

Among the methods for the usability evaluation, there are different techniques to be used according to the purpose of the evaluation, the type of prototype to be evaluated, and the characteristics of the participants, among others. Analyzing the needs of the evaluation allows identifying the appropriate method to carry out the usability test [21].

The types of prototypes for these tests are paper prototypes, diagrams of the screens without any functionality or with partial functionality, prototypes that appear to be functional, but a human reply behind the computer. Also, the tests can be carried out with final software versions, before its launch or with systems already implemented [18]. Regarding serious games, Olsen, Procci y Bowers [22] emphasize that evaluation should be carried out from the paper version [23] since the implicit objective of the game must be revised considerably.

In the present study, a prototype of a serious multimedia runner-type game was developed, *Play with SID* (SID for the acronym in Spanish for deficiency identification system), designed to detect deficiencies in cognitive functions in preschool children, before learning to read and write. The aim of this paper was to evaluate the usability of the interface of the serious game, as well as the tangible user interface, a teddy bear with motion sensors.

II. Materials and Methods

A. Design of the Evaluated Prototype

The design of the serious game for the identification of cognitive deficiencies in preschool children and its control interface was obtained from a study based on rapid contextual design and participatory design techniques, reported in [24]. The design of the exercises to evaluate different cognitive skills, such as sustained attention, memory, working memory, visuospatial abilities, and reaction time involved in literacy, is described in a report to be published.

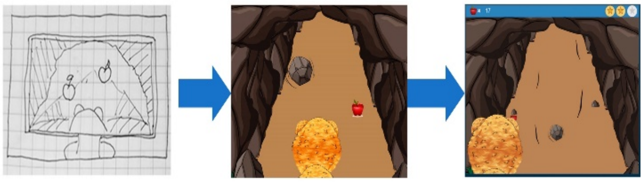


Fig 1. Serious Game Prototype Design Process.

The present study was limited to the use of one of the designed exercises, which evaluates attention and visuospatial ability. A low fidelity prototype was used, made with an authoring tool, as prototyping games using authoring tools is fast and provides immediate feedback [25]. This evaluation gives the possibility of identifying how users interact with the system, testing the controls or means of interaction, and discovering the reactions of the participants to the characteristics of the prototype.

The design process for the serious game prototype is shown in Fig. 1. It consists of a runner-type game, in which a character advances without stopping on a track that pretends to be endless, with limited movements.

The objective of the game is to walk the track, collecting as many apples as possible, and avoiding obstacles with the movement of the character shown on the screen, the bear SID. The role of the instructor is a bee, which gives the indications of what the user should do in the game (Fig. 2).

The process of designing the prototype of the tangible user interface for the game is shown in Fig. 3. This input device is a teddy bear with motion sensors, whose appearance is the character of the game interface.

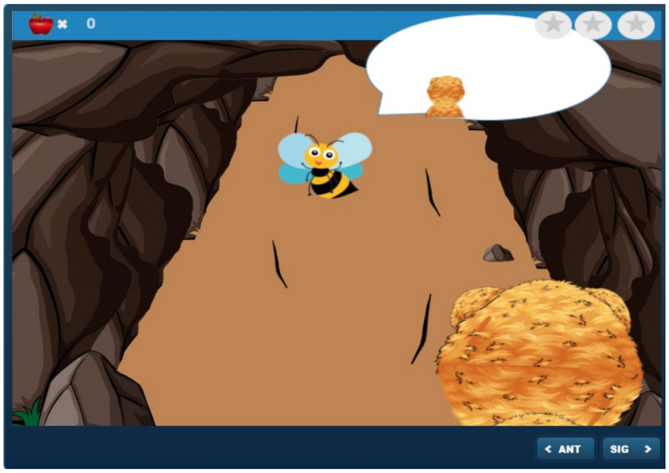


Fig 2. Serious Game Exercise Interface.

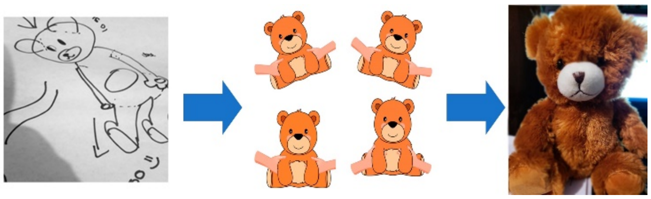


Fig 3. The Design Process of the Prototype of a Tangible User Interface.

B. Study Design

A usability study was conducted with 18 participants whose inclusion and exclusion criteria were as follows: preschool students between 4 and 6 years old, who had not been previously diagnosed with any cognitive deficiency or motor disability and who did not suffer from allergies due to textiles of the tangible user interface (even so, different t-shirts for the teddy bear were used, made with hypoallergenic fabric).

The recruitment was carried out in a public school of preschool level in the metropolitan area of Mexico City. Table I depicts the demographic information of the participants. For a usability test, only 5 participants are needed to find approximately 80% of the problems of using an interface [26], this number is recurrent since it is stated that the number of failures found may depend more on the type of tasks and the design, than on the number of users [27].

C. Data Collection

The Human-Computer Interaction technique called Wizard of Oz was used; this technique mainly consists of simulating functionality that has not yet been developed. [28]. The user perceives that he is interacting with the system when, in reality, he is interacting with a human being (magician or evaluator) who is the one who provides the answers [29]. This technique allows evaluating a prototype before the development stages.

The test consisted of playing the game using the teddy bear or the tangible user interface as an input device. With the movement of the bear, the participants controlled the virtual character within the game.

The movements were made from the observation of the children, by the human wizard, who simulated the control movements. The context of the use was a preschool classroom, in an environment without distractors, as the serious game system would be used. Regardless of this non-threatening environment [30], a non-participatory observation was made by a teacher and some parents. A computer, a monitor, and a web camera were used. Also an additional camera also recorded the test. The layout of the installation is shown in Fig. 4(a)-(d).

TABLE I. GENDER AND AGE DISTRIBUTIONS OF PARTICIPANTS

	Age distribution			Gender		Grade	
	4 year-old	5 year-old	6 year-old	Male	Female	2nd	3rd
Number	4	10	4	7	11	2	16
%	22.22	55.55	22.22	38.88	61.11	11.11	88.88

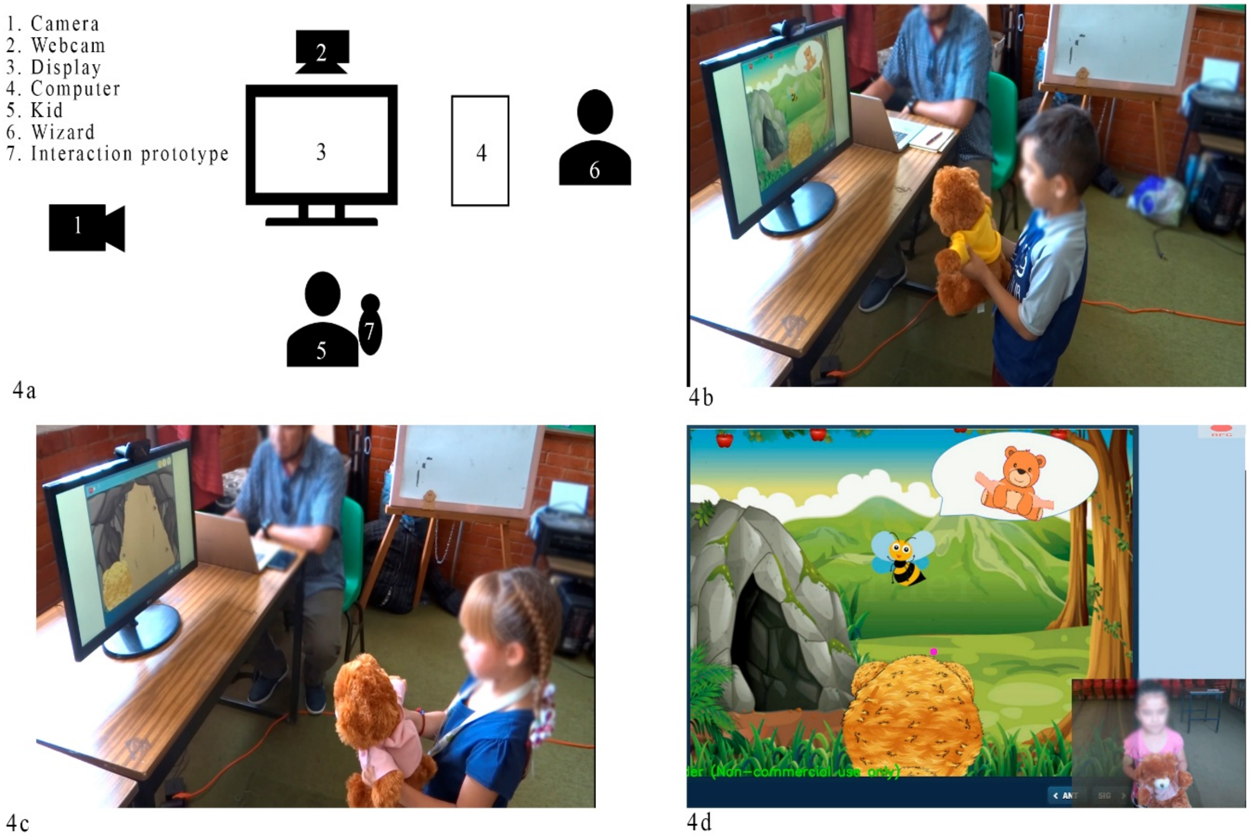


Fig 4. Installation Diagram for the Wizard of Oz Technique.

After explaining the test procedure to the parent or guardian, the child was instructed to play a game on the computer, using the bear. Subsequently, the informed consent of both was requested, also for the recording of the test. The teddy bear was given to the child as the input device, and he/she was given a T-shirt in the color of his/her choice. During the test, the instructor (bee) shows how to use the control (teddy bear) with animations, no more detailed indications of the exercise were given. The average test time with each participant was 10 minutes.

D. Observed Concepts

Four concepts were observed during the usability test to review the interactivity of the serious game interface. These are described in Table II.

The observation of each participant was carried out during the test, and each session was videotaped, to be evaluated later. An evaluation scale of 0 to 2 was proposed to quantify the observation data, based on [31], where 0 is equivalent to the fact that the participant failed to achieve the observed concept, 1 is equivalent to the fact that the participant managed to achieve difficulties and 2 means that the participant managed to achieve without any problem.

E. Usability Measure Instrument

The System Usability Scale (SUS) questionnaire was used to measure usability [32]. It consists of 10 statements (Table III) in which users rate the level of agreement or disagreement; the scores are on a scale of 1 to 5, where 1 corresponds to *totally disagree* and 5 to *totally agree*. In the present work, the SUS statements were adapted according to the age range from 4 to 6 years and the evaluated technological development, a serious game with a physical control interface. It was used in a Spanish version.

As it is stated for the evaluation of the SUS, results were carried out on with a scale of 0 to 4, obtained by subtracting a point from the odd statements and for even questions, the number given by the user in the answer must be subtracted from five. The sum of these results must be multiplied by 2.5 to obtain an evaluation percentage. This percentage is interpreted as not acceptable (<50%), marginal (50-70%) or acceptable (> 70%). Among these items, 4 and 10 are usually identified to refer to *learnability* and the rest to *usability* [27].

TABLE II. CONCEPTS OBSERVED DURING THE TEST TO EVALUATE INTERACTIVITY

Observed concept	Description
Interaction with the game	Identify that the movements of the bear control the character within the game
Fulfillment of the objective of the activity	Understand the purpose of the activity presented in the game
Reaction to stimuli	Identify that the necessary action is carried out for each stimulus, either pick apples or dodge obstacles
Game time without distraction	Maintaining attention in the game throughout the duration of the test

TABLE III. AFFIRMATIONS FOR THE SYSTEM USABILITY SCALE (SUS). ADAPTED FROM [33]

Affirmation
1. I would like to continue using this game
2. I found the game very complicated
3. I thought the game was easy to use
4. I think I would need support to be able to use this game
5. I found that the functions of the game were well done
6. I thought there were too many errors in the game
7. I imagine that most children would learn to use this game very quickly
8. I found the bear very difficult to use
9. I felt calm using the game
10. I needed to learn a lot of things before I could go on with the game

F. Eye-Tracking

Eye-tracking is a technique that allows evaluating eye movements and their sequence to understand the processing of the information received from the screen and the behavior during a usability test [34]. It has also been linked to the point of interest of attention in an interface and has previously been used in the study of serious games [35]. This technique was used to obtain additional information about the interactivity with the prototype. The eye-tracking software used was Gaze Recorder, with the webcam placed on the monitor.

G. Analysis of Data

Kruskall-Wallis tests were run for independent samples to determine statistically significant differences between the age groups for the results of the concepts observed during the test (Table II), as well as for the results of the usability test with the SUS (Table III). Statistical analysis was performed with *SPSS Statistics* software.

III. RESULTS

The results of each user for the concepts observed during the test to evaluate interactivity are depicted in Table IV.

The mean of the evaluation of the observed concepts for all the participants was 1.5 for the interaction with the game (SD = 0.57), 1.5 for the fulfillment of the activity objective (SD = 0.51), 1.61 for the reaction to the stimuli (0.51) and 1.66 (SD = 0.5) for the game time without distraction (SD = 0.59). The scores obtained with the System Usability Scale SUS are shown in Table V.

TABLE IV. RESULT OF THE CONCEPTS OBSERVED TO EVALUATE INTERACTIVITY

Age group	Game interaction	Activity objective	Stimulus reaction	Game time without distraction
4 year-old	Mean: 1.5 SD: 0.57	Mean: 1.75 SD: 0.5	Mean: 1.75 SD: 0.5	Mean: 1.5 SD: 0.57
5 year-old	Mean: 1.3 SD: 0.48	Mean: 1.4 SD: 0.51	Mean: 1.5 SD: 0.52	Mean: 1.7 SD: 0.67
6 year-old	Mean: 2 SD: 0	Mean: 1.5 SD: 0.57	Mean: 1.75 SD: 0.5	Mean: 1.75 SD: 0.5

TABLE V. SYSTEM USABILITY SCALE SCORE

	Mean	SD	Min	Max	Usability 95%-CI
Total	82.61%	15.82	53	100	[74.74%, 90.47%]
Female	79.09%	18.78	55	100	[66.47%, 91.71%]
Male	89.14%	8.33	73	100	[81.43%, 96.85%]

CI = Confidence interval

a)

(a) Confidence Intervals

b)

(b) boxplot

Fig 5. Results of SUS Evaluation by Age Group.

The mean of the SUS test results was 82.61% (SD = 15.82) with the participants in the age range. Fig. 5 shows the results of the evaluation of the prototype by age range. In the group of 4 years, the results were found from 65% to 100% with a mean of 86.5% (SD = 15.15); in the 5-year group, 53% to 100%, with a mean of 82% (SD = 17.34) and finally, the 6-year-old group had a minimum evaluation of 58% and reached a maximum of 95% with a mean of 80.25% (SD = 16.04). One of the volunteers was outside the age range; however, his participation was considered to contrast their answers illustratively, but it was not counted within the sample.

Table VI and Table VII show the results of the Kristall-Wallis test for independent samples between age groups for evaluation with SUS.

TABLE VI. KRISKALL-WALLIS TEST RANKS RESULTS FOR SUS EVALUATION AMONG AGE GROUPS, RANKS

	Age group	N	Mean rank
SUS evaluation	4 year-old	4	10.75
	5 year-old	10	9.50
	6 year-old	4	8.25
	Total	18	

TABLE VII. KRISKALL-WALLIS TEST STATISTICS FOR SUS EVALUATION AMONG AGE GROUPS, TEST STATISTICS A, B

	SUS Evaluation
Chi-square	0.443
df	2
Asymp. Sig	0.801

a. Kruskal-Wallis Test

b. Grouping variable: Age group

Fig. 6 depicts color maps of the zones where the sight of the participants was focused.

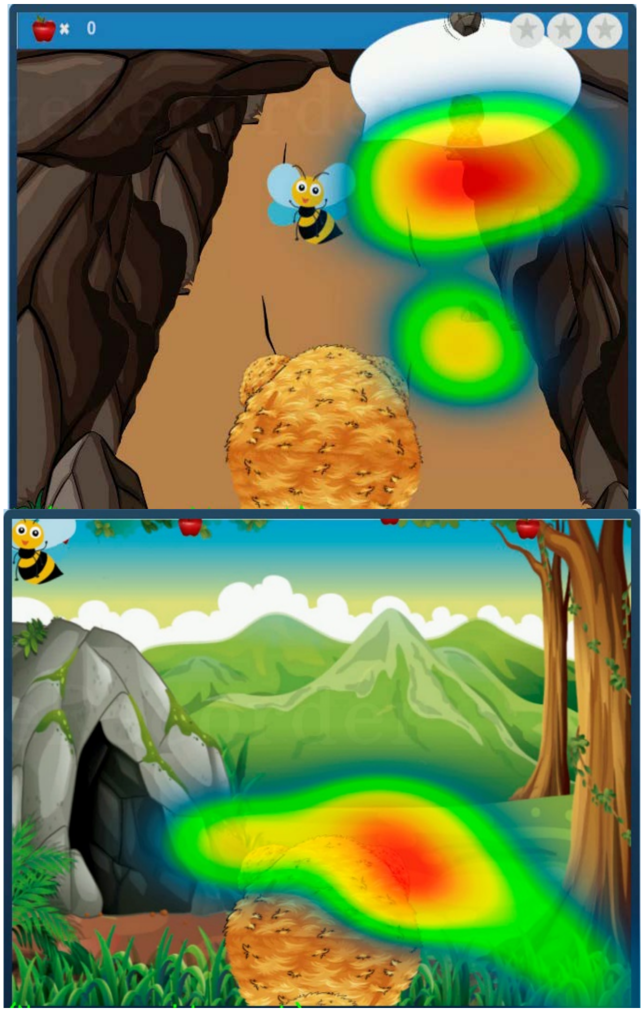


Fig 6. Heat Map of Attention Distribution.

IV. Discussion

The SUS confidence interval, obtained for the evaluation of the sample, between 74.74% and 90.47%, according to the evaluation scale, implies that the usability of the system is in the "acceptable" range, defined in [33]. Therefore, users evaluate the prototype favorably which meets the usability criteria. Regarding the ranking of adjectives, it can be classified between "good" and "the best imaginable", according to [36].

The results of the Kristall-Wallis test for independent samples indicated that there are no statistically significant differences between the age groups in the evaluation of the observed concepts, as well as in the evaluation with the SUS. Therefore, the result obtained for the sample used in this study can be generalized.

It is important to highlight that the only participant outside the age range obtained an evaluation of 0 in the concepts observed for interactivity, presented in Table II (interaction with the game, fulfillment of the activity objective, reaction to stimuli, and time of game without distraction). In contrast, its percentage evaluation of the game with the SUS was 20%, indicating a high score in the item that evaluated the difficulty of the game. Despite being a single participant under the age of 4, it is notorious that the serious game design is not aimed for this age range, under four years.

Regarding eye monitoring, bias was found in the experimental procedure, since not all participants were adequately captured due to the webcam used for this purpose and placement. In the case of child volunteers, with different heights, it is necessary to fix the face to homogenize the calibration of eye-tracking. However, this monitoring allowed studying the area of the game interface in which the attention of the participants was focused, in such a way it could be confirmed that there are distractors not considered in the design. The results obtained showed that there is a relationship of attention in the areas where the stimuli were presented. In the welcome screen where the control of the main character is explained, the attention was focused on the animation of the instructor character (the bee).

Among the limitations of this study, it was focused on evaluating the usability of a serious game using the proposed tangible user interface. Thence, not all the exercises that involve the evaluation of cognitive deficiencies for the detection of cognitive deficits were evaluated.

In this sense, although the SUS usability and learnability subscales have been commonly used according to the items identified for such, Lewis [33] recommended reporting it as a one-dimensional metric.

Regarding the use of SUS in Spanish, it has been successfully used in this language, although there is no validated Spanish version [37], [38]. On the other side, considering the use of the SUS with children, in a major part of these usability studies, children are accompanied by their parent or guardian. In certain applications, adult intervention is necessary, for example, with therapeutic education for children or adolescents and their caregivers [20], [39]–[41]. In our study, the specific user profile was in an age range between 4 and 6 years, and no adult intervened to explain the detail of the activity; there were no difficulties in completing it. Subsequently, no problems were applying the SUS.

V. Conclusions

With the completion of the Wizard of Oz test, the importance of evaluating the usability before the development of the systems can be identified, obtaining information on how to use it, the opinions of the participants, and identify characteristics that require improvement. The observation of the environment where the activity took place within the school and not in a controlled environment, allowed to know in greater detail the technical requirements of the system.

According to the usability evaluation with the SUS questionnaire, the prototype has good to excellent usability, with no statistically significant differences between the age groups.

Regarding observations during and after the usability test, show that participants were favorably evaluated in the observed concepts on the interaction with the game, causing the most conflict to understand each other. We assume that the function of the tangible user interface was not explained intentionally before starting the test, and the volunteers had to learn how to use it on the fly.

This study is part of the iterative design of the prototype. Therefore, the observations made will improve the high-fidelity prototype.

The observed concept with the highest score was game time without distraction, which means that the prototype design allowed participants to maintain their attention for the required time. This characteristic will allow evaluating sustained attention as a cognitive ability. In this regard, one of the participants presented a deficit in the development of laterality, which was observed, since the movement of the bear made it to the opposite side of the required one.

The use of observation-based usability evaluation techniques is useful to obtain information from the participants when they are preschool children because their communication skills are still developing, the expression of their perception in detailed form is limited.

On the other hand, carrying out the usability evaluation in low and medium fidelity prototypes, before development, provides a significant complement, which allows identifying user behavior according to the methodological process and the context of use.

In this study, the characteristics of a serious game and a tactile user interface for children were successfully evaluated. The Wizard of Oz test in such young children is outstanding to test prototypes, allowing them to collect information about users through observation since it is challenging to achieve extensive descriptions at such an early age.

We also successfully applied the SUS to evaluate the usability of a tangible user interface, all along with a serious game in preschool children, without the intervention of the parents or other adults.

As the aim of the study was to evaluate the user interface with the serious game, not all the exercises that involve the evaluation of cognitive deficiencies for the detection of cognitive deficits were evaluated. Though, useful improvements were achieved for the final design, including additional executive functions to be evaluated and the validation of the design for children between four and six-years-old.

Future studies in this subject would include the comparison of different usability measurement tools specifically for tangible interfaces and serious games for children, also for the case of special needs software.

ACKNOWLEDGEMENT

We acknowledge the support of Instituto Politécnico Nacional and Consejo Nacional de Ciencia y Tecnología (CONACyT). We also acknowledge the participants, their parents, teachers, and administrative staff involved in this study.

REFERENCES

[1] T. Tischler, M. Daseking, and F. Petermann, “Cognitive Abilities and Literacy: The Role of Intelligence in Reading Skills,” Kindheit und Entwicklung, vol. 26, no. 1, pp. 48–57, Jan. 2017.

[2] H. Rindermann, C. D. Michou, and J. Thompson, “Children’s writing ability: Effects of parent’s education, mental speed and intelligence,” Learn. Individ. Differ., vol. 21, no. 5, pp. 562–568, Oct. 2011.

[3] S. Shaul and M. Schwartz, “The role of the executive functions in school readiness among preschool-age children,” Read. Writ., vol. 27, no. 4, pp. 749–768, Apr. 2014.

[4] H. W. Catts, D. Compton, J. B. Tomblin, and M. S. Bridges, “Prevalence and nature of late-emerging poor readers.,” J. Educ. Psychol., vol. 104, no. 1, pp. 166–181, Feb. 2012.

[5] F. Cuetos, P. Suárez-Coalla, M. I. Molina, and M. C. Llenderrozas, “Test para la detección temprana de las dificultades en el aprendizaje de la lectura y escritura,” Rev Pediatr Aten Primaria, vol. 17, pp. 99–107, 2015.

[6] R. M. Tomé, J. M. Pereira, and M. Oliveira, “Using Serious Games for Cognitive Disabilities,” 2014, pp. 34–47.

[7] S. Valladares-Rodríguez, R. Pérez-Rodríguez, L. Anido-Rifón, and M. Fernández-Iglesias, “Trends on the application of serious games to neuropsychological evaluation: A scoping review,” J. Biomed. Inform., vol. 64, pp. 296–319, Dec. 2016.

[8] R. Dörner, S. Göbel, M. Kickmeier-Rust, M. Masuch, and K. Zweig, Eds., Entertainment Computing and Serious Games. Dagstuhl Castle, Germany: Springer International Publishing, 2015.

[9] D. Xu, “Tangible User Interface for Children An Overview,” Computing, 2005.

[10] S. Valladares-Rodriguez, M. J. Fernández-Iglesias, L. Anido-Rifón, D. Facal, C. Rivas-Costa, and R. Pérez-Rodríguez, “Touchscreen games to detect cognitive impairment in senior adults. A user-interaction pilot study,” Int. J. Med. Inform., vol. 127, pp. 52–62, Jul. 2019.

[11] H.-T. Jung et al., “Remote Assessment of Cognitive Impairment Level Based on Serious Mobile Game Performance: An Initial Proof of Concept,” IEEE J. Biomed. Heal. Informatics, vol. 23, no. 3, pp. 1269–1277, May 2019.

[12] T. Tong and M. Chignell, “Developing a serious game for cognitive assessment,” in Proceedings of the Second International Symposium of Chinese CHI on - Chinese CHI ’14, 2014, pp. 70–79.

[13] E. Shamilov, N. Gavish, H. Krisher, and E. Horesh, “Tangible user interface,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, vol. 10906 LNAI, pp. 471–479.

[14] B. Schneider, J. Wallace, P. Blikstein, and R. Pea, “Preparing for future learning with a tangible user interface: The case of neuroscience,” IEEE Trans. Learn. Technol., 2013.

[15] N. Bevan, J. Carter, and S. Harker, “ISO 9241-11 Revised: What Have We Learnt About Usability Since 1998?,” in HCI 2015: Human-Computer Interaction: Design and Evaluation, Springer, Cham, 2015, pp. 143–151.

[16] R. S. Pressman, Software engineering : a practitioner’s approach, 6th ed. New York, New York, USA: McGraw-Hill, 2005.

[17] B. Shneiderman and C. Plaisant, Designing the User Interface: Strategies for Effective Human-Computer Interaction (4th edition). Pearson Addison Wesley, 2005.

[18] J. Lazar, J. H. Feng, and H. Hochheiser, Research Methods in Human-Computer Interaction, vol. 278, no. 2. Morgan Kaufmann, 2017.

[19] S. Cano, C. A. Collazos, L. Flórez Aristizábal, C. S. Gonzalez, and F. Moreira, “Towards a methodology for user experience assessment of serious games with children with cochlear implants,” Telemat. Informatics, vol. 35, no. 4, pp. 993–1004, Jul. 2018.

[20] T. Sun et al., “In-hospital usability and feasibility evaluation of Panda, an app for the management of pain in children at home,” Pediatr. Anesth., vol. 28, no. 10, pp. 897–905, Oct. 2018.

[21] P. Markopoulos and M. Bekker, “How to compare usability testing methods with children participants,” Technology, 2002.

[22] T. Olsen, K. Procci, and C. Bowers, “Serious Games Usability Testing: How to Ensure Proper Usability, Playability, and Effectiveness,” Springer, Berlin, Heidelberg, 2011, pp. 625–634.

[23] V. Wattanasoontorn, I. Boada, R. García, and M. Sbert, “Serious games for health,” Entertain. Comput., vol. 4, no. 4, pp. 231–247, Dec. 2013.

[24] A. Sánchez-Morales, C. L. Martínez-González, F. L. Cibrian, and M. Tentori, “Interactive interface design for the evaluation of attention deficiencies in preschool children,” in Proceedings of the 7th Mexican Conference on Human-Computer Interaction - MexIHC ’18, 2018, pp. 1–4.

[25] J. Höysniemi, P. Hämäläinen, and L. Turkki, “Wizard of Oz prototyping of computer vision based action games for children,” in Proceedings of the 2004 Conference on Interaction Design and Children: Building a Community, IDC 2004, 2004, pp. 27–34.

[26] R. A. Virzi, “Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough?,” Hum. Factors J. Hum. Factors Ergon. Soc., vol. 34, no. 4, pp. 457–468, Aug. 1992.

[27] G. Lindgaard and J. Chattratichart, “Usability testing,” in Proceedings of the SIGCHI conference on Human factors in computing systems - CHI ’07, 2007, p. 1415.

[28] D. Salber and J. Coutaz, “Applying the Wizard of Oz technique to the study of multimodal systems,” in Human-Computer Interaction. EWHCI 1993. Lecture Notes in Computer Science Vol. 753, L. J. Bass, J. Gornostaev, and C. Unger, Eds. Springer, Berlin, Heidelberg, 1993, pp. 219–230.

[29] J. Preece, Y. Rogers, H. Sharp, D. Benyon, S. Holland, and T. Carey, Human-computer interaction. Addison-Wesley Pub. Co, 1994.

[30] P. (Panos) Markopoulos, Evaluating children’s interactive products : principles and practices for interaction designers. Morgan Kaufmann, 2008.

[31] L. García, A. Pernett, J. Cano, V. Rafael, and B. Herazo, “Estudio exploratorio de usabilidad para niños de Colombia,” 2017.

[32] P. W. Jordan, Usability Evaluation In Industry. Taylor & Francis, 1996.

[33] J. R. Lewis, “The System Usability Scale: Past, Present, and Future,” Int. J. Human–Computer Interact., vol. 34, no. 7, pp. 577–590, Jul. 2018.

[34] A. Poole and L. J. Ball, “Eye Tracking in HCI and Usability Research,” in Encyclopedia of Human Computer Interaction, IGI Global, 2006, pp. 211–219.

[35] M. Frutos-Pascual and B. Garcia-Zapirain, “Assessing Visual Attention Using Eye Tracking Sensors in Intelligent Cognitive Therapies Based on Serious Games,” Sensors, vol. 15, no. 5, pp. 11092–11117, May 2015.

[36] A. Bangor, P. Kortum, and J. Miller, “Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale,” J. Usability Stud., vol. 4, pp. 114–123, 2009.

[37] Y. N. Ortega-Gijón and C. Mezura-Godoy, “Usability evaluation process of brain computer interfaces,” in Proceedings of the IX Latin American

Conference on Human Computer Interaction, 2019, pp. 1–8.

[38] R. Juarez and V. M. Gonzalez, "Mental Models, Performance and Usability of a Complex Interactive System: The Case of Twitter," in 2013 Mexican International Conference on Computer Science, 2013, pp. 7–12.

[39] C. L. Martínez-González, M. C. C. Camargo-Fajardo, P. Segura-Medina, and P. Quezada-Bolaños, "Therapeutic Patient Education with Learning Objects Improves Asthma Control in Mexican Children," J. Med. Syst., vol. 44, no. 4, p. 79, Apr. 2020.

[40] M. M. Leung, K. F. Mateo, S. Verdaguer, and K. Wyka, "Testing a Web-Based Interactive Comic Tool to Decrease Obesity Risk Among Minority Preadolescents: Protocol for a Pilot Randomized Control Trial.," JMIR Res. Protoc., vol. 7, no. 11, p. e10682, 2018.

[41] A. Bernier et al., "New-Onset Diabetes Educator to Educate Children and Their Caregivers About Diabetes at the Time of Diagnosis: Usability Study.," JMIR diabetes, vol. 3, no. 2, p. e10, 2018.

Comparative Analytics of Classifiers on Resampled Datasets for Pregnancy Outcome Prediction

Udoinyang G. Inyang [1], Imo J. Eyoh[3], Chukwudi O. Nwokoro[5]
Department of Computer Science
Faculty of Science, University of Uyo, Nigeria

Francis B. Osang[2], Adenrele A. Afolorunso[4]
Department of Computer Science, Faculty of Science,
National Open University of Nigeria, Abuja

Abstract—**The main challenges of predictive analytics revolve around the handling of datasets, especially the disproportionate distribution of instances among classes in addition to classifier-suitability issues. This unequal spread causes imbalance learning and severely obstructs prediction accuracy. In this paper, the performances of six classifiers and the effect of data balancing (DB) and formation approaches for predicting pregnancy outcome (PO) were investigated. Synthetic minority oversampling technique (SMOTE), resampling with and without replacement, were adopted for data imbalance treatment. Six classifiers including random forest (RF) were evaluated on each resampled dataset with four test modes using Waikato Environment for Knowledge Analysis and R programming libraries. The results of analysis of variance performed separately using F-measure and root mean squared error showed that mean performance of classifiers across the datasets varied significantly (F=117.9; p=0.00) at 95% confidence interval, while turkey multi-comparison test revealed RF(mean=0.78) and SMOTE (mean=0.73) as having significantly different means. The RF model on SMOTE produced each PO class accuracy ≥0.89, area under the curve ≥ 0.96 and coverage of 97.8% and was adjudged the best classifier-DB method pair. However, there was no significant difference (F=0.07, 0.01; p=1.000) in the mean performances of classifiers across test data modes respectively. It reveals that train/test data modes insignificantly affect classification accuracy, although there are noticeable variations in computational cost. The methodology significantly enhance the predictive accuracy of minority classes and confirms the importance of data-imbalance treatment, and the suitability of RF for PO classification.**

Keywords—Imbalance learning; pregnancy outcome; random forest; SMOTE; imbalance data

I. Introduction

Complications among pregnant women occur frequently and are the obvious sources of maternal mortality (MM) in addition to poor or undesirable pregnancy outcomes (POs). The frequency of MM in developing economies is 50 to 100% higher than those witnessed in developed countries [1]. Pregnancy complications serve as predictors of MMs and other POs (i.e. stillbirth, miscarriage, preterm birth, full term birth etc). Miscarriage, which is an unexpected vaginal flow of blood before twenty-eight (28) weeks of pregnancy, is one of the anomalies noticed among pregnant women especially in Nigeria and other developing countries. Globally, around eighty percent (80%) of maternal deaths and about ninety eight percent (98%) of stillbirths have been linked to direct obstetric complications, like haemorrhage, sepsis, side effects of abortion, preeclampsia and eclampsia, and prolonged obstructed labour [1]. Childbirth complications, maternal infections in pregnancy, maternal syndromes (as pre-eclampsia and diabetes), foetal growth limit and inherited complications are the main reason for the occurrence of stillbirths. Preterm births are associated with multiple pregnancy complications and occurs in 5 to 18% of pregnancies and is also the adjudged cause of infant morbidity and mortality [2].

Improvements in maternal health care systems largely depend on the availability of pieces of knowledge required for the understanding of the effect of pregnancy risks factors, and greatly impact on the future of obstetric health care while attempting to curb maternal morbidity. Although, a significant progress has been recorded in the development of statistical predictive models for PO classification, with better results than clinical tests, there is still room for enhancements in terms of accuracy, interpretability of results and sensitivity to adverse outcomes [3]. Feature ranking and selection, and machine learning (ML) approaches are progressively being utilized for obstetrics outcome classification. However, the suitability of an algorithm to a particular problem domain may affect its performance — accuracy and computational costs. In addition, data from real-world domains are hardly perfect. Some are characterized by uneven distribution of target classes (i.e. some examples of classes may appear more frequently than others) and poses a challenge to data mining (DM) algorithms, as the effectiveness of any DM algorithm is reasonably dependent on the sensitivities to the less frequent (minority) target class [4,5]. Generally, DM algorithms are by default tailored for datasets with equal target class distribution (i.e, they were designed with the assumption of an evenly distributed target class samples), therefore producing poor or below optimal predictive results for the minority target class(es) when imbalanced datasets are encountered. This is because the built model was skewed towards the majority class because of their dominance in the training dataset [4]. The consequences of the class imbalance manifest when the built model is deployed to classify new sets of examples. External influences like missing data, inconsistencies or other forms of noise impact greatly on the imbalanced data distribution, than those that are balanced or near balanced, and produces a noisy classification model [5, 6].

The main focus of predictive modelling, especially in medical researches, is the prediction of the minority target class because of the vital and very useful pieces of knowledge it conveys, despite its paucity in the dataset. Hence the need to adopt methodologies capable of overcoming the class bias issues. Authors in [5], [7] and [8] describe three methods for correcting data imbalance anomaly: (1) data level through resampling, (2) algorithm modification-based approaches, and (3) the cost-sensitive approach. The widely adopted resampling approaches (data level approaches) are based on oversampling and under sampling techniques. This paper aims at determining the best classifier-resampling pair for the prediction of PO using maternal risk factors as predictors. The objectives of this work were twofold; firstly, to compare different resampling techniques based on their ability to address class imbalance and guarantee high accuracy of individual PO classification. Secondly, to assess and perform comparative analysis on six ML algorithms based on their ability to correctly classify PO instances, especially those of the minority class labels. This is achieved by evaluating and comparing classifiers' performances on resampled dataset for the purpose of predicting PO. The remainder of this paper is organized into four sections. Section II gives related works associated with classification methods, dataset imbalance and resampling methods. In Section III, the experimental workflow is described with emphasis on dataset description, pre-processing and resampling, and predictive modelling. The results of the best performing models are described in Section IV while conclusions and future directions are given in Section V.

II. Literature Review

A. Classification and Prediction Models

Classification is a data mining (DM) technique that assigns objects to targeted clusters. Although there are many types of algorithms available in DM for solving medical problems, random forests (RF), k-nearest neighbor (KNN), support vector machine (SVM), decision tree (DT), naïve Bayes (NB), and multi-layer perceptron (MLP) are considered in this paper for pregnancy outcome prediction (POP). SVM has been known to outperform many ML algorithms in many applications, in terms of prediction accuracy and computational cost [9]. Reference [10] employed SVM-based decision support system for preterm birth risks prediction. The model predicted when the birth is likely to occur and the possible outcome for the babies. The authors pointed out that SVM provided an excellent intelligent and comprehensive inference mechanism capable of enhancing the healthcare provided to pregnant women who are at risk with a true positive rate (TPR) of 83.9%, a false positive rate (FPR) of 0.27, and receiver operating characteristic (ROC) area of 0.79. Reference [11] utilized SVM-based decision support system for monitoring the process of child delivery. The data collected include data on heart rate, blood pressure, pulse, uterine contraction, cervical opening and urine volume from pregnant mothers in Indonesia. A total of 40 records were collected based on the earlier listed indicators and tested on the proposed SVM model. For all the selected indicators, an average accuracy of 97.5% was obtained. Author in [12], SVM was used to predict fetal distress using fetal heart rate parameters. A total of 909 data examples with nine parameters were collected and partitioned into 332 normal fetuses, and 577 diagnosis of various pathological conditions. Analysis of results showed that SVM was able to detect fetal distress with an accuracy of 83.0%.

Author in [13] compared two ML algorithms namely; SVM (with linear and non-linear kernels and logistic regression model for the prediction of preterm births. Data for the analysis were collected from a local hospital in India and included age, number of times pregnant, obesity, diabetes mellitus and hypertension with a 10-fold cross validation (10-FCV) for each run. The authors concluded that SVM provided a more accurate prediction with accuracy of 86% compared to the logistic regression model.

Author in [15] proposed a neonatal mortality prediction system using real time medical measurement data based on C4.5 model. The adopted indicators included mean blood pressure, serum, pH, immature/total neutrophil ratio, serum sodium, serum glucose, respiratory rate, heart rate, and pO_2 blood oxygen level. The C4.5 was applied to two sets of data; the summary observations obtained during the initial 12 hours of admission into the neonatal intensive health-care unit by the Canadian Neonatal Network from multiple NICU, and second was the data collected from Children's Hospital of Eastern Ontario (CHEO), Canada and consists of real time medical measurement from a single, out born-only NICU. Analysis of findings revealed that summary data for the first 48 hours of NICU admission provided the best results in the overall with mean sensitivity of 63% and mean specificity of 94%. The authors noted that the results obtained were very significant as the values exceeded the minimum requirement of their clinical partners. Author in [16] DT (C4.5) was applied for the prediction of levels of risk in pregnant women. According to the authors, the C4.5 was adopted because it is powerful, popular, and efficient and can handle the delicate nature of pregnancy problems. For the analysis, 600 pregnant women who went for monthly check-up in Bangalore district hospital were interviewed and two sets of data were obtained namely; unstandardized and standardized pregnancy datasets. The authors concluded that C4.5 classifier provided better results on standardized pregnancy dataset than unstandardized dataset with accuracy of 71.3% and 66.1% respectively. Author in [17] proposed a preterm birth prediction in symptomatic women using DT modelling for biomarkers. The purpose of their study was to use recursive partitioning to identify gestational age-specific and threshold values for infectious and endocrine biomarkers of every pending delivery. The preterm birth predictors considered were white blood cell count, cortisol, maternal age and corticotrophin-releasing hormone. Analysis of results from the DT showed that white blood cell greater than 12,000/mL prior to gestation of 28 weeks and corticotrophin-releasing hormone beyond 28 weeks provided more accurate biomarkers for the prediction of preterm birth within the first 48 hours.

Author in [18] utilized DT, NB, kNN, ANN and SVM for the prediction of high-risk pregnancy cases. The purpose of this study was the timely detection and provision of immediate intervention for these at-risk pregnancy women. In their analysis, DT outperformed other classifiers with accuracy of 97.01%, followed by ANN with accuracy of 93.40%, other classifiers performed in the average with SVM giving the worst accuracy of 76.39% in this context. Author in [19] adopted some DM tools to predict neonatal jaundice caused by hyperbilirubinemia. The aim of the work was to accurately identify neonates at risk of developing severe hyperbilirubinemia in order to offer early medical attention and treatment. Two hundred and twenty seven (227) healthy new-born infants with gestational age ≥ 35 weeks were enrolled for the experiment while bilirubin meter was used for capturing bilirubin levels from the time of birth to hospital discharge. An input space of 72 variables were collected and pruned to 62 via pre-processing. An interval of 8 hours was allowed between measurements for two months (February to March 2011) in Obstetrics Department of the Centro Hospitalar Tâmega e Sousa, E.P.E., North Portugal. The classifiers selected for the analysis included J48, simple classification and regression trees, NB, MLP, sequential minimal optimization (SMO) algorithm and simple logistic available in Waikato Environment for Knowledge Analysis (WEKA). The authors pointed out that only three classifiers namely; NB, MLP and simple logistic correctly predicted neonatal hyperbilirubinemia.

Author in [20] used nine ML tools for the prediction of fetal health status based on maternal clinical history. Ninety six (96) pregnant women between 18 and 41 years in Istanbul were involved for the experiment between January 17, 2015 and February 21, 2017 with 97 fetuses (95 single and 1 twins) and 23 input features. A 10-FCV was employed for training and testing using nine ML algorithms available in Azure ML system. These included averaged perceptron, boosted DT, Bayes point machine, decision forest, decision jungle, logistic regression, ANN and SVM. Result showed that features such as fetal age, age of mother, blood stereotype, test results, number of abortus, number of delivery and any illnesses of mother regarding pregnancy were significant factors that influenced fetal health status. Out of the selected algorithms, the authors pointed out that boosted DT, decision forest and decision jungle produced the best results with accuracy of 89.5%. In conclusion, the authors noted prediction systems are vital tools that could be employed by both clinicians and pregnant women to remotely predict fetal health status in an early stage.

Author in [21] proposed a hybrid system consisting of bijective soft set and back propagation ANN for the prediction of neonatal jaundice. The neonatal jaundice dataset comprising 808 instances with 16 attributes collected from January to December, 2007 in neonatal intensive care unit in Cairo, Egypt, was used for the experiment. The proposed system was compared with bijective soft set, back propagation ANN, MLP, decision table and NB and found to provide the best accuracy of 99.1%. Author in [22] utilized MLP to predict risk of diabetes mellitus that causes several complications during pregnancy. The experimental setting consisted of 394 pregnant women aged 21 years and above, eight attributes and 10-FCV test mode. Results revealed that MLP attained a precision of 0.74, Recall of 74.1%, F-measure (F_m) of 74.1%, and ROC area of 77.9%. The authors concluded that MLP is an excellent tool for predicting gestational diabetes mellitus.

The work reported in [23] employed SVM-based decision support system for preterm birth risks prediction. The SVM model predicted the likely to time birth occur and the possible outcome of babies' status. The results of the empirical experiment showcased SVM as the most performing in terms of intelligent and comprehensive inference mechanism regarding decision support for at risk pregnant women. The result produced true positive rate of 83.9%, a false positive rate of 0.27, and receiver operating characteristic (ROC) area of 0.79. Refs. [24-27] deployed decision support tools that would provide needed assistance to practitioners in ensuring safety of vaginal births after cesarean delivery for women of child bearing age and in the general management of PO. Refs. [28-29] have demonstrated the effectiveness of decision support systems in handling associations between two or more obstetric and neonatal emergencies. A comparative study of machine learning tools and statistical models was reported in [30] for the prediction of postpartum hemorrhage (PH) risks during labour with the aim of minimizing maternal morbidity and mortality. The experiments on data from 12 sites showed that all the models adopted in the study produced satisfactory results, although the extreme gradient boosting model (XGboost) had the best ability to discriminate among PH followed by random forests (RF) and lasso regression model. The effectiveness of ML methods in mining of electronic health data in the domain of atrial fibrillation (AF) induced risks prediction was reported in [31]. Out of a total of 2,252,219 women used for the study, 1,225,533 developed AF during a selected 6-month interval. Two hundred (200) widely used electronic health record features, (age and sex inclusive), and random oversampling approach implemented with a single-layer, fully connected ANN yielded the optimal prediction of six-month incident AF, with an area under the receiver operating characteristic curve (AUC) of 80.0% and an F1 score of 11.0%. The ANN model performed only slightly better than the basic logistic regression consisting of known clinical risk factors for AF, which had 79.4% and 79.0% as AUC and F1 value respectively. The results confirmed the effectiveness of machine learning algorithms in the prediction of AF in patients. The performance of Fuzzy approach, SVM, RF and Naïve Bayes (NB) for the prediction of cardiotocograph‐based labour stage classification from patients with uterine contraction pressure during ante‐partum and intra‐partum period, the proposed algorithm tend to be efficient and effective in terms of visual estimation to incorporate automated decision support system, which will help to reduce high risk of hospitalized patients. Author in [32] experimental results of the impact computational intelligence on the precision of cardiovascular medicine was presented. The method was applied to neonatal coarctation classification and prediction by analyzing genome-wide DNA methylation of newborn blood DNA using in 24 isolated, non-syndromic cases. Six machine learning algorithms including deep learning was used for detection. Deep learning achieved the optimal performance with an AUC and sensitivity of 95% and 98% specificity at 95% confidence interval. The related works considered were based on a single dataset test mode. The significance of this work is the assessment of each classifier on varying dataset test modes.

B. Data Resampling Approaches

Real-world modelling problems are characterized by uneven target class spread. These domains include but not limited to fraud, medicine, spam, web, telecommunications, education and churn customers. In the medical domains like obstetrics, the frequency of desirable outcomes is usually higher than the adverse ones, thereby resulting in a data imbalance problem (DIP). During model building, the infrequent target class(es) have limited representation in the built model because of paucity of training samples, and therefore lacks the classification and prediction competences regarding such class(es). The degree of imbalance is measured by the imbalance ratio (IR) — the ratio of the frequency of observations in the majority class to the tally of instances in the minority class [33], [34]. The notational description of DIP is as follows [35]. Given a dataset Q with m examples and *n* attributes, where $Q = \{x_i, y_i\}$, $i = 1,2, \dots, m$, and where $x_i \in X$ is a data-point in the attribute set $X = \{b_1, \dots, b_n\}$, and $y_i \in Y$ is an element in the set of target classes $Y = \{1, \dots, c\}$. A subset of the desirable (majority class) instances $G \subset$X, and subsets of minority class (adverse instances) $U \subset X$, where $|G| < |U|$. The preprocessing via resampling applied the maternal dataset has the goal of balancing the training and testing sets Q such that$|G| \equiv |U|$.

Since DM algorithms were designed to learn from balanced class training representatives, they produce models that are less equipped for classification of instances in minority class(es) whereas a good coverage is recorded for majority class elements, when confronted with DIP[5], [8]. Although three approaches — resampling, algorithm modification and cost-sensitive approaches, are recommended for imbalance anomaly correction [5], this paper investigated the effect of resampling methodologies on predictive performance of some selected DM classifiers. The rationale for choosing resample approach is due to its simplicity, cost efficiency and classifier independence. Resampling methods operate either by adding elements to the minority class (oversampling) or reducing representatives of the majority class (undersampling). It can also combine both oversampling and undersampling approaches [34],[36]. Synthetic minority oversampling technique (SMOTE), is an oversampling approach that increases the elements of the minor class(es) by generating simulated data items in the nearness of the existing minority class instances, with the goal of flattening IR. Author in [36] describes two key stages for SMOTE implementation: (1) Clustering data-points based on class labels and finding kNN using euclidean distance between every minority data-point with respect to all other minority data items. The *k* least distance examples are chosen as the nearest neighbours. Euclidean distance (D) between one object with the minority class label (x) and another sample with the minority class label (y) for all features is defined by Eq. 1 [36]. (2) New data-points are constructed by inserting points between any two elements belonging to the minority class. One of its *kNN* will be randomized to be candidates in new data construction process. Thereafter, original minor data element (*x*) and one chosen candidate (*y*) will be used to generate new values among *x* and *y*. The process is defined by Eq. 2

$$D_{x,y} = \sqrt{\sum_{c=1}^{n} (x_c - y_c)^2} \quad (1)$$

$$N_c(x, y) = x_c + t.(x_c - y_c) \; for \; 0 \leq t \leq 1 \quad (2)$$

where $\mathrm{N_c}(\mathrm{x,y})$ is the new data-point, *n* is the number of attributes and *r* is a random number between 0 and 1.

Undersampling approaches generate a subgroup of the original dataset by deleting instances with the majority class label. Random undersampling takes place when observations that are deleted are arbitrarily picked from majority class until the data set becomes balanced whereas informative undersampling adopts available rules to mark items for deletion [37]. However, undersampling techniques seemingly impact the multi-class imbalanced data classification performance negatively if useful instances in each majority class are eliminated [38],[39].

C. *Classifier Evaluation Metrics*

In predictive analytics, it is an essential task to assess the quality of the predictions in order to guide in classifier modelling for the specified problem domain. A contingency or confusion matrix (CM) is usually applied for such purposes, providing not only classification errors and accuracy, but also parameters to compute other measures [8],[35]. CM is actually not a performance measure as such, but the basis for deriving other measures. The basic CM for a binary classifier (Table I) uses four indicators (true positive (TP), false positive (FP), true negative (TN) and false-negative (FN)) to measure the classification performance of both classes independently.

TABLE I. CM FOR A BINOMIAL CLASSIFICATION PROBLEM

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

TP is the number of positive PO instances that are correctly classified while FP is the number of negative PO instances misclassified as positive. FN represents the tally of positive PO instances misclassified as negative whereas the negative instances that are correctly classified are defined by TN. These parameters are represented as percentages; TPR, FPR, true negative rate (TNR), and false negative rate (FNR) and defined in Equations 3 – 6 respectively, as follows.

$$TPR(sensitivity) = \frac{TP}{TP + FN} \quad (3)$$

$$TNR \; (specificity) = \frac{TN}{FP + TN} \quad (4)$$

$$FPR = \frac{FP}{FP + TN} \quad (5)$$

$$FNR = \frac{FN}{TP + FN} \quad (6)$$

TPR (or sensitivity) gives a measure of the proportion of actual positive examples which are correctly classified while FPR is the proportion of actual negative examples of PO which are incorrectly identified as positive PO instances. FNR is the percentage of positive PO instances which are wrongly classified as negative POs while the TNR is the fraction of actual negative PO examples which are correctly classified. Accuracy (ACC) has been the widely used metric [8], [40]. It quantifies the predictive capability of elements in the test dataset. Although, it is easy to implement and interpret, it ignores class distribution and frequently skews in the direction of the majority class. It is therefore not suitable for DIP scenario [35]. Apart from ACC (Eq. 7), there are other derivable measures that consider class inequality in their design — precision, recall and F_m given in Equations 8 - 10 respectively and are suitable when the positive class label is the key issue whereas the ROC and area under the curve (AUC) capture performances of minority and majority classes. Precision is a fraction of the predicted positive POs that are actually positive while F_m defines the harmonic mean between precision and sensitivity. The F_m is a more complete measure because it combines precision and recall.

$$ACC = \frac{TN + TP}{FP + TN + TP + TP} \tag{7}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F_m = \frac{2 \times precision \ \times recall}{precision + recall} \tag{10}$$

The ROC curve is a graph of TPR or sensitivity on the *y*-axis against FPR on the *x*-axis while the extreme values are 0 and 1. The total area enclosed by the ROC curve is described by AUC value and is given in Eq. 11. An AUC value of 100% depicts a perfect classification, the one close to 100% depicts a very good performance, while values lower than 50% depicts performance by chance or luck. Another widely used metric of interest, is the root mean squared error (RMSE) which measures the deviation between the classifier's output and actual values. It is defined in Eq. 12 [40].

$$AUC = \frac{1 + TPR - FPR}{2} \tag{11}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(output_c(i) - actual_c(i))^2} \tag{12}$$

where $output_c(i)$ denotes the prediction probability of instance i, which belongs to class c, and $actual_c(i)$ depicts the actual probability.

III. Experimental Settings

A. Dataset Source and Preprocessing

Data was acquired from secondary health facilities in Uyo, Nigeria. A total of one thousand six hundred and thirty-two (1,632) records were obtained from archives of retrospective observations of pregnant women recorded while they enrolled for antenatal care, with an input feature space of forty-two (42) attributes in excluding the target variable. Some of the attributes include; average maternal age, number of children delivered, previous medical history, abortion, miscarriage, prematurity, previous illness, number of attendances to antenatal care, antenatal registration, and mode of delivery, amongst other features. Attribute cleaning, aggregation and elimination of attributes with only a single domain value was performed. The resultant dataset which had thirty-five (35) attributes were subjected to feature ranking [41] and selection via PCA in WEKA software. Attributes with eigenvalue (EV) scores greater than or equal to unity [42] were thirteen (13) and together accounted for 67.13% variation of the target feature. Table II gives a description of attribute description and rank.

As shown in Table II, the average maternal blood pressure topped the list with EV of 3.86 (11.7% proportion of variance), followed by average maternal weight (EV = 2.77, proportion = 8.39%). The thirteenth rank attribute, average ascorbic acid level accounted for 3.17% variation with eigenvalue score of 1.05.

TABLE II. Rank And Description Of Significant Attributes

Rank	Attrib -ute	Description	EV	Prop-ortion (%)	Cum-ulative (%)
1	Maternal BP	Average maternal blood pressure	3.86	11.69	11.69
2	Maternal Weight	Average maternal weight	2.77	8.39	20.29
3	Hemoglobin Level	Average number of red blood cells count	2.37	7.18	27.47
4	PCV level	Average Packed Cell Volume count	1.92	5.82	33.29
5	Pulse Rate	Average number of heart beats per minute	1.54	4.67	37.67
6	Mode of Delivery	Delivery method vaginal delivery =1; caesarean section = 2	1.42	4.30	42.26
7	Malaria Frequency	Number of times maternal malaria Diagnosis	1.39	4.21	46.47
8	Hepatitis C	Indicates history of hepatitis C disease; presence=1, absence=2	1.26	3.82	50.29
9	Diabetes Status	Maternal Diabetic status non-diabetic=0 type1=1; type2=2, others=3	1.18	3.60	53.89
10	Herbal Ingestion	Use of herbal medicinal products during pregnancy	1.15	3.48	57.37
11	Respiratory disorder	Maternal respiratory disease status; presence=1, absence=2	1.12	3.39	60.76
12	Age	Maternal age during pregnancy	1.06	3.20	63.96
13	Ascorbic acid Level	Average amount of ascorbic acid in the body during pregnancy	1.05	3.17	67.13
14	Pregnancy outcome	Maternal delivery outcome miscarriage = 0; pre-term =1; full-term=2, stillbirth=3	-	-	-

Table II also reveals that PO consists of four distinct classes of instances — miscarriage, preterm, term and stillbirth. Out of the 1,632 records, 198 (12.1%) examples belong to miscarriage class, 65 (4.0%) are representatives of preterm births while 114 instances (7.0%) were stillbirths. Term births had majority of observations with a frequency of 1255 (76.9%) (while observations of preterm, still-births and miscarriage classes together have a tally of 23.1%). The distribution of class examples depicts a severe imbalanced situation where term births is the majority class whereas the other three classes (miscarriage, preterm, term and stillbirth) are in the minority with high IR values as follows; miscarriage (6.3), preterm (19.1) and stillbirth (11.0).

B. Resampling Methodology

The final stage of preprocessing implements three data resampling techniques based on oversampling and undersampling — SMOTE, resample with replacement (RRW) and resample without replacement (RRN). The implementation was performed in WEKA version 3.8.4 using default values of "weka.filters.supervised.instance.resample" function and R library. The distribution of class labels in the resultant datasets (Table III) depicts a substantial reduction in the severity of imbalance in the resampled datasets than the original datset (ORD). There is a uniform spread in the RRW method and a near balance distribution in the SMOTE dataset. The RRN approach randomly eliminated 847 (67.5%) instances of the majority PO class (term births) while other PO classes remained unchanged. The IR of the resampled dataset, given in Table IV and Fig. 1, reveals a maximum inter class IR deviation of 0.43 for SMOTE while zero (0) deviation is observed for RRW dataset.

TABLE III. DISTRIBUTION OF TARGET LABELS DATASET

Dataset Code	Resample Method	Misca-rriage (%)	Pret-erm (%)	Term (%)	Still Birth (%)	Total
ORD	Original data	198 (12.1)	65 (4.0)	1255 (76.9)	114 (7.0)	1632
SMOTE	Oversamp-ling (SMOTE)	1188 (25.5)	1300 (27.9)	1255 (27.0)	912 (25.2)	4655
RRN	Random resampling without replacement	198 (25.2)	65 (8.3)	408 (53.0)	114 (14.5)	785
RRW	Random resample with replacement	408 (25)	408 (25)	408 (25)	408 (25)	1632

TABLE IV. ANALYSIS OF IR IN THE ORIGINAL AND RESAMPLED DATASETS

Dataset Code	Miscarriage (%)	Preterm (%)	Term (%)	Still Birth (%)
ORD	6.3	19.3	1	11.0
SMOTE	1.09	1	1.03	1.43
RRN	2.06	6.27	1	3.58
RRW	1	1	1	1

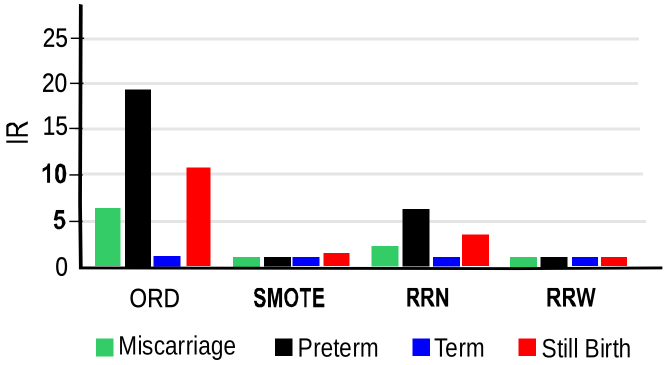


Fig. 1. Visualization of IR for ORD and Resampled Datasets

The RRN dataset has a maximum IR value of 6.27 for preterm class which was hitherto 19.3, while stillbirth drifted to 3.58 from 11.0. This produces a significant balance effect when compared with the ORD dataset.

C. Predictive Modeling and Performance Comparison

The input features correspond to the significant attributes selected during preprocessing with PCA while PO is the target variable. The predictive modeling was performed in WEKA 3.8.4 platform using six classifiers; DT, SVM, KNN, RF, NB and MLP. The default WEKA parameters of each classifier were used for model building and testing processes as follows;

- DT was implemented with C4.5 algorithm with 0.25 as the confidence level, the minimum number of item-sets per leaf was set to 2 while leaf pruning was utilized to get the final tree.
- SVM was trained with John Platt's SMO, Polykernel function, an internal parameter of 1.0 for the exponent of each kernel function and a penalty parameter at 1.0. The model adopted a batch processing mode with a bag-size of 100.
- kNN was invoked through Instance based learning (Ibl) function with one neighbour for returning the output class. Brute force search algorithm was used for nearest neighbours selection based on euclidean distance. The process was iterated with a batch processing size of 100.
- NB parameters were based on weight learning without kernel estimator and supervised discretization functions.
- MLP used backpropagation to learn a multi-layered perceptron. It used a learning rate of 0.3 and momentum of 0.2.
- RF constructed a forest of random trees with an unlimited depth and 100 as the maximum number of iterations.

The models were built and executed with each resampled dataset by adopting four test modes— two based on k-fold cross validation while the other two relied on percentage splitting ratio namely; 10-FCV, 5-fold cross validation (5-FCV), 80% split for training and 20% for testing (80-20) and 70% split for training and 30% for testing (70-30). Since all the classes of PO are of interest in this work, the performances of the classifiers were evaluated based on derivatives from ROC curve — sensitivity, specificity, recall, precision, AUC and other performance measures including kapa statistic (KS), RMSE and CM parameters [43]. Generally, for DIP treatment, measures such as F_m and AUC are recommended rather than traditional classification ACC. Since F_m, combines precision and recall (it eliminates the limitations of other single metric) and also imposes an enhanced inter-class performance equilibrium [44] while RMSE is widely used error measure for classifier evaluation, both measures are more suitable for real-world applications and therefore adopted for the comparative analysis. The overall results (Table V) depict F_m and RMSE values across dataset and classifiers for all the test modes while weighted averages across dataset and classifiers are presented in Tables VI and VII.

The results in Table V, show that the F_m and RMSE values were moderately high for ORD dataset. However due to the imbalance effect, predictions based on the ORD dataset will be biased towards the term births class. The performance of RF on SMOTE dataset ($F_m \geq 0.92$) was the best followed by kNN in RRW ($F_m \geq 0.81$) dataset. In terms of RMSE, the least error value was recorded by RF in SMOTE dataset (RMSE= 0.18) with 10-FCV test mode. The weighted averages in Tables VI and VII, which also appear graphically in Fig. 2 and 3, clearly exposed the performances of the resampled datasets across test modes and classifiers respectively. The average classification result from the resampled dataset is highest with SMOTE dataset ($F_m = 0.73$) in 10-FCV and 80-20 datasets while the least value ($F_m = 0.53$) was attributed to NB in 10-FCV and 80-20 datasets. The performances in terms of both F_m and RMSE for RF and DT are almost the same as evidenced in overlapped trajectories in Fig. 3 and are the topmost performing classifiers while NB is the least performing algorithm.

All F_m and RMSE differences across classifiers and datasets are significant with 95% confidence using a two-way analysis of variance (ANOVA) test in R programming environment. The interaction effect [41] between test modes, resampled datasets and classifiers provided evidence of the existence of a significant interaction between the effects of datasets and classifiers, (F=117.94; p=0.000), while interaction between test mode and other factors yielded no significant effect (F=0.07,0.01;p=1.000 respectively) at 95% confidence level.

A similar result was observed with RMSE as the response variable — classifier and dataset interaction produced (F=17.24; p= 0.00) while interaction involving test modes produced insignificant effects (F = 0.17, 0.14; p= 0.00). This implies that prediction accuracy varies significantly across classifiers and dataset only. Turkey's multiple comparison test [45] showed that the difference between means of dataset pairs and classifier pairs are significantly different. The confidence intervals for SMOTE (mean = 0.73), RF (mean=0.78) are different from others in their respective groups with similar trend significantly observed with RMSE as the dependent variable. However, test modes do not affect the quality of predictions significantly.

TABLE V. PERFORMANCE COMPARISON CLASSIFIERS ON RESAMPLED DATASETS AND TEST MODES

Classifier	Resampling Method	**Test Modes**							
		5-FCV		**10-FCV**		**Train (70%)**		**Train (80%)**	
		F_m	RMSE	F_m	RMSE	F_m	RMSE	F_m	RMSE
NB	ORD	.73	.32	.73	0.32	.70	.33	.71	.32
	SMOTE	.32	.53	.29	.53	.33	.53	.30	.54
	RRN	.59	.38	.60	.38	.61	.38	.59	.41
	RRW	.53	.40	.52	.41	.52	.41	.52	.41
MLP	ORD	.77	.26	.76	.26	.80	.30	.79	.30
	SMOTE	.66	.34	.66	.34	.63	.35	.67	.33
	RRN	.56	.67	.58	.36	.63	.37	.55	.38
	RRW	.58	.37	.60	.37	.60	.36	.58	.36
KNN	ORD	.74	.34	.74	.34	.71	.35	.69	.36
	SMOTE	.89	.23	.89	.23	.88	.24	.88	.23
	RRN	.52	.47	.52	.47	.53	.47	.51	.47
	RRW	**.82**	**.27**	**.84**	**.26**	**.81**	**.29**	**.84**	**.27**
SVM	ORD	.76	.35	.75	.35	.68	.35	.69	.36
	SMOTE	.71	.40	.71	.40	.72	.39	.71	.40
	RRN	.63	.38	.64	.38	.66	.38	.65	.38
	RRW	.50	.40	.50	.39	.49	.40	.49	.39
RF	ORD	.79	.29	.79	.29	.80	.30	.76	.33
	SMOTE	**.94**	**.20**	**.94**	**.18**	**.92**	**.20**	**.93**	**.19**
	RRN	.56	.39	.56	.39	.59	.39	.52	.41
	RRW	**.84**	**.24**	**.85**	**.23**	**.82**	**.26**	**.83**	**.24**
DT	ORD	.76	.33	.81	.26	.78	.31	.80	.30
	SMOTE	.88	.23	.88	.23	.86	.25	.87	.23
	RRN	.58	.36	.58	.37	.70	.35	.55	.37
	RRW	.79	.29	.80	.28	.73	.32	.79	.28

TABLE VI. WEIGHTED AVERAGE OF F_M AND RMSE ACROSS DATASETS

Dataset	**5-FCV**		**10-FCV**		**Train (70%)**		**Train (80%)**	
	F_m	RMSE	F_m	RMSE	F_m	RMSE	F_m	RMSE
ORD	0.76	0.32	0.76	0.31	0.73	0.33	0.74	0.33
SMOTE	0.73	0.32	0.73	0.32	0.72	0.33	0.73	0.32
RRN	0.57	0.45	0.58	0.39	0.63	0.39	0.57	0.40
RRW	0.68	0.35	0.65	0.34	0.63	0.36	0.64	0.34

TABLE VII. WEIGHTED AVERAGE ACROSS CLASSIFIERS

Dataset	**5-FCV**		**10-FCV**		**Train (70%)**		**Train (80%)**	
	F_m	RMSE	F_m	RMSE	F_m	RMSE	F_m	RMSE
NB	0.54	0.41	0.53	0.41	0.54	0.41	0.53	0.42
MLP	0.64	0.41	0.65	0.33	0.67	0.35	0.65	0.34
KNN	0.74	0.33	0.75	0.33	0.73	0.34	0.73	0.33
SVM	0.65	0.38	0.65	0.38	0.64	0.38	0.64	0.38
RF	0.78	0.28	0.79	0.27	0.78	0.29	0.76	0.29
DT	0.77	0.29	0.78	0.27	0.78	0.30	0.77	0.28

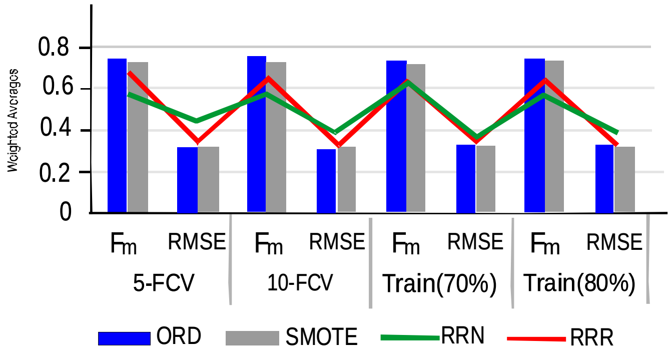


Fig. 2. Graph of Weighted Averages of F_m and RMSE Across Datasets.

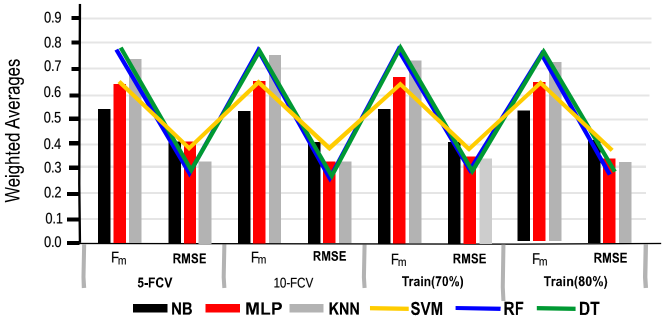


Fig. 3. Graph of Weighted Averages of Fm and RMSE Across Classifiers.

IV. EVALUATION OF RF CLASSIFICATION AND DISCUSSION

The results obtained from PO predictions on all classes using RF classifier and SMOTE dataset in all test mode are reported in Tables VIII and IX.

TABLE VIII. RF CLASS PREDICTIONS WITH SMOTE ACROSS TEST MODES

Test mode	KS	Coverage (%)	ACC (%)	Time (secs)
5-FCV	.89	98.1	92	1.57
10-FCV	.90	97.8	93.1	1.5
Train (70%)	.88	98.3	91.9	0.16
Train (80%)	.89	98.0	93.0	.07

TABLE IX. RF CLASS PREDICTIONS EVALUATION WITH SMOTE

Test mode	Class	TPR	FPR	PR	RE	FM	AUC
5-FCV	Stillbirth	.90	.02	.92	.90	.91	.97
	Term	.87	.06	.84	.87	.86	.96
	Preterm	.96	.01	.98	.96	.97	.99
	Miscarriage	.92	.02	.93	.92	.93	.99
10-FCV	Stillbirth	.94	.02	.93	.91	.92	.96
	Term	.89	.06	.85	.89	.87	.96
	Preterm	.96	.01	.98	.96	.97	.99
	Miscarriage	.93	.02	.94	.93	.94	.99
Train (70%)	Stillbirth	.88	.03	.88	.88	.88	.97
	Term	.88	.07	.82	.88	.85	.95
	Preterm	.96	.01	.98	.96	.97	1.0
	Miscarriage	.91	.01	.96	.91	.93	.99
Train (80%)	Stillbirth	.86	.02	.92	.86	.89	.96
	Term	.89	.07	.82	.89	.85	.95
	Preterm	.96	.01	.97	.96	.96	.99
	Miscarriage	.93	.013	.96	.93	.95	.99

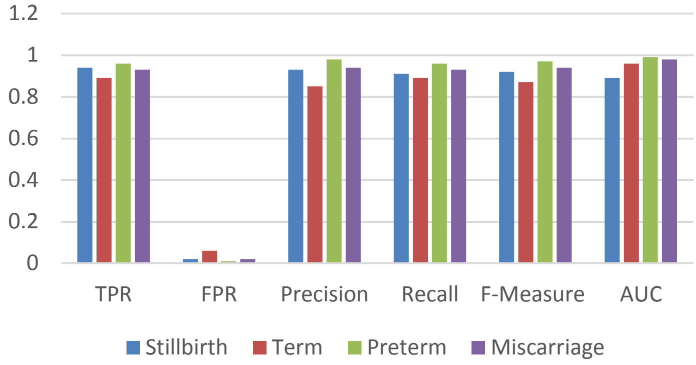


Fig. 4. Graph of RF 10-FCV Performance of PO Classes with SMOTE.

Excellent coverage of instances is expressed in the results with (ACC ≥ 91.9% and coverage ≥ 97.8%) across the test modes. The time used ranges from 0.07 seconds to 1.57 seconds with 80-20 dataset split having the least time due to the number of testing instances used. Average class predictions were greater than 91.8% with 10-FCV having the highest average ACC of 93.1% although computationally expensive. As shown in Table IX, preterm class has the highest sensitivity of 96% in all test modes while the least score is observed for Term class in all test modes except 10-FCV (89%). A similar trend is observed for F_m where term birth earned the least score of 85% in both Train (70%) and Train (80%) test modes. All the performance measures reported in this work depict very good results therefore confirming the suitability of the approach.

The summary of RF predictions using 10-FCV test mode — since it had the highest ACC and KS values (Table VIII) and least classification error (RMSE=0.18) as shown in Table V, is given in Fig. 4 while associated ROC curves for all the PO classes are presented in Fig. 5 to 8.

The sensitivity, precision, recall and F_m for all classes are excellent (ACC ≥ 89). The RMSE=0.01 is least for preterm births in almost all the test modes while term-birth was the least performing class — RMSE = 0.07. The AUC for each class of PO (still birth = 96.36%, term birth = 95.84%, preterm = 99.12% and miscarriage = 98.76%) depict an enhancement in the results as compared to the ORD dataset. The sensitivity also recorded very good results in all categories of PO (still =94%, Term =89%, preterm =96%, miscarriage =93%) with insignificant FPR in each class (0.01≤ FPR ≥0.06).

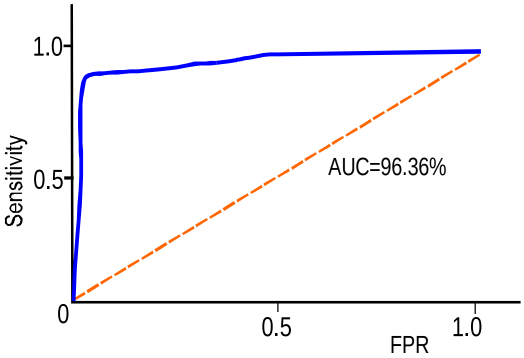


Fig. 5. ROC Curve for Still Birth Prediction.

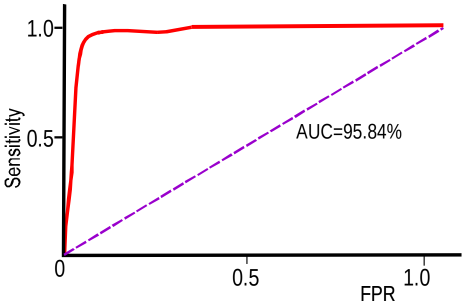


Fig. 6. ROC Curve for Term Birth Class Prediction.

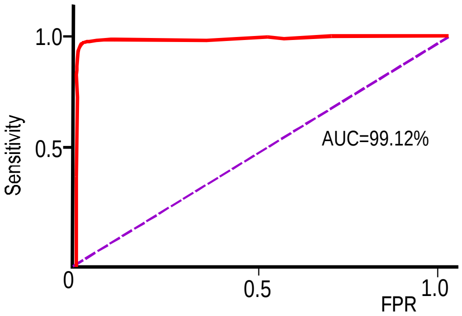


Fig. 7. ROC Curve for Preterm Birth Class Pprediction.

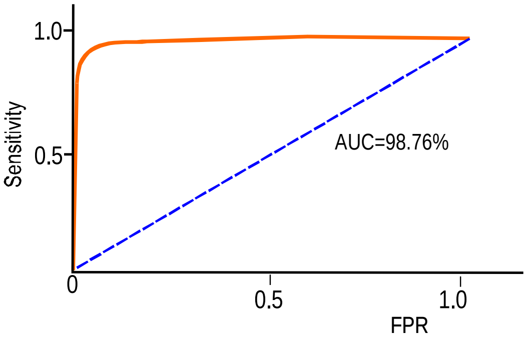


Fig. 8. ROC Curve for Miscarriage Class Prediction.

V. Conclusion

The work reported in this paper implemented a comparative predictive analytics on six machine learning algorithms (RF, DT, NB, SVM, MLP, kNN) and three data imbalance treatment approaches (RRW,RRN, SMOTE) for the prediction of POs using four test dataset modes (10-FCV, 5-FCV, Train (70%), Train (80%)). The aim was to identify the best classifier for PO classification using pregnancy risk factors. The process commenced with data collection and preprocessing — data cleaning, integration, feature selection and imbalance treatment. Feature rank analysis identified 13 principal attributes based on EV scores from PCA, which the other analytic stages depended on. SMOTE, RRN and RRW datasets drastically reduced the IR when compared to ORD dataset and were used for classification and prediction by the six ML algorithms. The experiment was conducted on four different test modes while derivatives of CM and other standard metrics were used to evaluate the performances of the different classifiers.

The results of ANOVA performed separately using F_m and RMSE showed that mean performance of classifiers across the datasets varied significantly (F=117.94; p=0.00) at 95% confidence interval, while turkey multi-comparison test revealed RF (mean=0.78) and SMOTE (mean=0.73) as having outstandingly significant means. In addition, RF model on SMOTE dataset produced ACC ≥ 0.89, AUC ≥ 0.96 and coverage of 97.8% for each PO class which depict a very good performance and was the best performing classifier. However, there was no significant difference (F=0.07, 0.01; p=1.000) in the mean performance of classifiers and datasets across test data modes respectively. The results significantly enhance the predictive accuracy of all the classes (especially adverse PO class) and confirmed the importance of data-imbalance treatment and the suitability of RF for PO classification. In terms of the adopted resampling methods, SMOTE produced the least IR among the various classes while RF and DT were the two most performing classifiers. This implies that oversampling is better than random unsdersampling methodology in the treatment of DIP maternal health domian. The results further proved that train/test data modes insignificantly affect classification accuracy in a balanced data setting, although there are noticeable variations in computational cost. The results of preprocessing identified 13 pregnancy risk factors that significantly impact on PO, therefore provide the right information for the early diagnosis and treatment of the adverse POs thereby reducing MM. The performance of these models on binary classification problems and discovery of optimal classifiers' parameters for improved performance are directions for future work.

References

[1] Goldenberg, Robert L., Elizabeth M. McClure, and Sarah Saleem. "Improving pregnancy outcomes in low-and middle-income countries." Reproductive health 15(1), (2018): 88.

[2] Romero, Roberto, Sudhansu K. Dey, and Susan J. Fisher. "Preterm labor: one syndrome, many causes." Science 345, no. 6198 (2014): 760-765.

[3] Mu, Yu, Kai Feng, Ying Yang, and Jingyuan Wang. "Applying deep learning for adverse pregnancy outcome detection with pre-pregnancy health data." In MATEC Web of Conferences, vol. 189, p. 10014. EDP Sciences, 2018.

[4] Y. Sun, A. K. Wong, and M. S. Kamel, “Classification of imbalanced data: A review," Int. J. Pattern Recognition Artificial. Intelligence., vol. 23, no. 4, pp. 687_719, 2009.

[5] Ebenuwa, Solomon H., Mhd Saeed Sharif, Mamoun Alazab, and Ameer Al-Nemrat. "Variance ranking attributes selection techniques for binary classification problem in imbalance data." IEEE Access 7 (2019): 24649-24666..

[6] B. A. G. Nguyen Hoang and S. Phung, ``Learning pattern classi_cation tasks with imbalanced data sets," in Pattern Recognition, P.-Y. Yin, Ed., Vukovar, Croatia: InTech, 2009.

[7] Yıldırım, Pınar. "Pattern classification with imbalanced and multiclass data for the prediction of albendazole adverse event outcomes." Procedia Computer Science 83 (2016): 1013-1018.

[8] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, ``An insight into classi_cation with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," Inf. Sci., vol. 250, pp. 113_141, Nov. 2013.

[9] Kim, Sangwook, Zhibin Yu, Rhee Man Kil, and Minho Lee. "Deep learning of support vector machines with class probability output networks." Neural Networks 64 (2015): 19-28.

[10] Moreira, Mario WL, Joel JPC Rodrigues, Guilherme AB Marcondes, Augusto J. Venancio Neto, Neeraj Kumar, and Isabel de la Torre Diez. "A Preterm Birth Risk Prediction System for Mobile Health Applications Based on the Support Vector Machine Algorithm." In 2018 IEEE International Conference on Communications (ICC), pp. 1-5. IEEE, 2018.

[11] Sulistiyanti, A., Farida, S., & Widodo, S. “Decision Support System To Monitoring Maternity Process Using Support Vector Machine Method”. International Journal of Research in Engineering and Science, 6(8), (2018). 45-49.

[12] Lunghi, F., G. Magenes, L. Pedrinazzi, and MARIA GABRIELLA Signorini. "Detection of fetal distress though a support vector machine based on fetal heart rate parameters." In Computers in Cardiology, 2005, pp. 247-250. IEEE, 2005..

[13] Prema, N. S., and M. P. Pushpalatha. "Machine learning approach for Preterm Birth Prediction Based on Maternal Chronic Conditions." In Emerging Research in Electronics, Computer Science and Technology, pp. 581-588. Springer, Singapore, 2019.

[14] Sahin, Hakan, and Abdulhamit Subasi. "Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques." Applied Soft Computing 33 (2015): 231-238.

[15] Gilchrist, Jeff, Colleen M. Ennett, Monique Frize, and Erika Bariciak. "Neonatal mortality prediction using real-time medical measurements." In 2011 IEEE International Symposium on Medical Measurements and Applications, pp. 65-70. IEEE, 2011.

[16] Lakshmi, B. N., T. S. Indumathi, and Nandini Ravi. "A Study on C. 5 decision tree classification algorithm for risk predictions during pregnancy." Procedia Technology 24 (2016): 1542-1549..

[17] Hill, Jacquelyn L., M. Karen Campbell, Guang Yong Zou, John RG Challis, Gregor Reid, Hiroshi Chisaka, and Alan D. Bocking. "Prediction of preterm birth in symptomatic women using decision tree modeling for biomarkers." American journal of obstetrics and gynecology 198, no. 4 (2008): 468-e1.

[18] Mehta, Rutvij, Nikita Bhatt, and Amit Ganatra. "A survey on data mining technologies for decision support system of maternal care domain." International Journal of Computers and Applications 138, no. 10 (2016): 20-4.

[19] Ferreira, Duarte, Abílio Oliveira, and Alberto Freitas. "Applying data mining techniques to improve diagnosis in neonatal jaundice." BMC medical informatics and decision making 12, no. 1 (2012): 143.

[20] Akbulut, Akhan, Egemen Ertugrul, and Varol Topcu. "Fetal health status prediction based on maternal clinical history using machine learning techniques." Computer methods and programs in biomedicine 163 (2018): 87-100.

[21] Azar, Ahmad Taher, H. Hannah Inbarani, S. Udhaya Kumar, and Hala Shawky Own. "Hybrid system based on bijective soft and neural network for Egyptian neonatal jaundice diagnosis." International Journal of Intelligent Engineering Informatics 4(1), (2016): 71-90.

[22] Moreira, Mário WL, Joel JPC Rodrigues, Neeraj Kumar, Jianwei Niu, and Arun Kumar Sangaiah. "Multilayer Perceptron Application for Diabetes Mellitus Prediction in Pregnancy Care." In International Conference on Frontier Computing, pp. 200-209. Springer, Singapore, 2017.

[23] Moreira, Mario WL, Joel JPC Rodrigues, Guilherme AB Marcondes, Augusto J. Venancio Neto, Neeraj Kumar, and Isabel de la Torre Diez. "A Preterm Birth Risk Prediction System for Mobile Health Applications Based on the Support Vector Machine Algorithm." In 2018 IEEE International Conference on Communications (ICC), pp. 1-5. IEEE, 2018.

[24] Kuppermann, Miriam, Anjali J. Kaimal, Cinthia Blat, Juan Gonzalez, Mari-Paule Thiet, Yamilee Bermingham, Anna L. Altshuler, Allison S. Bryant, Peter Bacchetti, and William A. Grobman. "Effect of a Patient-Centered Decision Support Tool on Rates of Trial of Labor After Previous Cesarean Delivery: The PROCEED Randomized Clinical Trial." Jama 323, no. 21 (2020): 2151-2159.

[25] Vinks, Alexander A., Nieko C. Punt, Frank Menke, Eric Kirkendall, Dawn Butler, Thomas J. Duggan, DonnaMaria E. Cortezzo et al. "Electronic Health Record–Embedded Decision Support Platform for Morphine Precision Dosing in Neonates." Clinical Pharmacology & Therapeutics 107, no. 1 (2020): 186-194.

[26] López-Martínez, Fernando, Edward Rolando Núñez-Valdez, Vicente García-Díaz, and Zoran Bursac. "A Case Study for a Big Data and Machine Learning Platform to Improve Medical Decision Support in Population Health Management." Algorithms 13, no. 4 (2020): 102.

[27] Løhre, Erik Torbjørn, Morten Thronæs, Cinzia Brunelli, Stein Kaasa, and Pål Klepstad. "An in-hospital clinical care pathway with integrated decision support for cancer pain management reduced pain intensity and needs for hospital stay." Supportive Care in Cancer 28, no. 2 (2020): 671-682.

[28] Pick, R. A.: Benefits of decision support systems." Handbook on Decision Support Systems 1. Springer, Berlin, Heidelberg, 719-730. (2008).

[29] Ekong, V. , Inyang, U. G., and Onibere, E. A.: Intelligent decision support system for depression diagnosis based on neuro-fuzzy-CBR hybrid." Modern Applied Science 6.7. (2012)

[30] Venkatesh KK, Strauss RA, Grotegut CA, Heine RP, Chescheir NC, Stringer JS, Stamilio DM, Menard KM, Jelovsek JE. Machine Learning and Statistical Models to Predict Postpartum Hemorrhage. Obstetrics & Gynecology. 2020 Apr 1;135(4):935-44.

[31] Tiwari P, Colborn KL, Smith DE, Xing F, Ghosh D, Rosenberg MA. Assessment of a Machine Learning Model Applied to Harmonized Electronic Health Record Data for the Prediction of Incident Atrial Fibrillation. JAMA network open. 2020 Jan 3;3(1):e1919396-.

[32] Bahado-Singh RO, Vishweswaraiah S, Aydas B, Yilmaz A, Saiyed NM, Mishra NK, Guda C, Radhakrishna U. Precision cardiovascular medicine: artificial intelligence and epigenetics for the pathogenesis and prediction of coarctation in neonates. The Journal of Maternal-Fetal & Neonatal Medicine. 2020 Feb 4:1-8

[33] García, Vicente, José Salvador Sánchez, and Ramón Alberto Mollineda. "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance." Knowledge-Based Systems 25(1) (2012): 13-21.

[34] Burnaev, Evgeny, Pavel Erofeev, and Artem Papanov. "Influence of resampling on accuracy of imbalanced classification." In Eighth International Conference on Machine Vision (ICMV 2015), vol. 9875, p. 987521. International Society for Optics and Photonics, 2015.

[35] Beckmann, Marcelo, Nelson FF Ebecken, and Beatriz SL Pires de Lima. "A KNN undersampling approach for data balancing." Journal of Intelligent Learning Systems and Applications 7(4), (2015): 104.

[36] [Fahrudin, Tora, Joko Lianto Buliali, and Chastine Fatichah. "Enhancing the Performance of SMOTE Algorithm by Using Attribute Weighting Scheme and New Selective Sampling Method for Imbalanced Data Set." Int. J. Innov. Comput. Inf. Control 15(2) (2018).

[37] Kaur, Prabhjot, and Anjana Gosain. "Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise." In ICT Based Innovations, pp. 23-30. Springer, Singapore, 2018.

[38] Q. Gu, Z. Cai, L. Zhu, and B. Huang, "Data mining on imbalanced data sets," in International Conference on Advanced Computer Theory and Engineering (ICACTE '08), 2008, pp. 1020-1024.

[39] Jeatrakul, Piyasak, and Kok Wai Wong. "Enhancing classification performance of multi-class imbalanced data using the OAA-DB algorithm." In The 2012 International Joint Conference on Neural Networks (IJCNN), pp. 1-8. IEEE, 2012.

[40] Liu, Yangguang, Yangming Zhou, Shiting Wen, and Chaogang Tang. "A strategy on selecting performance metrics for classifier evaluation." International Journal of Mobile Computing and Multimedia Communications (IJMCMC) 6, no. 4 (2014): 20-35.

[41] Inyang, Udoinyang G., and Oluwole Charles Akinyokun. "A hybrid knowledge discovery system for oil spillage risks pattern classification." Artificial intelligence Research 3(4), (2014): 77-86.

[42] Akinyokun, Oluwole Charles, and Udoinyang G. Inyang. "Experimental study of neuro-fuzzy-genetic framework for oil spillage risk management." Artif. Intell. Research 2(4), (2013): 13-36.

[43] Bin Othman, Mohd Fauzi, and Thomas Moh Shan Yau. "Comparison of different classification techniques using WEKA for breast cancer." In 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, pp. 520-523. Springer, Berlin, Heidelberg, 2007.

[44] Busa-Fekete, Róbert, Balázs Szörényi, Krzysztof Dembczynski, and Eyke Hüllermeier. "Online F-measure optimization." In Advances in Neural Information Processing Systems, pp. 595-603. 2015.

[45] Kim, Hae-Young. "Statistical notes for clinical researchers: post-hoc multiple comparisons." Restorative dentistry & endodontics 40, no. 2 (2015): 172-176.

Neuro-fuzzy System with Particle Swarm Optimization for Classification of Physical Fitness in School Children

Jose Sulla-Torres[1], Gonzalo Luna-Luza[2], Doris Ccama-Yana[3], Juan Gallegos-Valdivia[4]
Escuela de Ingeniería de Sistemas
Universidad Nacional de San Agustín de Arequipa
Arequipa, Perú

Marco Cossio-Bolaños[5]
Departamento de Ciencias de la Actividad Física
Universidad Católica del Maule
Talca, Chile

Abstract—**Physical fitness is widely known to be one of the critical elements of a healthy life. The sedentary attitude of school children is related to some health problems due to physical inactivity. The following article aims to classify the physical fitness in school children, using a database of 1813 children of both sexes, in a range that goes from six to twelve years. The physical tests were flexibility, horizontal jump, and agility that served to classify the physical fitness using neural networks and fuzzy logic. For this, the ANFIS (adaptive network fuzzy inference system) model was used, which was optimized using the Particle Swarm Optimization algorithm. The experimental tests carried out showed an RMSE error of 3.41, after performing 500 interactions of the PSO algorithm. This result is considered acceptable within the conditions of this investigation.**

Keywords—Classification; ANFIS; particle swarm optimization; physical fitness; RMSE

I. Introduction

The World Health Organization (WHO) has highlighted some important key facts about people's sedentary attitude and their relationship to some health problems due to physical inactivity. Some of these show that a lack of physical activity is among the top 10 risk factors for death and is also significantly related to other diseases [1]. Besides, a low level of physical activity is related to low levels of physical fitness. However, Physical Fitness (PF) is a reliable indicator of health in childhood years as well as in adulthood [2].

In this perspective, the evaluation of physical fitness is important. In general, physical fitness tests within the school education system are an important tool to measure the achievements of the learning standards associated with physical education Test is understood as the instrument or procedure that measures an observable response, in this case, that of physical fitness this can be measured through flexibility, jumping, agility tests, considering age, Body Mass Index (BMI), maximal oxygen consumption, Maximum Expiratory Flow (MEF), among others [3], that can be measured with the use of measuring instruments of the different types of physical fitness [4].

There are some studies on the classification of physical activity and physical fitness developed with supervised machine learning algorithms (SML) [5] and neural networks [6]. They have also been applied with Neuro-fuzzy systems that combine fuzzy logic and neural networks [7] that monitor human physical activity [8], and computational intelligence techniques to evaluate anthropometric indices [9]. In these works, a classification process is carried out using the ANFIS model, and the authors propose fuzzy sets and fuzzy rules based on anthropometric indicators and physical fitness tests for their classification.

However, to date, no studies have been found that seek to optimize these neuro-fuzzy models with some evolutionary algorithm, which allows optimizing the error when classifying physical fitness in school children. These evolutionary algorithms such as genetic algorithms, ant colony, swarms of particles would allow optimizing the results obtained in a system of classification of physical fitness.

In the present work, the question is raised that if the implementation of a hybrid neuro-fuzzy system (ANFIS) with Particle Swarm Optimization (PSO) will optimize the classification error of physical fitness in school children.

To answer the question posed, the classification system of physical activity in male and female school children between six to twelve years of age in the Arequipa-Peru region was designed; the system has as input attributes: age, weight, height and BMI, maximal oxygen consumption, MEF; and as output the classification of low PF, standard PF, and high PF. Training and classification tests are carried out with the Matlab r2018a ANFIS neuro-fuzzy model, using the *genfis3* function, with a fuzzy Sugeno-type model, an FCM clustering technique (fuzzy c-means clustering), with unsupervised learning; The FIS, created from ANFIS, is evaluated to optimize the mean square error obtained in the classification by using the PSO algorithm.

After this introduction, this article is organized as follows: The related works to this article are explained in Section II, the methodology in Section III, and the results of experimentation in Section IV, and finally the conclusions and future work in Section V.

II. Related Work

In recent years, work on classification using machine learning and neural network techniques has been used. The classification of the first part refers to the works related to physical fitness and then review works related to neuro-fuzzy, fuzzy c-means clustering, PSO, and ANFIS-PSO.

In 2020, Cai et al. [10] proposed a machine learning prediction model for successful aging (SA) based on physical fitness tests. Four machine learning models (logistic regression, deep learning, random forest, and gradient boosting decision tree) were applied to develop the prediction models, and the analyzed sample was 890. The accuracy and area under the curve of all four machine learning models was >85%

In [6], they used massive artificial neural networks to detect complicated patterns from vast amounts of input data to learn classification models. This paper compares several cutting-edge classification techniques for automatic recognition of activities between people in different settings that vary widely in the amount of information available for analysis. Neural networks performed better, achieving 60% overall prediction accuracy.

Shihabudheen and G. N. Pillai [11] mentioned that neuro-fuzzy systems are part of flexible computing (soft-computing) that encompass a set of techniques that have in common the robustness in handling imprecise and uncertain information that exists in problems related to In the real world, flexible computing techniques can be combined to take advantage of their advantages. ANFIS is a method that allows creating the rules base of a fuzzy system, using the backpropagation training algorithm from collecting data from a process. Its architecture is functionally equivalent to a Sugeno-type rule base.

Particle Swarm Optimization (PSO) is a population metaheuristic that has been successfully applied to solve optimization problems. It is inspired by the social behavior of the flight of flocks of birds or the movement of schools of fish. The PSO algorithm was developed by Kennedy and Eberhart based on a social metaphor approach [12], and it is based on the factors that influence the decision making of a particle that is part of a set of similar particles. The decision of each particle is made according to a social component and an individual component, through which the movement of this particle is determined to reach a new position in the space of solutions. Metaheuristics try to simulate this behavior to solve optimization problems.

Clustering can be defined as the process of grouping a set of abstract or physical objects into similar classes. Clustering is an unsupervised learning technique, and a suitable clustering method should identify clusters that are as compact as they are separated from each other, that is, they have high intra-cluster similarity and low inter-cluster similarity [13]. The clustering methods used in ANFIS are Subtractive Clustering and fuzzy c-means clustering (FCM). Subtractive Clustering is a fast, one-step algorithm to estimate the number of groups and centers of clusters in a data set, and it is implemented using the *subclust* function, the *genfis2* function uses this method to generate a fuzzy inference system (FIS). FCM is a grouping method developed by Dunn in 1973 and improved by Bezdek in 1981, and this method allows determining the membership of a data in a cluster, based on its degree of membership in each of the predefined clusters and the distance of the data to each of the centers of the clusters, through an optimization function; this clustering method is used by *genfis3* to generate a fuzzy inference system [14].

In [15], they applied a multiple classification support vector machine algorithm optimized by Particle Swarm Optimization to identify five types of conventional human postures. Experimental results show that our overall classification accuracy is 92.3%, and Measure F can reach 92.63%, indicating that the human activity recognition system is accurate and effective.

In 2019, Sivaram [16] proposed an Advanced Expert System Using Particle Swarm Optimization Based Adaptive Network-Based Fuzzy Inference System to Diagnose the Physical Constitution of Human Body. The comparative results with the ANFIS system and proves that BSO-ANFIS matches well with the physician's report than the ANFIS system.

In [17], they made an ANFIS training with a modified PSO algorithm. The proposed model is applied to identify a nonlinear dynamic system. ANFIS uses the least-squares method to calculate the error. The modified PSO algorithm removes the worst particle from the swarm and replaces it with two particles generated by a crossover operator from two particles, one selected at random. Moreover, the other is selected for the characteristic of being the worst local best of its generation. The modified PSO algorithm managed to improve the error of the original FIS. The idea of improving the error found in a FIS was taken from this article, using the PSO algorithm to improve the classification error for both training and test data.

In [18] they focused on the problem of recognizing physical activity, that is, the development of a system that can learn patterns from the data in order to detect what physical activity a certain user is carrying out, for this They propose a hybrid system that combines particle swarm optimization for clustering characteristics and genetic programming combined with evolutionary strategies for the evolution of a population of classifiers, in the form of decision trees. This worked significantly well for the user's specific case.

In [9] developed a valid prediction model, a modern hybrid approach was built, combining a fuzzy inference system based on adaptive networks and particle swarm optimization (ANFIS-PSO) for the prediction of changes on anthropometric indices, including waist circumference, waist-hip ratio, thigh circumference, and upper-middle arm circumference, in female athletes. The results of the ANFIS-PSO analysis were more accurate compared to SPSS. From the mentioned article, the idea of optimizing the FIS error for the classification of Physical Activity using PSO is taken, the cost function to be optimized will be the mean square error.

As described, most works vary about the use of some classifiers for physical fitness, besides, the use of ANFIS and the PSO metaheuristic. From the works studied, we saw that the vast majority of the works obtain an acceptable precision when classifying physical activity. Considering that most of these works are implemented for other types of activities that is why in this work, we intend to use the combination ANFIS PSO in the classification of physical activity in school children little discussed in the literature. The objective is to validate if these techniques are also obtaining good results with RMSE.

III. Methodology

The study was descriptive in cross-section. The sample was selected 1813 children in a probabilistic way (stratified), with 988 men and 825 women of average socioeconomic status from public schools in the urban area of the city of Arequipa-Peru (2320 meters above sea level).

Internationally standardized protocols were used because they offer a higher degree of reliability for anthropometric and physical activity variables.

The study seeks to classify Physical Activity using anthropometric data and tests of Physical Activity in school children between 6 and 12 years of age in the Arequipa region, Peru. Using a neuro-fuzzy system (ANFIS), whose root mean square error (MSE) will be optimized with the Particle Swarm Optimization (PSO) algorithm. The system will be made in Matlab R2018a.

The methodology followed for the following work is made up of the stages shown in Fig. 1.

A. Understanding the Problem

The classification of the Physical fitness will be done on a database of 1813 records, and each record has four input attributes:

- age,
- gender,
- weight,
- height,
- Body Mass Index (BMI)
- Oxygen saturation,
- Maximum Expiratory Flow (FEM)

and as an output attribute:

- Classification of Physical fitness.

Data preprocessing was first performed with anthropometric data cleaning and arrangement and the data from the following tests:

- Flexibility (cm): The flexibility of the dorsal-lumbar region, sitting, and modified reach was measured.
- Horizontal jump (cm): The horizontal jump was measured the number of times in the "kangaroo" position.
- Agility 10 x 5 m: (second) It was evaluated in a 5-meter run ten times. It was evaluated in seconds with a chronometer.

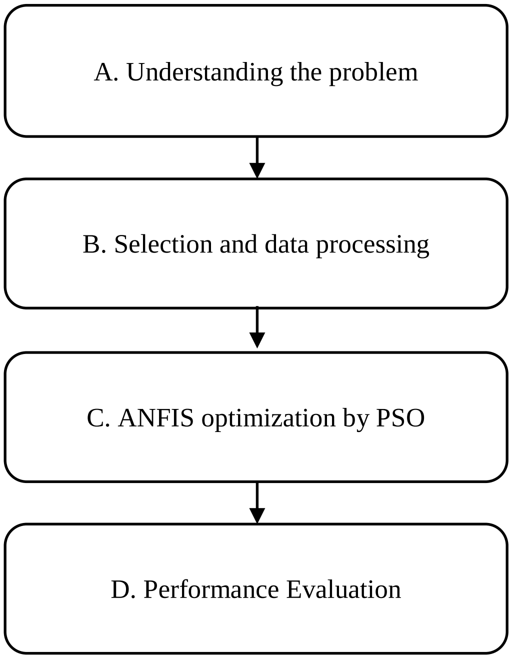

A. Understanding the problem
B. Selection and data processing
C. ANFIS optimization by PSO
D. Performance Evaluation

Fig. 1. The Proposed Methodology for the Classification of Physical Fitness.

TABLE I. Percentile Values Of The Flexibility Test Of Men And Women From 6 To 12 Years Old, From Arequipa, Peru

Flexibility (cm)														
	6		**7**		**8**		**9**		**10**		**11**		**12**	
	M	**W**	**M**	**W**	**M**	**W**	**M**	**W**	**M**	**W**	**M**	**W**	**M**	**W**
P5	34.85	35.00	15.40	18.00	14.8	15.10	14.70	15.85	16	18.00	18.1	21.2	41.6	42.1
P10	36.10	35.52	17.40	20.00	16.0	18.00	18.00	18.7.	20	21.80	20.2	23.4	41.7	42.7
P25	39.00	36.50	18.50	21.00	24.0	22.25	21.75	25.00	24	29.00	26.5	27.0	42.0	44.0
P50	41.50	39.50	22.00	29.75	36.0	38.00	32.50	40.00	36	41.00	38.0	39.5	46.0	47.0
P75	43.05	43.80	36.05	39.38	44.0	46.75	43.00	47.00	43	47.00	44.5	46.5	47.0	53.0
P90	46.85	45.00	43.10	46.48	47.4	49.95	46.65	50.00	46	49.00	49.0	50.0	48.8	54.0
P95	47.08	45.44	44.15	49.22	48.2	51.98	50.95	53.06	49	51.08	50.45	51.9	49.4	55.0

TABLE II. PERCENTILE VALUES OF THE JUMP TEST OF MEN AND WOMEN FROM 6 TO 12 YEARS OLD, FROM AREQUIPA, PERU

Horizontal Jump														
	6		**7**		**8**		**9**		**10**		**11**		**12**	
	M	**W**	**M**	**W**	**M**	**W**	**M**	**W**	**M**	**W**	**M**	**W**	**M**	**W**
P5	65.5	54.0	61.4	56.7	69.2	64.1	80.0	73.4	80.3	80.0	74.05	62.2	111	101.0
P10	70.5	56.4	65.8	63.7	70.6	70.0	89.1	77.4	90.0	85.0	86.0	70.8	112	101.0
P25	82.75	70.0	76.5	69.75	87.0	80.0	99.25	90.0	98.0	90.0	105.0	90.0	115	105.0
P50	91.0	81.0	93.0	82.0	110.0	95.0	110.0	101.5	115.0	105.0	120.0	105.0	120	111.0
P75	97.0	88.0	104.0	88.5	117.0	105.0	118.5	111.0	125.0	112.0	130.25	114.0	120	114.0
P90	106.0	97.8	113.6	100.3	125.0	112.0	129.3	114.3	140.0	115.2	140.0	121.8	120	129.4
P95	110.5	102.0	115.0	104.3	135.0	114.8	135.0	122.15	145.0	126.0	144.4	133.0	120	142.2

TABLE III. PERCENTILES VALUES OF THE AGILITY TEST OF MEN AND WOMEN FROM 6 TO 12 YEARS OLD, FROM AREQUIPA, PERU

Agility (10 x 5 m)														
	6		**7**		**8**		**9**		**10**		**11**		**12**	
	M	**W**	**M**	**W**	**M**	**W**	**M**	**W**	**M**	**W**	**M**	**W**	**M**	**W**
P5	24.03	26.07	22.32	24.32	19.43	20.12	19.50	19.51	14.71	18.46	18.22	19.16	19.5.0	21.33
P10	24.68	26.21	23.02	24.75	20.03	20.35	20.17	20.01	15.21	19.35	18.81	19.98	19.89	21.78
P25	25.83	27.78	24.4	25.24	22.06	21.81	21.77	21.57	18.43	20.74	20.10	22.12	21.04	22.40
P50	28.03	28.4	27.12	26.58	24.19	24.13	23.50	21.70	21.49	23.89	21.80	24.00	21.11	23.30
P75	28.03	31.09	31.00	28.75	25.88	27.49	24.98	25.25	23.96	24.90	23.45	24.89	21.13	23.41
P90	31.89	32.58	37.18	30.22	28.03	28.90	26.19	26.48	26.62	26.93	24.62	26.40	21.14	24.57
P95	32.69	35.586	38.208	31.24	29.722	30.98	27.007	27.939	27.966	27.88	25.997	26.767	21.15	24.96

Tables I, II, and III show the results of the Tests evaluated with their corresponding P5, P10, P125, P50, P75, P90, and P95 percentiles. The cutpoints used were based on percentiles, similar to NASPE batteries [19], considering that values that approach the percentile (P90) are rated as excellent and in the percentile (P10) as deficient, depending on the physical test or the battery objective (concerning health or physical performance). For this, the following cut-off points are suggested for the diagnosis of physical fitness: <P10 = deficient, P10 to P25 = Poor, P25 to P50 = Regular, P50 to P75 = Good, P75 to P90 = Very good and> 90 Excellent

For our work, in a practical way, we considered the Physical fitness (PF) classification: low PF (P <25), standard PF (25 <= P <75), and high PF (P> = 75).

The research work was implemented in Matlab r2018a using the *genfis3* tool, based on the PSO algorithm of [20], to which modifications were made to adapt it to the nature of the problem addressed in this article.

B. *Selection and Data Processing*

At this stage, for proper processing, the 1813 records were stored in a file (.mat), utilizing a script (.m) in charge of converting data to this format, this file (.mat) is a Matlab structure that is used to store data, which internally consists of 2 Matrices: Input Matrix and Output Matrix.

The Input Matrix consists of Age, BMI, Oxygen Saturation and the Maximum Expiratory Flow (MEF), the Output Matrix includes of the Flexibility in Physical fitness taking into account adequately the following values low PF: P <25 (value 0), standard PF: 25 <= P <75 (value 5) and high PF: P> 75 (value 10) according to the percentiles previously found in Tables I, II and III.

For balancing, the records were first randomly disordered, then 70% were chosen from these for training and 30% for tests.

This 70% was divided into: "Training inputs" and "Training outputs"; and 30% in "Test inputs" and "Test outputs."

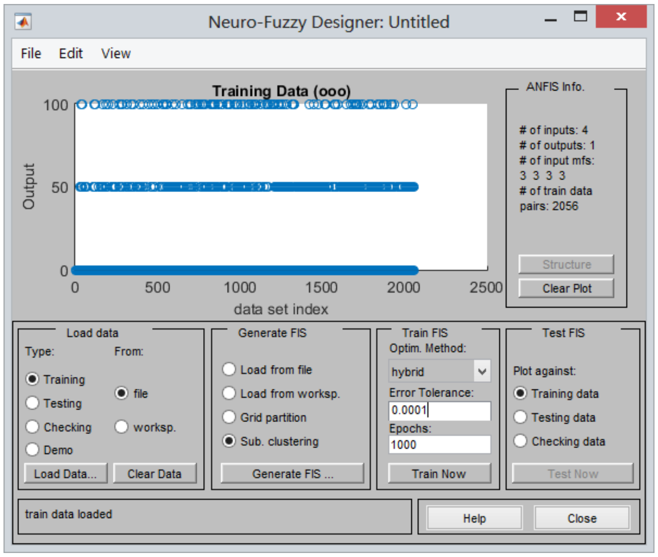
C. *ANFIS Optimization by PSO*

The ANFIS network [21] has the Matlab Neuro-fuzzy Designer application for its implementation. This is shown in Fig. 2.

It is observed that it consists of four distinct parts, such as:

- Load Data
- Generate FIS
- Train FIS
- Test FIS

Based on these steps or stages, an app optimized by PSO was built that tries to improve this process, which can be seen in Fig. 3. It is detailed in the following section.



ANFIS Info.: # of inputs: 4; # of outputs: 1; # of input mfs: 3 3 3 3; # of train data pairs: 2056

Load data — Type: ● Training ○ Testing ○ Checking ○ Demo; From: ● file ○ worksp.

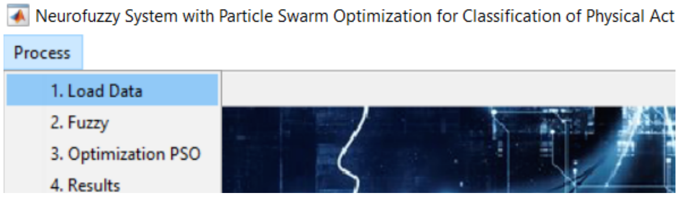
Generate FIS: ○ Load from file ○ Load from worksp. ○ Grid partition ● Sub. clustering

Train FIS — Optim. Method: hybrid; Error Tolerance: 0.0001; Epochs: 1000

Test FIS — Plot against: ● Training data ○ Testing data ○ Checking data

train data loaded

Fig. 2. ANFIS Network Implemented with Neuro-Fuzzy Designer.


Neurofuzzy System with Particle Swarm Optimization for Classification of Physical Act

Process

1. Load Data
2. Fuzzy
3. Optimization PSO
4. Results

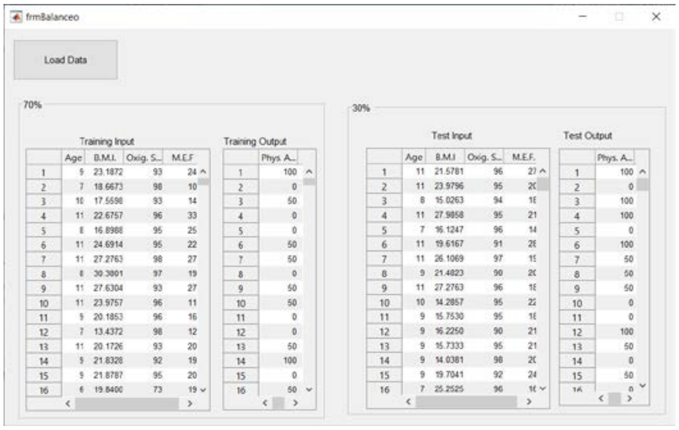
Fig. 3. Physical Activity Classification App Optimized with PSO.

1) *Load Data:* Unlike the Neuro-fuzzy Designer tool, this process is improved by incorporating data balancing features.

Fig. 4. The balancing process of the data described in section B is observed.

It should be noted that the separation between inputs and outputs is due to the fact that it facilitates the process of fusing or generating the Fuzzy Inference System (FIS) in the following section.

The results of loading and balancing the data show an amount of 1069 records for both training and testing and 744 records for training and testing outputs.


frmBalanceo

Load Data

70%

Training Input

	Age	B.M.I.	Oxig. S...	M.E.F
1	5	23.1872	93	24
2	7	18.6673	98	10
3	10	17.5598	93	14
4	11	22.6757	96	33
5	8	16.8988	95	25
6	11	24.6914	95	22
7	11	27.2763	98	27
8	8	30.3001	97	19
9	11	27.6304	93	27
10	11	23.9757	96	11
11	5	20.1853	96	16
12	7	13.4372	98	12
13	11	20.1726	93	20
14	5	21.8328	92	19
15	5	21.8787	95	20
16	6	19.8400	73	19

Training Output

	Phys. A...
1	100
2	0
3	50
4	0
5	0
6	50
7	50
8	0
9	50
10	50
11	0
12	0
13	50
14	100
15	0
16	50

30%

Test Input

	Age	B.M.I	Oxig. S...	M.E.F.
1	11	21.5781	96	27
2	11	23.9796	95	2
3	8	15.0263	94	1
4	11	27.9858	95	21
5	7	16.1247	96	14
6	11	19.6167	91	2
7	11	26.1069	97	1
8	9	21.4823	90	2
9	11	27.2763	96	1
10	10	14.2857	95	2
11	9	15.7530	95	1
12	9	16.2250	90	21
13	9	15.7333	95	21
14	9	14.0381	98	2
15	9	19.7041	92	24
16	7	25.2525	96	1

Test Output

	Phys. A...
1	100
2	0
3	100
4	100
5	0
6	100
7	50
8	50
9	50
10	0
11	0
12	100
13	50
14	0
15	50
16	0

Fig. 4. Balancing the Data.

2) *Generate FIS:* For the generation of the FIS, we chose to use fuzzy c-means clustering (FCM), since this allows us to generate a more adaptable model to the problem of the classification of flexibility in school physical fitness, as it is known that this problem has different ranges. For each specific age, that is, a 6-year-old child cannot be treated in the same way as a 12-year-old, due to the difference in the percentiles of the anthropometric indicators with age.

An alternative was to create 7 FIS models of the Sugeno-type with trapezoidal membership and linear output functions, for records grouped by age.

The way to improve this creation of 7 FIS models was proposed, for this, an unsupervised fuzzy classification was chosen that with a single FIS model that covers all the ages considered, that is, a fuzzy clusterization that automatically inferred the necessary membership functions that may apply to school children.

The solution was to use the Fuzzy c-means fusification model through Matlab's *Genfis3* function.

Genfis3 generates a fuzzy inference system (FIS) from previously provided data using fuzzy c-means (FCM) clustering, which by extracting a set of rules models the behavior of the data. The function requires separate sets of input and output data.

When there is only one output, genfis3 can be used to generate an initial FIS for ANFIS formation. The extraction method first uses the FCM function to determine the number of antecedents and consequent membership rules and functions.

Equation (1) shows the parameters taken by *genfis3*, *Xin* is the matrix of the input data, *Xout* is the matrix of the output data, *type* specify the type (Sugeno or Mamdani), *cluster_n* specifies the number of groups and grouping specifications for the FCM algorithm is detailed in *fcmoptions*.

$$fismat = genfis3(Xin, Xout, type, cluster\_n, fcmoptions) \quad (1)$$

The input membership function type is 'gaussmf'. By default, the output membership function type is 'linear.' However, if the type is specified as 'Mamdani,' then the output membership function type is 'gaussmf'.

The parameters for *genfis3* are shown in Table IV.

Fig. 5 shows the generation of the fuzzy FCM model for clusters utilizing the call to the *genfis3* function.

TABLE IV. PARAMETERS FOR GENFIS3

Parameters	**Value**
Clusters	10
Number of Rules	10
Exponent U	2
Max. Number of Iterations	100
Minimum amount of improvement	1e-5

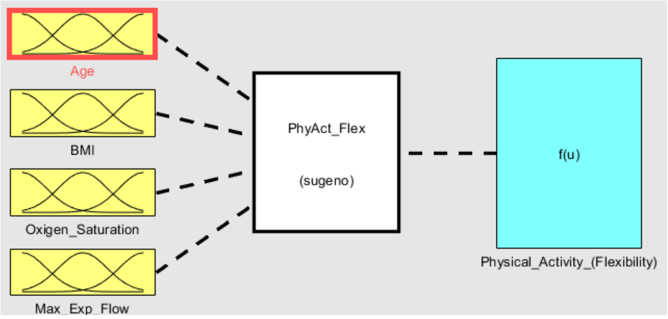


Fig. 5. Fuzzy c-means for Clusters.

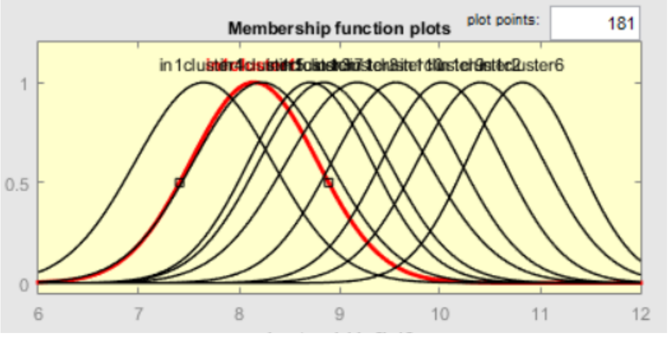


Fig. 6. Membership Function for Variables.

Fig. 6 shows the generation of the *gaussian Sugeno* membership function for the variables.

3) Train FIS: In this step, the PSO algorithm was implemented. The PSO algorithm is described as: Individuals living in society have an opinion that is part of a set of beliefs (the search space) shared by all possible individuals. Each individual can modify their own opinion based on three factors*:*

- The knowledge about the environment (its value of *fitness*).
- Historical knowledge or previous experiences (memory).
- Historical knowledge or experiences of individuals located in your neighborhood.

A PSO algorithm maintains a cluster of particles, where each particle represents a solution to the problem. Particles fly through a multidimensional search space, where the position of each particle is adjusted according to its own experience and the experience of its neighbors.

PSO is initialized with a group of random particles (solutions) and then searches for optimal ones by updating iterations. In each iteration, each particle is updated by the following two "best" values. The first of these is the best solution (fitness) that has been achieved so far and is represented as *Pbest*. The Other best value is the best solution obtained so far by any particle in the population, this is represented as *Gbest*. Each particle knows the best value to date (*Pbest*) and the best value in the group (*Gbest*). The particles try to modify their position using the current speed and distance, from *Pbest* to *Gbest*. The adjusted velocity and the adjusted position of each particle can be calculated using the formulas in equations (2) and (3).

$$V_i^{t+1} = w^t V_i^t + c_1 r_1^t (Pbest_i - X_i^t) + c_2 r_2^t (Gbest_i - X_i^t) \quad (2)$$

$$X_i^{t+1} = X_i^t + V_1^{t+1} \quad (3)$$

Where:

V_i^{t+1}: Adjusted speed

W^t: Inertia of the movement itself.

c_1: Confidence coefficient in the experience.

c_2: Confidence coefficient in the group experience.

$Pbest_i$: The best previous position of *i*

X_i^t: Current position of *i*

$Gbest_i$: Best previous position found by the group

r_1^t, r_2^t: Random operators between 0 and 1

X_i^{t+1}: Particle *i* position after adjustment.

In more detail, the PSO method is described as follows:

a) Initialize the population. The position of each of the particles is determined randomly.

b) The best previous position is matched to the current position.

c) Each position is evaluated in the fitness function to determine the quality of the solution.

d) The aptitude of the current position is compared with the best previous one.

e) Assign informants (neighborhood) of size k to the particle.

f) Determine the best particle in the neighborhood.

g) Adjust speed.

h) Adjust the position.

i) Check if the stopping criterion is met.

j) If not met, return to step e.

The values that PSO takes for an initial training optimization test are detailed in Table V. It should be noted that more tests will be carried out in the "Results" section to reach the conclusions.

TABLE V. PARAMETERS FOR PSO

Parameters	**Value**
Number of Iterations	100
Population	25
Inertia Weight	1
Damping factor	0.99
Personal training coefficient C1	1
General training coefficient C2	2

4) Test FIS: For the evaluation of performance, the mean square error (MSE) was used. The mean squared error (MSE) of an estimator measures the average of the squared errors, that is, the difference between the estimator and what is estimated. The MSE is a risk function, corresponding to the expected value of the loss of the squared error or quadratic loss. The difference occurs due to randomness or because the estimator does not take into account information that could produce a more accurate estimate.

The MSE is the second moment (about the origin) of the error and therefore incorporates both the variance of the estimator as well as its bias. For an unbiased estimator, the MSE is the variance of the estimator. Like variance, the MSE has the same units of measurement as the square of the quantity being estimated. In an analogy with the standard deviation, taking the square root of the MSE produces the error of the mean square root or the deviation of the mean square root (RMSE or RMSD), which has the same units as the estimated quantity; For an unbiased estimator, the RMSE is the square root of the variance, known as the standard deviation.

IV. Results

In Table VI. The results for the training are shown considering a population of 25 and at different levels of iterations.

Fig. 7 shows the comparison between the output (black color) and the expected results (red color) for the training with the best result obtained, which was 500 iterations.

Table VII shows the results for the tests considering a population of 25 and at different levels of iterations:

TABLE VI. Training Results

ITERATIONS				
	100	**200**	**500**	**1000**
Cost	3.54	3.35	3.10	3.47
MSE	12.53	11.25	9.63	12.06
RMSE	3.54	3.35	3.10	3.47
Medium error	-0.11	0.039	-0.02	-0.007
Error St. D.	3.54	3.35	3.10	3.47

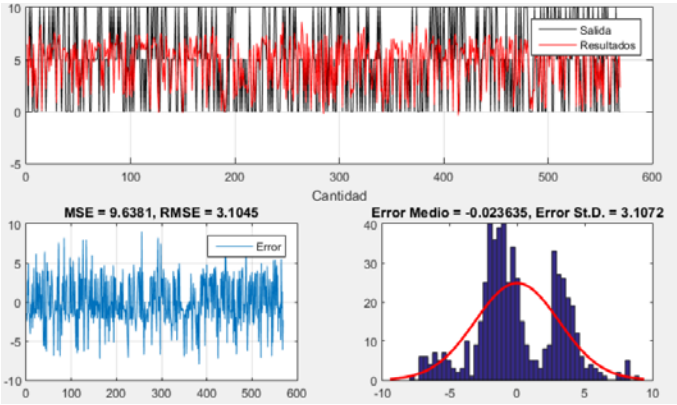


Fig. 7. Training Result after 500 Iterations.

TABLE VII. Tests Results

ITERATIONS				
	100	**200**	**500**	**1000**
Cost	3.52	3.52	3.41	3.46
MSE	12.41	12.43	11.65	11.98
RMSE	3.52	3.52	3.41	3.46
Medium Error	0.21	0.42	0.36	0.29
Error St. D.	3.5	3.51	3.40	3.45

Table VII shows that the Cost and RMSE results decrease as the iterations increase until from 500 onwards it grows again.

V. Conclusion

Based on the table of percentiles generated for the physical fitness tests, an application was created using ANFIS-PSO that processed the data for its classification

The results obtained in Tables VI and VII about the classification of the physical fitness of school children using ANFIS-PSO show that the error, as well as the cost, decreases to a greater number of interactions, more specifically when it reaches 500 iterations reaching at an RMSE of 3.10, on the other hand, if the number of iterations continues to increase, it is seen that the results decline, because it is considered that an overtraining has occurred, being adjusted to very specific characteristics of the training data that are not causally related with the objective function.

For future work, this study can be extended to improve PSO training by increasing the swarm size and parallelizing the algorithm. It is also suggested that the age range should be widened, and other physical tests increased. The implementation of the Fuzzy c-means made it more manageable to address the problem of physical fitness classification.

This approach can save physical education teachers time when trying to evaluate large populations of children.

These results can be considered as a baseline to make future comparisons and observe changes over time.

Through particle swarm optimization calculations, physical fitness can be classified by the Neuro-fuzzy System validly and reliably, since the RMSE values of 3.10 are acceptable.

Acknowledgments

To the ‘Universidad Nacional de San Agustín de Arequipa’, who has financed the project “Normative proposal for assess the levels of physical activity of school children in the province of Arequipa”, with contract number 15-2016-UNSA for the financing granted to carry out the article.

References

[1] WHO, “Physical activity,” 2018. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/physical-activity. [Accessed: 30-Mar-2020].

[2] F. B. Ortega, J. R. Ruiz, M. J. Castillo, and M. Sjöström, "Physical fitness in childhood and adolescence: A powerful marker of health," International Journal of Obesity. 2008.

[3] J. Raistenskis, A. Sidlauskiene, B. Strukcinskiene, S. Uğur Baysal, and R. Buckus, "Physical activity and physical fitness in obese, overweight, and normal-weight children," Turkish J. Med. Sci., 2016.

[4] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," Expert Syst. Appl., vol. 35, no. 1, pp. 82–89, 2008.

[5] B. Sheng, O. M. Moosman, B. Del Pozo-Cruz, J. Del Pozo-Cruz, R. M. Alfonso-Rosa, and Y. Zhang, "A comparison of different machine learning algorithms, types and placements of activity monitors for physical activity classification," Meas. J. Int. Meas. Confed., 2020.

[6] Y. Saez, A. Baldominos, and P. Isasi, "A comparison study of classifier algorithms for cross-person physical activity recognition," Sensors (Switzerland), 2017.

[7] F. Sgrò et al., "A neuro-fuzzy approach for student module of physical activity ITS," in Procedia - Social and Behavioral Sciences, 2010.

[8] M. Nii, Y. Kakiuchi, K. Takahama, K. Maenaka, K. Higuchi, and T. Yumoto, "Human activity monitoring using fuzzified neural networks," in Procedia Computer Science, 2013.

[9] M. Kazemipoor, M. Rezaeian, M. Kazemipoor, S. Hamzah, and S. K. Shandilya, "Computational Intelligence Techniques for Assessing Anthropometric Indices Changes in Female Athletes," Curr. Med. Imaging Former. Curr. Med. Imaging Rev., 2019.

[10] T. P. Cai, J. W. Long, J. Kuang, F. You, T. T. Zou, and L. Wu, "Applying machine learning methods to develop a successful aging maintenance prediction model based on physical fitness tests," Geriatr. Gerontol. Int., 2020.

[11] K. V. Shihabudheen and G. N. Pillai, "Recent advances in neuro-fuzzy system: A survey," Knowledge-Based Syst., 2018.

[12] J. Kennedy and R. Eberhart, "Particle swarm optimization," Neural Networks, 1995. Proceedings., IEEE Int. Conf., vol. 4, pp. 1942–1948 vol.4, 1995.

[13] A. Pérez-Suárez, J. F. Martínez-Trinidad, and J. A. Carrasco-Ochoa, “A review of conceptual clustering algorithms,” Artificial Intelligence Review. 2019.

[14] M. Li, K. C. Kwak, and Y. T. Kim, "Estimation of energy expenditure using a patch-type sensor module with an incremental radial basis function neural network," Sensors (Switzerland), 2016.

[15] Y. Zhu, J. Yu, F. Hu, Z. Li, and Z. Ling, "Human activity recognition via smart-belt in wireless body area networks," Int. J. Distrib. Sens. Networks, 2019.

[16] M. Sivaram, A. S. Mohammed, D. Yuvaraj, V. Porkodi, V. Manikandan, and N. Yuvaraj, "Advanced Expert System Using Particle Swarm Optimization Based Adaptive Network Based Fuzzy Inference System to Diagnose the Physical Constitution of Human Body," in Communications in Computer and Information Science, 2019, vol. 985, pp. 349–362.

[17] V. Seydi Ghomsheh, M. Aliyari Shoorehdeli, and M. Teshnehlab, "Training ANFIS structure with modified PSO algorithm," in 2007 Mediterranean Conference on Control and Automation, MED, 2007.

[18] A. Baldominos, C. Del Barrio, and Y. Saez, "Exploring the application of hybrid evolutionary computation techniques to physical activity recognition," in GECCO 2016 Companion - Proceedings of the 2016 Genetic and Evolutionary Computation Conference, 2016.

[19] S. G. Zieff, A. Lumpkin, C. Guedes, and T. Eguaoje, "NASPE sets the standard: 35 years of national leadership in sport and physical education," J. Phys. Educ. Recreat. Danc., vol. 80, no. 8, pp. 46–49, 2009.

[20] M. K. Heris, "Evolutionary ANFIS Training in MATLAB - Yarpiz," 2020. [Online]. Available: https://yarpiz.com/319/ypfz104-evolutionary-anfis-training. [Accessed: 30-Mar-2020].

[21] "A PID-like ANFIS Controller Trained by PSO Technique to Control Nonlinear MIMO Systems," Aust. J. Basic Appl. Sci., 2013.

An Efficient Classifier using Machine Learning Technique for Individual Action Identification

G.L.Sravanthi[1], M.Vasumathi Devi[2], K.Satya Sandeep[3], A.Naresh[4], A.Peda Gopi[5]
Vignan's Nirula Institute of Technology & Science for Women, Peda Palakaluru, Guntur-522009, Andhra Pradesh

***Abstract*—Human action recognition is an important branch of computer vision and is getting increasing attention from researchers. It has been applied in many areas including surveillance, healthcare, sports and computer games. This proposed work focuses on designing a human action recognition system for a human interaction dataset. Literature research is conducted to determine suitable algorithms for action recognition. In this proposed work, three machine learning models are implemented as the classifiers for human actions. An image processing method and a projection-based feature extraction algorithm are presented to generate training examples for the classifier. The action recognition task is divided into two parts: 4-class human posture recognition and 5-class human motion recognition. Classifiers are trained to classify input data into one of the posture or motion classes. Performance evaluations of the classifiers are carried out to assess validation accuracy and test accuracy for action recognition. The architecture designs for the centralized and distributed recognition systems are presented. Later these designed architectures are simulated on the sensor network to evaluate feasibility and recognition performance. Overall, the designed classifiers show a promising performance for action recognition.**

Keywords—Human action recognition; machine learning; neural networks

I. Introduction

Human action recognition has a variety of applications which require automated recognition of human behaviors. By using motion recognition, movements of a person in image frames can be detected and regenerated into a 3D model, and this can be useful for sports experts to analyze the performance of athletes [3]. Besides, it can be implemented in surveillance system which can automatically monitor human presences and behaviors at public areas like shopping mall, train station and airport to detect abnormal or suspicious activities. The capability to understand the meaning of human interactions by machines enables the development of advanced human-computer interfaces which can be used for computer games or designs. Apart from that, analysis of human activities can also be used in video conferencing, healthcare monitoring, and virtual reality.

Much research has been done in the field of human action recognition. Before 2000, researchers paid more attention to body tracking and posture recognition. W. Kay et al. [1] proposed a model-based approach by which stick figures are obtained from ordinary images to derive a model of the lower limbs. C. Gu et al. [2] introduced an approach to estimate human postures by infrared images in real-time. Space-time methods have been widely used which can model 3D volumes of human actions by combining 2D images in a space-time dimension. R. Goyal et al. [3] introduced a spatio-temporal approach which can recognize a wide range of actions by using over-segmented videos with correlation techniques.

To recognize human activities from image sequences, state-space [7, 8] and template matching [9, 10] are also applied by many researchers. Traditional approaches usually need a series of specifically designed algorithms to preprocess the images and extract feature sets [11]. Designing of such complicated algorithms can be quite time consuming. Besides, these manually designed algorithms cannot be easily reproduced for new use cases. Compared with traditional approaches, machine learning methods do not need complex preprocessing algorithms and only depend on good training datasets. Apart from that, when more patterns need to be considered, modification of the classifier can be easily applied on machine learning methods [12]. This is why traditional approaches are gradually replaced by machine learning methods.

The purpose of our project is to design human action recognition.

In order to produce dynamic visuals to interact with the human, the installation operates a visualization program which simply loads the depth images, renders them by visual lights and displays them on the screen in order [13]. This program works well and enables simple interactions between human and the installation. However, it can only conduct binary classifications on presence of human or absence of human. The natural elements are depicted in Fig. 1. To analyze human actions in the depth images, a better solution needs to be found.

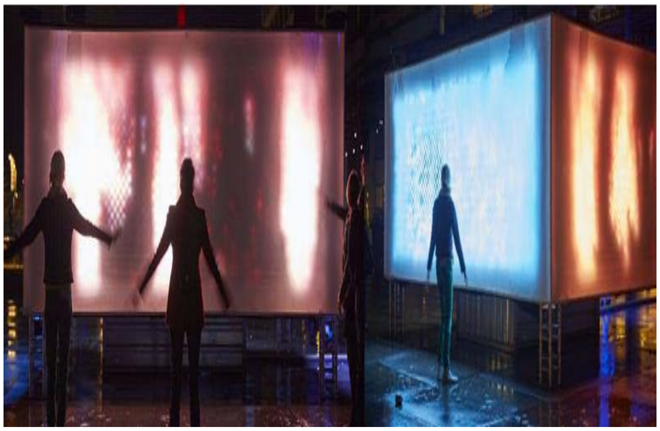


Fig 1. Natural Elements.

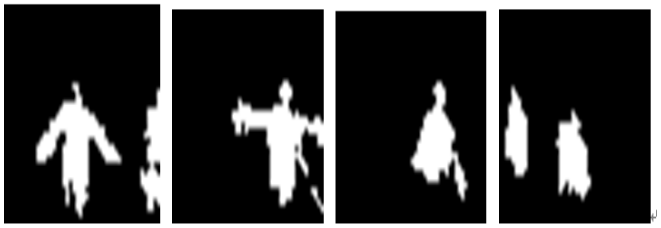


Fig 2. Depth Images.

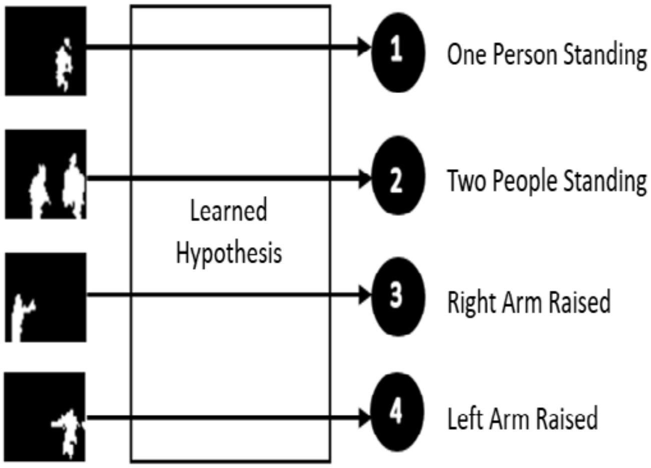


Fig 3. A Example for Mapping Input Data to Predetermined Posture Classes.

The depth images are depicted in Fig. 2 that represents natural images. As an example, we use four single images as inputs of the classifier. The classifier uses a learned "hypothesis work" for mapping inputs to predetermined classes. Fig. 3 shows an example for mapping input data to predefined posture classes.

To be specific, the classifier implements a mathematical model to read input data, transform the data, and compare the result with standard values to determine which action it belongs to. As the original data is not labeled, it cannot be directly used as input to the classifier in the learning stage. Before recognition, the original data needs to be visualized to investigate what kinds of human actions can be found. The recognizable human actions need to be classified and each action is noted by a unique label. Then the images are processed by algorithms and manually labeled to generate training examples for the classifier.

A. Machine Learning Models

Machine learning [14] has been widely used in spam detection, speech recognition, robot control, object recognition and many other domains. It is a type of artificial intelligence technology which enables computers to learn without being explicitly programmed [15]. Machine learning can be considered as the development of a system that can process a large volume of data, extract meaningful and useful information and exploit such information in practical problems [16]. Machine learning has two common learning styles: supervised learning and unsupervised learning.

Supervised learning [17] is usually used for regression, classification and ranking. In case of classification, the task of supervised learning is to establishing a relationship function from training data which consists of labeled training examples to their respective labels. Each training example contains input data and a corresponding labeled output value. The output of the relationship function is a logical value used for classification. The trained function should be able to predict the correct value of any input data. By analyzing each pair of training examples, the relationship function is produced and adjusted step by step until the underlying relationship between inputs and outputs can be appropriately expressed by the function. Common approaches for supervised learning includes logistic regression, Bayesian statistics and artificial neural networks [16].

Unsupervised learning [18] can be used for clustering, anomaly detection and dimensionality reduction. Unlike supervised learning, the task of this machine learning method is discovering the hidden structure of unlabeled training data [19]. As the training data is unlabeled, unsupervised learning does not have such a reward system to evaluate predicted output values [20]. It mainly focuses on exploring hidden patterns or intrinsic structure of training data [21]. The intrinsic structure can be used to organize these unlabeled data into similarity groups, which are also called clusters. K-means, hierarchical clustering and hidden Markov models are common cluster algorithms for unsupervised learning [22] [23].

The goal of this proposed work is to develop a human action recognition system based on the interaction dataset [24]. The algorithm used for this system should be able to recognize each human action which needs to be predefined before training. Thus, compared with unsupervised learning, supervised learning which can use labeled training set for classification tasks is the optimal solution for designing such a recognition system [25].

Fig. 4 shows the workflow of supervised learning methods. For classification tasks, the classifier is developed by three phases: training, validation and performance test. Feature extraction is the first step in the training phase. Raw data are preprocessed by feature extraction algorithms to get the feature matrix. Then the correct outputs are created for the derived feature data.

After that, feature data together with correct outputs are randomly selected and separated into three datasets: training set, validation set and test set. The training set is fed to the model iteratively to train the parameters of the model [26]. The validation set is used to determine when to stop the training by estimating the performance of the model during the training process. The test set is a set of examples that never take part in the training process, which means that is totally new data for the trained model and is finally used to evaluate the quality and the generalizability of the trained model by calculating the prediction accuracy on it. The work flow of the supervised learning is depicted in Fig. 4.

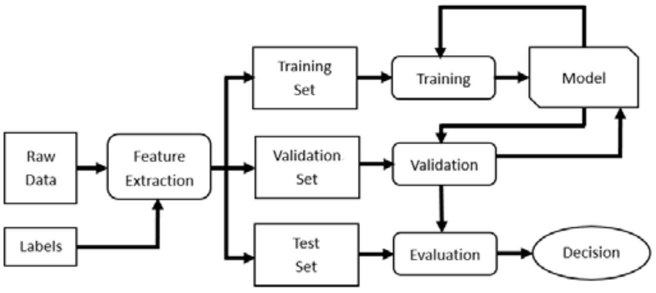


Fig 4. Workflow for Supervised Learning.

According to the literature study, some supervised learning methods are selected for our classifier. The reason why these methods are suitable for our project is discussed as follows.

Logistic Regression (LR) is one of the most commonly used method for applied statistics and discrete data analysis and often works surprisingly well as a classifier [27]. When the target outcome has only two types, it can be called a binary classification problem [28]. For binary classification, this algorithm learns from the relationship between the target outcome and a given set of predictors by estimating probabilities using the logistic function [29]. When the target outcome has more than two types, it is called a multiclass problem. Multinomial Logistic Regression (MLR) is a classification method which generalizes logistic regression to multiclass problems. We implement the MLR model as the classifier for human action recognition tasks [30].

II. Literature Survey

H. Zhao, et al. [4] present Motion History Image (MHI) to represent human motion. An MHI is a kind of temporal template which can be created by recording the motion in an image sequence. Human postures in the image sequence are accumulated in a way that records the corresponding motion history. MHI can be used in our project to extract motion features from humans. Aaron F. Bobick, et al. use a matching based method to recognize predefined motions by comparing new templates with labeled MHI instances. This matching approach performs well because of the high quality of motion examples.

Training examples for different motions are designed and created by the algorithm developers as reference [31]. However, we want to make the classification algorithms suitable for training and test datasets that are randomly collected, with unpredictable variations for each kind of motion [32]. Our algorithms will classify human actions by learning from data without the need of standard templates created by experts.

M. Monfort et al. [5] apply ANN and Bayesian framework for human action representation and classification on a multi-camera setup. Action recognition is achieved by using Multilayer Perceptron (MLP) which is a feed forward neural network model. The ability of this ANN based method to correctly classify human actions is shown by the experiments on multi-view database in which the highest recognition rate is 94.87%. The proposed method demonstrates the capacity of ANN in human action representation and recognition and shows the effectiveness of ANN model in challenging experimentations. Thus, we envision that ANN is a good candidate classifier for human action recognition.

S. Lai et al. [6] present a vision-based technique for static hand gesture recognition. Multilayer neural networks with back-propagation algorithms are used to recognize gesture features into predefined classes. The neural network based method performs well in the testing experiment and reaches sufficient recognition accuracy. What we can learn from this paper is how they analyze the problems and determine proper techniques for each step of recognition process. They divide the process into four steps: data gathering, data processing, feature extraction and classification. This type of work flow can be applied in the development of algorithms. Apart from that, this paper gives a detailed introduction of how to train the multilayer neural networks and how to design the experiments to test the classifier.

K. Tang et al. [7] propose a real-time human action recognition system based on fusion features of depth images and inertial signals. The system is trained by a public human action dataset and evaluated for real-time and offline recognition performance. The recognition accuracy is more than 97% which demonstrates the effectiveness of the system. This paper gives us a good example of data analysis in which bar charts, tables and confusion matrix are used to analyze different aspects of recognition performance.

J. Wu et al. [9] conduct investigation on different types of distributed neural networks in terms of communication cost and memory usage. They propose centralized, horizontal and vertical decomposition approaches for distributing neural networks in a Wireless Sensor Network (WSN). Compared with vertical decomposition, horizontal decomposition gets a more promising result for communication costs. This article gives a good example of how neural network can be modified to a distributed structure.

III. Proposed Method

According to the requirement of training examples, original data should be processed to fit the format of input layer of the classifier. For human posture recognition, still images are used as posture representation. For human motion recognition, a motion history image algorithm is applied to generate a motion representation from successive posture images. Pixel values of generated depth images can be directly used as features to train on machine learning models [33]. The classifier based on pixel features performs well on Intel Core i5 platform but it is not applicable on embedded platform which is limited on computational workload and memory capacity. Thus, it needs feature extraction methods to extract higher level human action features from depth images [34].

A projection-based feature extractor is presented to generate smaller feature matrix for the classifier used on embedded platforms. Machine learning models are built on TensorFlow platform. MLR is used for action recognition tasks followed by the SGD-MLR as a contrast method. A multi-layer ANN is also implemented as the classifier which uses flexible configuration of parameters to optimize recognition performance. This ANN model is then modified to a distributed structure which can be deployed on sensor networks.

A posture separation method is designed to detect posture of a single person in the image, separate it out and resize the image to a predefined size. The resized posture images are used as training examples for the human posture recognition. Pixel values are extracted as features of human postures. The feature matrix created by all feature vectors of training examples will be used as the training dataset for machine learning models. Since large matrix computation is hard to implement on embedded devices, a projection-based feature extraction method is used to extract high level features from pixel features. By using this feature extraction method, the size of the motion feature vector shrinks from 1728 to 84 and the size of posture feature shrinks from 480 to 44. Thus, this method can reduce computation workload and memory capacity of machine learning models and satisfy the requirement of embedded devices.

After data processing, four training datasets are created – two for human posture recognition and two for human motion recognition. These datasets can be easily used on machine learning models for training, validation and testing. The detail of data processing and feature extraction are explained in the following sections.

A. Data Preprocessing

At the beginning of the implementation phase, raw data should be transformed to a format which can be easily processed and readily retrieved. Data preprocessing methods are applied to solve this problem. The raw data are zipped log files which contain depth values of human interactions [31]. By using data preprocessing methods, these zipped log files are converted to binary value depth images and saved in text files. Pixel values are extracted from each frame and written in a single line of a text file [32]. Timestamps related to the frames are saved in another text file as index of frames. All empty frames are filtered and only valid frames are kept.

B. Feature Extraction

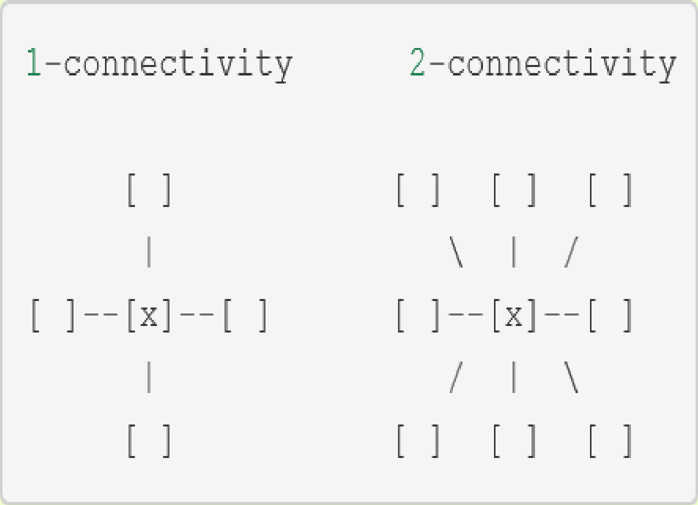
As mentioned before, features need to be extracted from labeled images. The classifier is first trained and used on Intel Core I5 platform and then transplanted to embedded platform which has limited computation capacity. Thus, feature extraction methods that require complex computations should not be applied. Since raw data has already been converted to binary value depth images, those methods designed for the RGB image are not applicable in our project. According to the literature study, learning methods can directly learn from pixel values of images. Thus, pixel values of each training example are extracted as the feature of human action patterns.

Since the size of a depth image is 36x48, the corresponding feature vector can be denoted as x[1,k] (k=1728) where k represents the number of features. When we have 1000 training examples for posture recognition, the input matrix can be denoted as X[n,k] (n=1000,k=1728) where n represents number of examples. If we use MLR as the classifier, the weight parameter used in LR is denoted by W[k,m] (k=1728,m=4) where m represents the number of output classes. The output can be denoted by hW(XW) where $h(\cdot)$ is the hypothesis work. During the training process, partial derivatives of cost function are calculated to update the weight parameter. The scale of feature matrix affects the speed of matrix operation and the larger feature matrix needs longer computation time.

In order to reduce computational workload and accelerate convergence of the objective function, we introduce a posture separation method that can detect a human body and cut it out by a bounding box. Only pixel values in the detected area are considered as effective features. This method can be achieved by image processing algorithm based on the Scikit-image library [4]. First, connected regions in the depth image are set to different colors. As can be seen in Fig. 5, connectivity refers to the maximum distance between neighbors. In our project, connectivity is set to 2 which means that the pixels with same value in 2 hops can be considered as neighbors. A connected region is a complete set of neighbors in which pixels are linked in range of connectivity. The connectivity of the pixels for behavior analysis is depicted in Fig. 5.

After coloring the connected regions, the number of pixels in each filled area is calculated and the largest region (except the background) is considered as the target region. The target region is then detected and cut out by a bounding box vector (min_row,min_col,max_row,max_col). The separated image cannot be used by learning models because they vary in size and contain different number of pixels. To solve this problem, we resize the separated image to 24x20 which is the best resolution compatible for most images. Pixel values of resized images can now be extracted as features.

Depth images only contain static information of human postures which is not sufficient for motion representation. According to [2], motion history image, a temporal template, can be used for motion feature representation. MHI not only records the presence of motion but also saves the history of a movement from the start frame to the end frame in sequence of images. MHI is created by past successive images using a weighted sum algorithm. Latest image contributes most and produces the brightest part of the MHI. The algorithm is shown below:



```
1-connectivity     2-connectivity

     [ ]           [ ]  [ ]  [ ]
      |              \  |  /
[ ]--[x]--[ ]      [ ]--[x]--[ ]
      |              /  |  \
     [ ]           [ ]  [ ]  [ ]
```

Fig 5. Connectivity of Pixels.

$$H_T(\mathrm{x},\mathrm{y},\mathrm{t}) = \frac{1}{\gamma}\begin{cases} \tau & if\ D(x,y,t) = 1 \\ \max(0, H_T(\mathrm{x},\mathrm{y},\mathrm{t}-1)-1) & otherwise \end{cases}$$

where $D(x,y,t)$ is a depth image, x and y are locations of pixels, t is the index of frame, τ is the duration, and $H\tau(x,y,t)$ is the generated MHI. Basically, MHI is a vector-image which contains direction information of a motion by combining and vanishing past images step by step. In order to create desirable MHIs which have clear borders and complete motions, key parameters of the algorithm should be determined.

In our implementation, the batch size is tested in the range of {5,10,20,50,100,200} to see how it affects the performance. The mini-batch selection procedure is shown below:

Offset=step*batch$_{\mathrm{size}}$ %(n- batch$_{\mathrm{size}}$)

batchData = trainDataset[offset: (offset+ batch$_{\mathrm{size}}$),:]

In our MLP neural networks, the neuron is a Rectified Linear Unit (ReLU) which employs the rectifier as a non-linear activation function. This rectifier function can be donated by

Relu(x)=max (0,x)

where x is the input of a neuron and the output is the maximum value between 0 and x. This means that it only uses positive values as activations and sets all negative values to 0.

For logistic regression, the cost function with L2-regularization can be denoted by

$$d(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Cost}(\mathrm{h}\,\theta\big(x(i), y(i)\big) + \frac{\beta}{2\eta}\sum_{i=1}^{m} h\,\theta_i^2 co$$

Fitting the training set too well is a problem for the classifier. When the classifier is over-fitting to the training examples, it will decrease the prediction accuracy on new examples. To solve this problem, we use regularization to prevent over-fitting and improve the generalization of the classifier. Regularization is one of the most common optimization techniques. It adds a penalty term associated with weight parameters to the cost function of hypothesis work. In this way, it makes a tradeoff between weight shrinking and minimum cost to find the model which has optimal prediction performance on all possible input examples.

IV. Results

This proposed work explains the experimental methods that were to find an optimal classifier and evaluate the results of the experiments conducted using different classifiers. Key parameters of these the classifiers are tested on Intel I5 platform to find the best configuration. To verify the modified classifiers based on architecture 2 and 3, the neural network model is redesigned in the c language and simulated on the sky notes by using Cooja. Performance evaluation is conducted for the simulated classifiers.

This section gives a brief introduction for experiments and explains the setups and measurements. The experiments were tested on 4-class posture recognition, 3-class motion recognition and 5-class motion recognition. Examples of four classes of postures and five classes of motions are shown in figure. The 3-class motion recognition is a simplified version of 5-class motion recognition. We consider "Moving Left" and "Moving Right" as the same class – "Moving", and consider "Waving Up" and "Waving Down" as the same class – "Waving". The third class of motion is "Standing".

At the beginning of the experiment, labeled examples of human actions are grouped into three parts: training set, validation set and test set. The training set is a set of examples used to tune the parameters of the classifier. The validation set is a set of examples used to estimate the performance of the model during the training process. The test set is a set of examples used to assess the performance of a fully-trained classifier. Table I illustrates the configuration of datasets for each recognition tasks. For 3-class and 5-class human motion recognition, the training/validation/test set ratio is 80:10:10 which is a commonly used settings. For 4-class human posture recognition, as there exists big variances between different examples of the same posture, to guarantee the generalizability of the classifier, the training dataset is divided with more validation and test examples.

Fig. 6 shows the result of experiment 1. We choose pixel values as feature of training examples and test the number of training step on 4-class posture recognition task. As can be seen in the figure, when the number of steps increases, the classifier achieves higher validation accuracy and test accuracy. Training accuracy is fluctuating because it is affected by the regularization term which tradeoff training accuracy to guarantee the generalizability. When the training step is larger than 1600, the test accuracy stops increasing and stays around 94.8%. Thus, for this specific task, 1600 steps is sufficient for training the classifier.

TABLE I. CONFIGURATION OF DATASETS

Recognition Tasks	Actions	No.of Examples in the Training set	No.of Examples in the Validation set	No.of Examples in the Test set
4-Class Human Posture Recognition	Left Arm Raised	200	100	150
	Right Arm Raised	200	100	150
	Single Person Standing	200	100	150
	Two people Standing	200	100	150
4-Class Human Motion Recognition	Moving	320	40	40
	Waving	320	40	40
	Standing	320	40	40
4-Class Human Motion Recognition	Moving Left	160	20	20
	Moving Right	160	20	20
	Waving up	160	20	20
	Waving down	160	20	20
	Standing	160	20	20

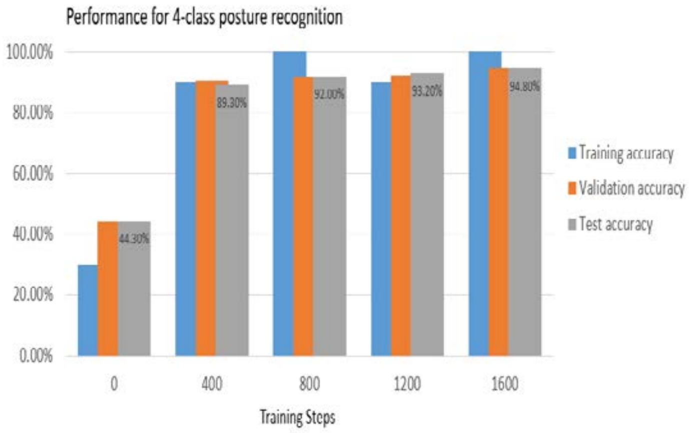


Fig 6. Impact of Training Steps on the Performance of the Classifier.

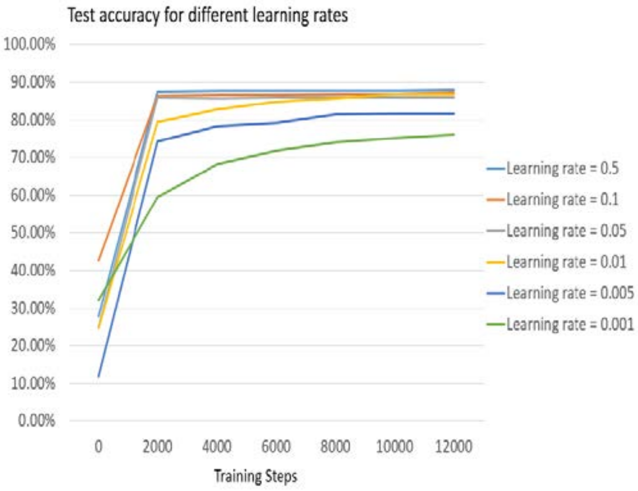


Fig 7. Impact of Learning Rate on the Test Accuracy.

When the learning rate (alpha) is set to a large value like 0.5, the model is unable to find the optimal spot so the resulting test accuracy keeps stably at 90.7%. When the alpha is set to a small value like 0.001, the training process is quite slow and needs more than 12000 steps. We find that 0.1 is a good value for learning rate that reaches the best test accuracy (91.2%). Fig. 7 represents the test accuracy levels.

Batch size is tested on 3-layer neural network for 3-class motion recognition problem using pixel features. When the batch size is set to an extremely low value like 1, the classifier is under-fitting and only gets 33.3% for test accuracy. For other batch sizes, the classifier gets similar performance about 95% ~ 97%. We find 10 is a good value for batch size that leads to the optimal test accuracy (97.5%) on this specific recognition task. Fig. 8 indicates the Impact of beta on the performance of the SGD-MLR classifier.

For a 4-layer neural network, the best configuration is for hidden nodes which leads to the optimal performance – 88% test accuracy. Fig. 9 represents the Impact of dropout on the performance of 3-layer neural networks

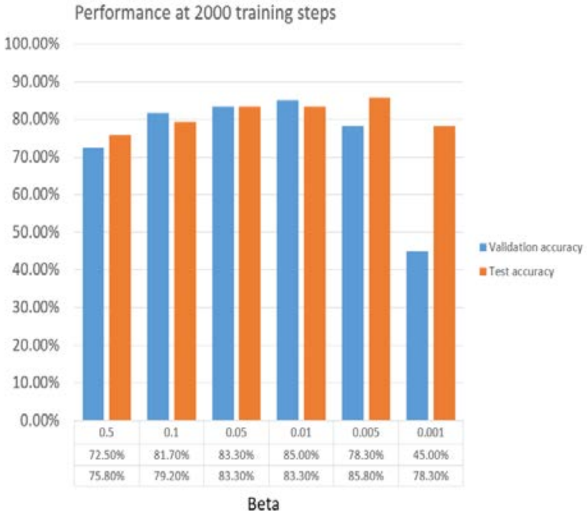


Fig 8. Impact of Beta on the Performance of the SGD-MLR Classifier.

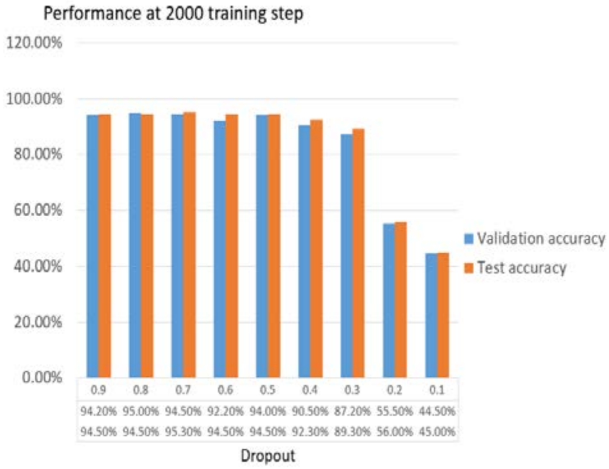


Fig 9. Impact of Dropout on the Performance of 3-Layer Neural Networks.

However, when we add one more hidden layer to the neural network, the performance does not improve a lot. In addition, tuning a 5-layer neural network is much tougher because of the various possible combinations of parameters. Thus, the 4-layer neural network is sufficient for this recognition task. The accuracy levels of the proposed method are illustrated in Table II.

TABLE II. ACCURACY LEVELS

No.of Hidden Layers	No.of Hidden Nodes	Validation Accuracy	Test Accuracy	No.of Hidden Layers	No.of Hidden Nodes	Validation Accuracy	Test Accuracy
1	5	75.0%	72.0%	1	50	89.0%	83.0%
1	20	84.0%	81.0%	1	75	90.0%	84.0%
2	10,5	80.0%	76.0%	2	50,20	87.0%	86.0%
2	20,10	79.0%	83.0%	2	75,50	84.0%	88.0%
3	20,10,5	77.0%	78.0%	3	75,20,10	85.0%	86.0%
3	50,20,10	82.0%	85.0%	3	75,50,20	87.0%	88.0%

V. Conclusion

The first task of the proposed work is to investigate human action patterns and process training examples. By using the visualization tool, the most common postures and motions are found. Examples of postures and motions are quickly selected and labeled by using the labeled tool. An image processing method is designed to improve the performance of posture recognition. The MHI is implemented to properly represent the human motion. For feature extraction, a projection based feature is presented to efficiently extract high level features from pixel values. Compared with the pixel feature, it reduces the computation workloads and the memory footprints of recognition process. Secondly, the classifiers based on machine learning models are successfully created and trained by using the TensorFLow platform. For both posture and motion recognitions, the trained classifiers show high performance in terms of validation accuracy and test accuracy. Seven experiments are conducted on Intel I5 platform and the optimal configurations of the key parameters are found based on each specific recognition tasks. Several modifications are designed for the neural network model to implement the classifier in a distributed way. We evaluate the performance of the modified neural network models, and it shows a high recognition accuracy for human posture and motion recognitions. This proposed work presents the architecture designs for centralized and distributed recognition systems and evaluate them in terms of extra-functional requirements including the performance and scalability. The designed architectures are simulated on the sensor network using the Cooja simulator and the Sky nodes. We evaluate the simulation tasks by using the confusion matrix and the result shows a quite good recognition accuracy for the distributed recognition system. However, as the Sky node has limited computation capacity, the distributed recognition system has lower throughput and longer response time compared with the centralized one. This problem can be solved in the future work by using more powerful devices and applying better communication mechanisms in the system.

References

[1]. W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev et al., “The kinetics human action video dataset,” arXiv preprint arXiv:1705.06950, 2017.

[2]. C. Gu, C. Sun, S. Vijayanarasimhan, C. Pantofaru, D. A. Ross, G. Toderici, Y. Li, S. Ricco, R. Sukthankar, C. Schmid et al., “Ava: A video dataset of spatio-temporally localized atomic visual actions,” arXiv preprint arXiv:1705.08421, 2017.

[3]. R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag et al., “The” something something” video database for learning and evaluating visual common sense,” in Proc. ICCV, 2017.

[4]. H. Zhao, Z. Yan, H. Wang, L. Torresani, and A. Torralba, “Slac: A sparsely labeled dataset for action classification and localization,” arXiv preprint arXiv:1712.09374, 2017.

[5]. M. Monfort, B. Zhou, S. A. Bargal, T. Yan, A. Andonian, K. Ramakrishnan, L. Brown, Q. Fan, D. Gutfruend, C. Vondrick et al., “Moments in time dataset: one million videos for event understanding.”

[6]. S. Lai, W.-S. Zhang, J.-F. Hu, and J. Zhang, “Global-local temporal saliency action prediction,” IEEE Transactions on Image Processing, vol. 27, no. 5, pp. 2272–2285, 2018.

[7]. K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller, “Shifting weights: Adapting object detectors from image to video,” in Advances in Neural Information Processing Systems, 2012.

[8]. S. Satkin and M. Hebert, “Modeling the temporal extent of actions,” in ECCV, 2010.

[9]. J. Wu, I. Yildirim, J. J. Lim, W. T. Freeman, and J. B. Tenenbaum, “Galileo: Perceiving physical object properties by integrating a physics engine with deep learning,” in Advances in Neural Information Processing Systems, 2015, pp. 127–135.

[10]. J. Hou, X. Wu, J. Chen, J. Luo, and Y. Jia, “Unsupervised deep learning of mid-level video representation for action recognition,” in AAAI, 2018.

[11]. Fatima, I.; Fahim, M.; Lee, Y.K.; Lee, S. A unified framework for activity recognition-based behavior analysis and action prediction in smart homes. Sensors **2013**, 13, 2682–2699.

[12]. Chen, L.; Nugent, C.D.; Wang, H. A knowledge-driven approach to activity recognition in smart homes. IEEE Trans. Knowl. Data Eng. **2012**, 24, 961–974.

[13]. Aloulou, H.; Mokhtari, M.; Tiberghien, T.; Biswas, J.; Yap, P. An adaptable and flexible framework for assistive living of cognitively impaired people. IEEE J. Biomed. Health Inform. **2014**, 18, 353–360.

[14]. Krüger, F.; Nyolt, M.; Yordanova, K.; Hein, A.; Kirste, T. Computational state space models for activity and intention recognition. A feasibility study. PLoS ONE 2014, 9, e109381, doi:10.1371/journal.pone.0109381.

[15]. K. K. Reddy and M. Shah, “Recognizing 50 human action categories of web videos,” Machine Vision and Applications Journal, 2012.

[16]. Kurakin, Z. Zhang, and Z. Liu, “A real-time system for dynamic hand gesture recognition with a depth sensor,” in EUSIPCO, 2012.

[17]. O. Kliper-Gross, T. Hassner, and L. Wolf, “The action similarity labeling challenge,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 3, 2012.

[18]. H. S. Koppula, R. Gupta, and A. Saxena, “Learning human activities and object affordances from rgb-d videos,” International Journal of Robotics Research, 2013.

[19]. G. Yu, Z. Liu, and J. Yuan, “Discriminative orderlet mining for real-time recognition of human-object interaction,” in ACCV, 2014.

[20]. Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, “Exploiting feature and class relationships in video categorization with regularized deep neural networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 2, pp. 352–364, 2018. Available: https://doi.org/10.1109/TPAMI.2017.2670560.

[21]. Arepalli, Gopi & Erukula, Suresh & Gopi, A.P. & Nagaraju, Chiluka. (2016). Secure multicast routing protocol in MANETs using efficient ECGDH algorithm. International Journal of Electrical and Computer Engineering (IJECE). 6. 1857-1865. 10.11591/ijece.v6i4.9941.

[22]. K. Sarada, V. Lakshman Narayana,(2020), “Improving Relevant Text Extraction Accuracy using Clustering Methods”, TEST Engineering and Management, Volume 83, Page Number: 15212 – 15219.

[23]. K. Sarada, V. Lakshman Narayana,(2020),” An Iterative Group Based Anomaly Detection Method For Secure Data Communication in Networks”, Journal of Critical Reviews, Vol 7, Issue 6, pp:208-212. doi: 10.31838/jcr.07.06.39.

[24]. Banavathu Mounika, P. Anusha, V. Lakshman Narayana,(2020), “ Use of BlockChain Technology In Providing Security During Data Sharing”, Journal of Critical Reviews, Vol 7, Issue 6, pp:338-343. doi: 10.31838/jcr.07.06.59.

[25]. Lakshman narayana, b. Naga Sudheer,(2020),” fuzzy base artificial neural network model for text extraction from images”, journal of critical reviews, vol 7, issue 6,pp:350-354, doi: 10.31838/jcr.07.06.61.

[26]. V. Lakshman Narayana, A. Peda Gopi,(2020),” Accurate Identification And Detection Of Outliers In Networks Using Group Random Forest Methodoly”, Journal of Critical Reviews, Vol 7, Issue 6,pp:381-384, doi: 10.31838/jcr.07.06.67.

[27]. Sandhya Pasala, V. Pavani, G. Vidya Lakshmi, V. Lakshman Narayana,(2020),” Identification Of Attackers Using Blockchain Transactions Using Cryptography Methods”, Journal of Critical Reviews, Vol 7, Issue 6,pp:368-375, doi: 10.31838/jcr.07.06.65

[28]. C.R.Bharathi, Vejendla. Lakshman Narayana , L.V. Ramesh, (2020),” Secure Data Communication Using Internet of Things”, International Journal of Scientific & Technology Research, Volume 9, Issue 04,pp:3516-3520.

[29]. Lakshman Narayana Vejendla and Bharathi C R ,(2018),“Multi-mode Routing Algorithm with Cryptographic Techniques and Reduction of Packet Drop using 2ACK scheme in MANETs”, Smart Intelligent Computing and Applications, Vo1.1, pp.649-658. DOI: 10.1007/978-981-13-1921-1_63 DOI: 10.1007/978-981-13-1921-1_63

[30]. Chaitanya, K., and S. Venkateswarlu,(2016),"Detection Of Blackhole & Greyhole Attacks In Manets Based On Acknowledgement Based Approach." Journal of Theoretical and Applied Information Technology 89.1: 228.

[31]. Lakshman Narayana Vejendla , A Peda Gopi and N.Ashok Kumar,(2018),“ Different techniques for hiding the text information using text steganography techniques: A survey”, Ingénierie des Systèmes d'Information, Vol.23, Issue.6,pp.115-125. DOI: 10.3166/ISI.23.6.115-125

[32]. A Peda Gopi and Lakshman Narayana Vejendla (2018), “Dynamic load balancing for client server assignment in distributed system using genetic algorithm”, Ingénierie des Systèmes d'Information, Vol.23, Issue.6, pp. 87-98. DOI: 10.3166/ISI.23.6.87-98

[33]. B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles, “Activitynet: A large-scale video benchmark for human activity understanding,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 961–970.

[34]. G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, “Charades-ego: A large-scale dataset of paired third and first person videos,” arXiv preprint arXiv:1804.09626, 2018.

A Survey on Detection and Prevention of Web Vulnerabilities

Muhammad Noman[1]
Department of Computer Science
Bahria University Karachi Campus

Muhammad Iqbal[2]
Department of Computer Science
Bahria University Karachi Campus
& School of Information Sciences & Technology
Southwest Jiaotong University, Chengdu, China

Amir Manzoor[3]
Department of Management Science
Bahria University Karachi Campus

Abstract—**The Internet provides a vast range of benefits to society and empowers the users in a variety of ways to use web applications. Simply, the internet has become the most transformative and fast-growing technology ever built, but it also brings new security challenges to web services in internet applications because of the scattered and open nature of the internet. A simple vulnerability in the program code could favor/benefit an attacker to obtain unauthorized access and perform adversary actions. Hence, the security of web applications from a hacking attempt is of paramount importance. This paper focuses on a literature survey recapitulating security solutions and major vulnerabilities to promote further research by systemizing the existing methods, on a bigger horizon. The data is collected from an absolute of 86 primary studies that are taken from well-known digital libraries. Different methods comprising secure programming, static, Dynamic, Hybrid analysis, and machine learning classify the data from articles. The quantity of references or the significance of a developing strategy is kept in account while selecting articles. Overall, our survey suggests that there is no way to alleviate all the web vulnerabilities therefore more studies is desirable in the area of web information security. All methods' complexity is addressed and some recommendations regarding when to use the application of given methods are provided. Finally, we typify the experience gained and examine future research openings in web application security.**

Keywords—Web security survey; web vulnerabilities; detection and prevention techniques

I. Introduction

Web-based applications are the best network-based solution to provide standard facilities. It has revolutionized the way standard facilities can be offered. Developing modern web applications is now the best mode. These applications are developed with the combination of a client and server-side development. The server-side portion uses different programming languages (.Net, PHP, Python, and Ruby) and front-end is a client-side portion, which runs on the user's web browser with different programming languages such as JavaScript and CSS/HTML. These two portions are frequently interconnected through HTTP or HTTPS protocol through asynchronous XML (AJAX) and JavaScript [1]. Fig. 1 describes the architecture of server-side and client-side of the website.

The availability of web applications has made them an integral part of everyone's daily life. This is because of their primarily free and internet-accessible availability and ability to handle sensitive data such as banking and payment for e-commerce. Because of their increased popularity, web applications are also the primary focus of hackers [2]. The popular uses of web applications, such as web blogs, social media, banking, and e-commerce, and their vulnerabilities are the focus of hackers to hack web applications with vulnerabilities. The weakness, bug, and loophole in the web application that can be exploited by hacker are called vulnerability [3].

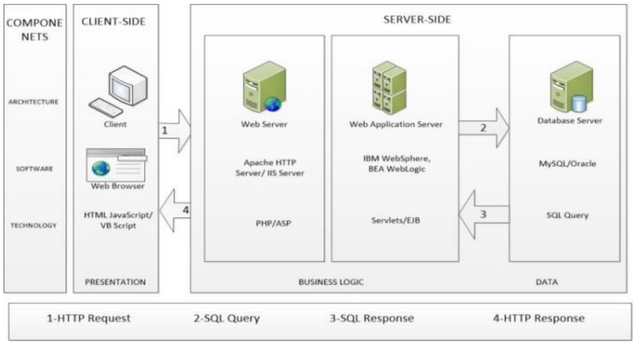


Fig. 1. Overview of Web Architecture.

The most critical vulnerabilities are cross-site scripting (XSS), SQL injections (SQLI), and cross-site request forgery (CSRF) that are listed in the top 10 web vulnerabilities by OWASP. The hackers can use the information of these vulnerabilities to compromise the website. Therefore, website requires security countermeasures to secure web application. A variety of techniques is being used around the globe to overcome these vulnerabilities and these techniques assist to identify the website vulnerabilities. There is a strong need for frequent testing to prevent and minimize web vulnerabilities. However, it requires that the tester have adequate experience given that the testing procedure itself is an extended manual process [4]. Therefore, some other approaches need to be explored to prevent vulnerabilities.

The methodology to cope with security issues is to find out the bugs before discovery and exploitation by hackers. One of the keen approaches is the use of a white-box Technique. It consists of an analysis of website source code. However, there is a problem with massive false positives and the web application's source code may not be available. There is another procedure called black-box testing to help analyzers and overcome the method of white box testing. The strategy is to examine the vulnerabilities of the application by giving some input for specific vulnerability output. Many researchers have effectively analyzed black box scanner in vulnerability detection. Furthermore, they find out its constraints by repeatedly testing numerous black-box scanners against a wide range of vulnerable applications. A lot of work in this direction is focused on fuzzing. It deals with testing (semi)-random values [5]. Another important method to prevent web vulnerabilities is data mining and machine learning. These learning methods with a variety of web applications are considered a unique approach. However, it can also be used in source code to identify vulnerabilities [6].

We tend to survey the last ten years of existing web vulnerabilities in this study. The goal is to systematize the present methods into a vast picture that supports future research. We categorized the review of web vulnerability detection methods using hybrid analysis, dynamic analysis, static analysis, data mining, and techniques of machine learning. Initially, with traditional approaches, we outline the web vulnerability discovery and analysis difficulty. We also briefly explain the web vulnerabilities and their types. Hybrid analysis, dynamic analysis, Static analysis, and machine learning are different approaches to prevent web vulnerabilities. After that, we discuss each method in detail with the definition, prevention, advantages, and challenges.

We structured the paper as follows. We initially explained the working of a web application with distinctive qualities. Section II describes the classification of web vulnerabilities along with the methods to secure it. Then, discuss and categorize each existing countermeasures in Section III. Section IV, we arrange the analysis method to detect vulnerabilities with the table and discussion. At that point, in Section V, The connected work is debated. In Section VI conclusion of this survey.

II. Background: Web Vulnerabilities Analysis and Methods

We describe the classification of web vulnerabilities and methods to secure web applications. A term vulnerability is a defect referred to error and bug that arises due to defects in the coding of a web application. This result in a severe type of damage to web application upon exploitation [4, 7]. Table I present five types of web vulnerabilities and we categorized these vulnerabilities into three main sections such as improper authentication, improper input validation, and improper session management and. It has been further divided into four web vulnerability categories: Query manipulation, Client-side, Path injection, and session management.

The main issue in security for web applications may be an inappropriate validation of user input. This Input enters into a web application via entry points ($_GET in the PHP language) and hackers can utilize web vulnerability through MySQL query. The major number of attacks occur with the combination of simple input and metadata like ‘And, OR’. Therefore these websites can frequently ensure the input of the user and validate the path and entry points [8].

A. *Improper Input Validation*

The web application is must validate or sanitize user input properly before its utilization in the web servers. Usually, web developers exercise sanitizing practices (i.e., sanitizers) for the transformation of inputs by the user into trusted data through filtration. For example, an HTML page may include JavaScript code (a PHP document may contain static HTML labels just as PHP declaration [2, 9].

TABLE I. Vulnerability Classes Split by Vulnerability Categories

VULNERABILITY Class	**VULNERABILITY CATEGORY**	**OVERVIEW**	**VULNERABILITY Name**
Improper input validation	Query Manipulation/injection	Vulnerabilities that are related to structures that are store information in the databases.	SQL Injection, NoSQL Injection, Xpath and LDAP Injection
	Client-side injection	Vulnerabilities associated with malicious code injected by a client-side such as JavaScript and processed by the server-side.	Cross-site scripting
	File and Path injection	Class of vulnerabilities that manipulate the relative path and redirect to a different location.	Remote document \local record consideration, Path/Directory Injection and Remote Code infusion
Improper session management	Session Management	Sort of malicious exploit of a site where unapproved directions are transmitted from a client that the web application trusts.	Cross-site request forgery
Improper authentication and authorization (Logic Flaw)	Logic Flaw	Vulnerabilities that can be manipulated with the coding of web application and changing them.	Unreliable Direct Object Reference, missing Functional access Control, Invalidated Redirects and Forwards or application rationale susceptibilities

1) Query Manipulation

Query manipulation is a vulnerability related to structures that store data like databases and where malicious code manipulates queries and changing them. With the help of these web vulnerabilities, a hacker easily manipulatesthe parameter of user input. As a result, the attacker becomes able to change the query's syntax. When the validation of these parameters is not proper, the maliciously infected parameters enter the reliable website due to which unsafe and unreliable information enters the web applications and damage its security. Hence, the missing or improper affirmation of controllable user data is the leading causee for injection vulnerability. There are different types of web vulnerabilities such as SQLi, LDAPI, and NoSQL. These vulnerabilities are related to the construction of filters and queries that are operated by some kind of engine example DBMS. SQL injection is considereda famous and exploited vulnerability. The other vulnerabilities are the same as SQLi, i.e. If a query involves sanitized user inputs with malicious characters then the behavior of the query performed can be altered [2, 9].

B. Client-Side Injections

Client-side injection enables malicious code to be executed by an attacker like JavaScript payloads on victim browsers without a server request. There are different vulnerabilities in this category such as XSS, remote code execution (RCE), and email injection (EI) [6].

1) File and Path Injection Vulnerability

In this class of vulnerability, a hacker manages the entrance to records from web applications or a document framework and URL areas not quite the same as the web application. These are the weakness which has a place with this gathering are RFI, LFI, and Directory traversal (DT) otherwise Path Traversals (PT) [6]. In this category, we have only considered Local file inclusion and remote file inclusion for this study.

C. Improper Authentication and Authorization (Logic Flaw)

Improper authenticating and authorizing procedures imply the invalid exercise of protocols like access control policies also known as "ACPS" as well as functions of authenticating. The logic of web application is generally executed by applying the application's control flow and saving by protecting sensitive information. One can achieve this situation or condition directly by keeping safety measures and checks to the coding of source or indirectly by the path directions provided to users like interface screening. Unsuitable implementation of business logic represents the logic errors, which force the application to behave in different ways as expected from it which results in dropping standard in "QOS" known as quality of service, losing both finance and information through the leakage. Three out of 10 top security-related hazards about applications of web [OWASP Top 10] be able to refer missing Insecure Direct Object Reference, Functional access Control, and Invalidated Readdresses and simply application logic susceptibilities [2, 9].

D. Improper Session Management

Web applications use the web session to recognize and associate multiple web entries from a single user within a specific period. A collection of web sessions is referred to as a session of a web, it may be utilized by the website for keeping the details, path of states from the past web requests and may change the further operations. In web application development, the management of the session is achieved by the cooperation of the client and the server with each other. The general tactic to do this is that an exclusive identifier (like a session ID) sent to the client by the server after the successful verification of the user. Securing alone the session ID will not be enough for managing the protected session. Session hijacking is performed by hackers through a malicious request linked to the authentic session ID. CSRF is a well-known outbreak in this category as listed in OWASP top 10 web vulnerabilities. The vulnerable web application on risk could not identify if web requests are infected or malicious until these are associated with valid session information [2, 9].

1) Session Management

Website use web session to recognize and associate multiple web entries from a single user within a specific period [2, 9]. The vulnerabilities that belong to this group are clickjacking, CSRF, Session fixation, and the hijacking of a session[10]. In CSRF, hacker submitsa malicious request as a legitimate user to web application.Clickjacking is a type of attack that invites a person to click or appeal on objects placed in infected pages and by doing this, some undesirable actions may happen without any consensus of the authentic person . Session fixation and hijacking are those attacks that aim for the user's session ID, on the other cross-Site request forgery and hand clickjacking also CSRF focus on the fact that illegal request on behalf of user [2, 11].

III. Related Work

Existing secondary studies on the topic of securing web applications are discussed (survey papers, review articles). Fig. 2 presents relevant reviews of the literature published over the past 13 years. Much work has been published to identify a taxonomy for vulnerabilities in software. Delgado et al. [12] built up a scientific classification for ordering the runtime programming flaw observing methodologies and monitoring them in light of three elements: component utilized for checking program execution and language.

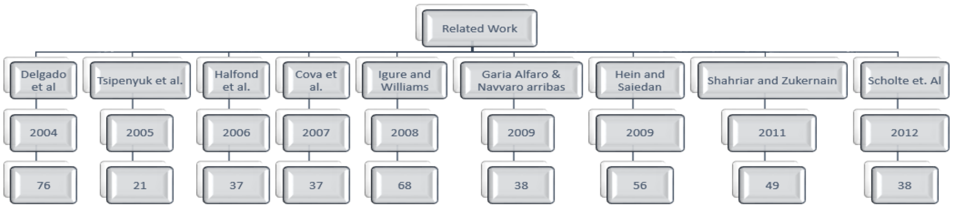


Fig. 2. Related Work.

Tsipenyuk et al. [7] arranged common flaws causing web vulnerabilities, seven out of eight categories are related to environmental and configuration issues. Many attacks and vulnerabilities are classified with various taxonomies developed and submitted in Igure and Williams' comprehensive survey[13]. Krsul [14] classifications classify vulnerabilities in software. The examinations by Halfond et al. [15], Chandrashekhar et al. [16], and Garcia-Alfaro and Navvaro-Arribas [17] give an audit on the strategies for relieving the most dangers vulnerabilities such as SQLI and XSS. The study by Cova et al. [18] features the advantages and disadvantages of weakness examination instruments accessible to secure the website. Fonseca et al. investigation [19] outline the coding flaws that should be avoided in C#, Java, and PHP. In another study, Shahriar and Zulkernine [20] gave the best in class approaches accessible for discovery and the aversion of hack attempts on applications under operation. Furthermore, they discussed the methodologies for moderating web vulnerabilities atthe program level In their study, Hydara et al. [21] discuss the methods of cross-site scripting vulnerability. The XSS systematic literature review highlights various systems for discovering and avoiding XSS attacks.

Wedman et al. [10] presented a definite survey of vulnerabilities aimed at launching session hijacking attacks and available mechanisms to protect users from such attacks for the protection of web applications from vulnerabilities, various methods are utilized and described in Li and Xue [9]. All the previously mentioned audits concentrate on any of the accompanying perspectives such as (I) building up a scientific categorization for characterizing attacks and vulnerabilities, (ii) detect the coding flaws that are abused for propelling attacks, and (iii) categorizing the flaws checking methodologies. The survey on SQLI distributed in 2012 does not take after a methodical strategy that confines the range of their investigation.

Deepa and Santhi [2] provided up-to-date approaches to web vulnerability prevention. This paper is divided into different phases of the software development life cycle with 86 primary studies. There is different web vulnerabilities research paper such as in case of XSS is thirty-five, in case of SQL injection is seventeen and in case of logical bug is thirty-five. Buczak, Anna, and Erhan [22] describe a literature survey on data mining and machine learning for intrusion detection. The latest review by Ghaffarian, Seyed, and Hamid [23] provided a detailed review of the many different methods based on machine learning that analysis and discovery of software vulnerabilities.

IV. Categorize Existing Countermeasures

Numerous researchers around the globe are working on several different ways to detect web vulnerabilities. The following sections present different technique/methods to find web vulnerabilities, such as static analysis, fuzzing and dynamic analysis, hybrid, machine learning technique, and secure programming,

The basic significance of the issue of web vulnerabilities is that many methodologies are researched and proposed. The suggested approaches are not absolute; All of them either need soundness or they are incomplete. Subsequently, all research is working to urge an enhanced approach contrasted with past works, referring to a particular part of the procedure of web vulnerability examination and revelation/discovery; like coverage of vulnerability, discovery exactness, runtime efficiency. Shahriar and Zulkernine [24] presents an extensive review to prevent web vulnerabilities reported during 1994 to 2010.

A. *Secure Programming*

Secure programming allows programmers to follow secure practices when they are developing the web application. Secure Programming protects coding practices by coding properly, checks the input data; encode correctly the user input, its type further by setting the query's parameter, also by bringing stored procedures to work. Query statements are named to those queries whose parameters are set with placeholders like "?" for referring to user data. SQL code handling placeholders in the string, which is attacking just like input. Queries that are parameterized and procedures already saved bear the same outcome however great measures are considered when programmed. Moreover, in developing of website, SQLIA's still a problem[2, 9].

To protect web apps from attackers, it is important to keep a close eye on the security features at every stage of the lifecycle while developing the web application. It is referred to as SDLC. After setting up a web application we must furnish the secondary security layer [2, 3]. Now day's operating systems are even more secure from the systems years back. The reason for this is the placement of automatic tools of safeguarding and protection within the compilers, core library alike DEP, and .NET respectively. In Linux and windows, stack or canaries cookie may also be used frequently [25]. These tools or systems stops a wide range of attacks without considering about the programmer practicing secure programming practices or not.

The writing of a safe program code has made clear on developers by the deployment of the website. Furthermore, due to the utilization of the stones library, it would be resolved in Java-based applications and Juillerat [26] that applies this technique. This library allows hackersto use databases using OOP and JavaScript payload instead of SQL payloads. The direct replacement of input data provided by the user as a string cannot be possible because it only goes via suitable procedures. Hence the programmers don't have to do much, limiting the additional work as the security features are controlled by the library. It can easily get rid of unsafe string code practice and when the number of queries framed. It can be performed by placing in the data and code a visible partition and Johns et al [27] accomplished. They achieved this by representing query syntax by the ELET (embedded language encapsulation type) introduction. To prevent attacks of XSS, Grabowski et al. [28] The created type system used in Java programming, implement directions of secure and safe programming.

A study was carried out by [29] to allow safe web development by using swift programming model language formed on the Jif language. This language confirms the integrity and confidentiality of information within the program code or declaring annotations description. The locations of the server or client can be recognized to secure placement data.

Another study proposed by Vikram et al., [30]. They give a new method Ripley, a replacement of Swift programming, to evade irregularities within the logic of business across both ends an impression of computational logic is the site on the side of the server that is present on the position of the client. Ripley confirms the reliability of RIAs and prevents from the extra work of within code annotation addition. However, information privacy cannot be guaranty and also it enforces memory, network overhead, and the reason for this is that it moves and positions from client to server every event. A language runtime used for the applications based on the PHP and python known as "Resin" permits the developers to use the already present code of application again to generate assertions that allocate the security policies. A comprehensive study conducted by Yip [31] to avoid Missing Access Control, XSS, and SQLI like multiple issues.

To develop new and secure web application an enormous frameworks of coding are created to preserve the data and important information present in web application with their reliability. To support authorizing rules or directions of web applications as acting like interpreting authority. Additionally, type checker an intermediate coding language created by Jia et al. [32] called "AURA". To allocate and verification that security policies applied properly or not by the integration of information flow and access control for web application Swamy et al. [33] enforced a system kind known as Apologue. To build a secure and safe multi-level web app a coding language is created as SELinks by incorporating language links with a fable type system and this is done by Corcoran et al. [34]. In this type, SELinks compiles the code relating to the implementation of policy to functions within the database defined by the user while fable finds the missed authorization checks. It does not guarantee the security policies relating to the state of the web application is tackle by Swamy et al. [35]. They give stateful approval approaches to the application. Krishnamurthy et al. [36] Proposed a method capluse to secure the web application with secure practices.

Another method proposed by [37] is the intelligent static examination that coordinates static investigation into the Integrated Development Environment (IDE). Additionally, provide secure programming support in-situ that helps developers stop vulnerabilities while building code. There is no need for further training and there are no hypotheses as to how programs are being developed. His work is inspired in portion from the observations that are the number of vulnerabilities introduced because many knowledgeable developers fail to practice secure programming. They have employed an interactive tool for prototype static investigation similarly as a module for Java in Eclipse. Kang and Park [38] suggested a smart fumigation system made in connection with the black box and white box test that could effectively detect/distinguish software weaknesses.

Na Meng [39] study served a wide reception of the validation and approval highlights gave by Spring Security - an outsider system intended to make sure about big business applications. They found that programming difficulties are generally identified with APIs or libraries, including the entangled cross-language information treatment of cryptography APIs. Moreover, discoveries uncover the deficiency of secure coding help and documentation, just as the gigantic hole between security hypothesis and coding rehearses.

The most recent study conducted by Bangani, S et al [40] proposes the educating of secure programming through a bit by bit approach. Our methodology incorporates the distinguishing proof of utilization hazards and secure coding rehearses as they identify with one another and to fundamental programming ideas. We explicitly mean to control instructors on the most proficient method to show secure programming in the .Net condition.The most recent study conducted by Agrawal, A et al [41] proposes an integrated and prescriptive framework intended to identify and mitigate vulnerabilities and provide suggestions for writing a more secure code.The detailed research review on a secure programming method to find XSS, SQLi, CSRF, LFI/RFI, and some other vulnerabilities presented in Table II.

TABLE II. SECURE PROGRAMMING EXISTING STUDIES

			Area of Focus					**Web Vulnerabilities**			
Research Article	***Language/Framework/Tool***	***Year***	***Vulnerability Detection***	***Vulnerability Prevention***	***Attack Detection***	***Attack Prevention***	***Vulnerability Predication***	***Query Manipulatio n/injection***	***Client side injection***	***File and Path injection***	***Session Management***
Chong, Vikram [29]	Swift	2007		✓		✓		✓	✓	✓	✓
Jia et al. [32]	Aura	2008		✓		✓					✓
Swamy et al. [33]	Fable	2008		✓		✓					✓
Vikram et al. [30]	Riply	2009		✓		✓		✓	✓	✓	✓
Corcoran et al. [34]	SELinks	2009		✓		✓					✓
Swamy et al. [35]	FINE	2010		✓		✓					✓
Krishnamurthy et al. [36]	Capsules	2010		✓		✓					✓
Zhu et al. [37]	ASIDE	2014	✓		✓			✓	✓		
Kang and Park [38]	WVF	2017	✓				✓	✓	✓	✓	✓
Na Meng et al. [39]	Empirical Study	2018	✓					✓	✓		
Bangani, S et al. [40]	Study	2019	✓					✓	✓		
Agrawal, A et al[41]	Framework	2019		✓				✓	✓	✓	✓

1) Discussion

There are numerous existing studies on preventing and detecting vulnerabilities in web applications through secure programming. Developers and firms need to focus on testing each bit of software and each application in their portfolios. By doing this right on time in the website improvement process, both can decrease, the expenses related to security. Application firewalls can be utilized as countermeasures to those attempting to hack information from an IP address. Other encryption, antivirus, antispyware, and confirmation software can be particularly utilized. To protect web applications from attackers its important keeping close eye on the security features at every stage of lifecycle while developing the application software. It is also referred as SDLC further after setting up website and it must be furnished with secondary security layer. XSS, LFI/RFI, RCE, SQLI and CSRF [2]. There are some different approached for safe coding which are distrust user input, input validation and magic switches and some tools to perform automatic source code analysis Rats, Flawfinder and ITS4. From the existing studies we conclude that Agrawal, A et al [41], Kang and Park [38], Zhu et al. [37] useful approaches in secure programming. The methods of secure programming is summarized year wise shows in the Fig. 3.

B. Analysis Method to Detect Web Vulnerabilities

There are different methods to prevent web vulnerabilities such as white box testing, blackbox, and fuzzing methods.

1) WHITE Box Testing

In the white box, the tester accesses the software code and knows the web source code's internal process. While it is possible to check how the input value of the software deduces the result value, however. This test allows access to possibly hidden source codes and code errors. The benefit of this process is that the input value can be easily predicted and a test scenario can be made. However, the white box test requires experienced skills and it is not possible to guarantee that the test specifications are met [38]. The method proposed by Jovanovic et al. [42] is suggested pixy tool.

2) Black-box Testing

The black box test depends on the software for the tester. The tester is unaware of how the software operates internally. It only tests with the result value deduction corresponding to the function-based input value. This method is advantageous for the tester as it does not require source code information or technical skills. However, while testing input values for a short time, some limitations that can not deduce logical errors and make a test case difficult without the knowledge of clear functional specifications [38].

C. Static Analysis

Mechanism of static analysis tools inspecting either binary or intermediate source code. Static examination means to look for potential vulnerabilities by inspecting the code of web applications without executing it [43]. The principal papers right now center around old vulnerabilities, for example, race conditions and buffer overflow. Later this kind of investigation has stretched out for executable programming without source code [44].

Programmers typically use static analysis tools during the development of software, checking if the code does not have vulnerabilities. In any case, these instruments just pursuit and identify the vulnerabilities. These apparatuses are program to, scanning for examples and utilizing rules for the sort of examination that they execute. As a result of this reality, the devices don't distinguish newfound classes of vulnerabilities in source code, potentially imaging the applications with bugs, creating false negatives – a weakness that exists not detailed. The false positives are additionally a worry, however in the feeling of causingwaste of time, since the software engineers need to review the code looking for non-existent bugs. Static investigation procedures arranged in two primary classes, to be specific lexical examination and semantic examination [43]. Next, these strategies are displayed, with more accentuation on taint analysis, a type of semantic examination. Lexical Analysis is a strategy to discover web vulnerabilities from source code. It's examined to scan for library capacities or framework calls that are not viewed as dependable touchy sinks.

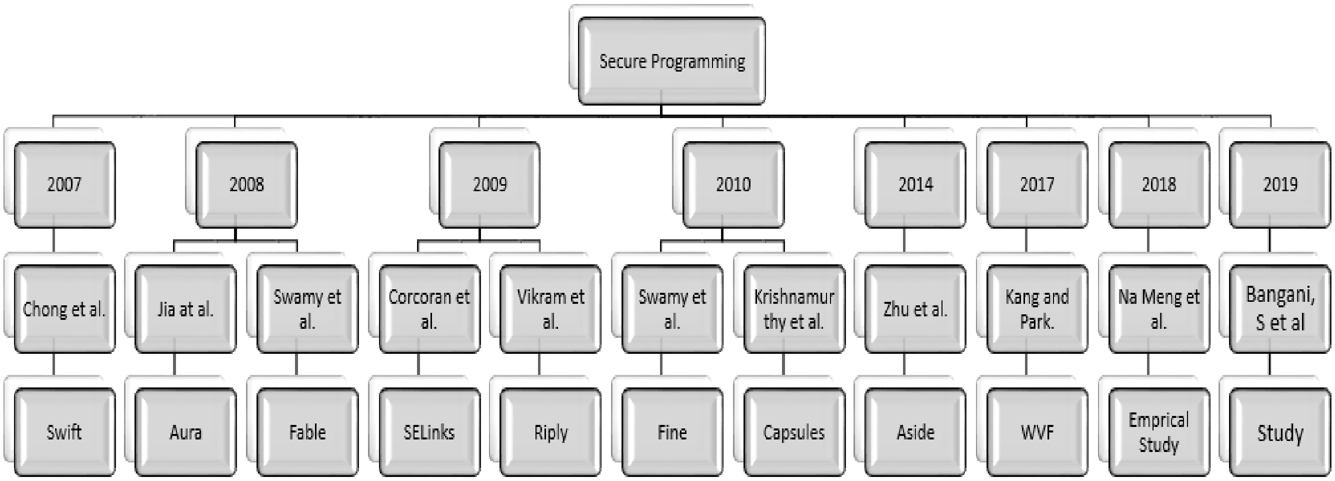


Fig. 3. Methods of Secure Programming.

This investigation includes a lot of three principal methods: control stream examination, type checking, and information stream investigation. In a study by [45], they established a static Analysis scanner WebSSARi to find vulnerabilities in web applications. This scanner provides intra-procedural and flow-sensitive reports established on the base of the lattice model. This broadens the PHP coding including two type states, known as tainted and untainted, also finds every type state of variables in it. Runtime sanitizing objects are introduced at the place the tainted data approaches the sinks. Numerous language features, like recursive functions and array elements, have not been supported, however.Using string taint analysis suitability of sanitizing procedures can be confirmed. Furthermore, Issermann and Su [46] to enhance Minamide[47] string analysis with taint support utilize this. It has analyzed the info string to perceive spoiled substring esteems to prevent any suspicious content from running by the JavaScript mediator. As it needs an understanding of the semantics of the site page the strategy can't discover DOM-based XSS.

In another study by Xie and Aiken [48] to do a reverse interpretation of fundamental blocks, methods, and the complete program to detect SQL vulnerabilities due to injection. The method they have is capable of automatically deriving the set of variables and sanitized after which function invocated by utilizing symbolic execution. Nevertheless, their static analysis technique is bound to a specific set oflanguage features. In a study done by Halfond and Orsob[49] Suggested method AMNESIA to joins static analysis and runtime monitoring to prevent SQL injection web vulnerabilities. In one more study, Shahriar and Zulkernineb[26] put forward an information-theoretic method to prevent SQL injection web vulnerabilities. The entropy of each SQL statement is calculated based on the tokens probability.

In their study, Thomas and Williams [50] SAFERPHP utilizes static analysis to find specific semantic vulnerabilities in PHP code: nullification of administration on account of vast circles, and approving tasks in databases [51]. According to the disavowal of administration, the device utilizes corrupt examination to discover circles, and afterward utilizes representative execution investigation to decide whether assailants can forestall the end of the circles. PHPSAFE (Fonseca and Vieira, [52].; Nunes et al., [63].) taints examination to scan for vulnerabilities in PHP code.Shahriar and Zulkernine [53] proposed a method to detect the vulnerabilities based on static anaylysis. Another study proposed by Shar and Tan [54] to prevent the cross-site scripting XSS method is called XSSsafer and Scholte et al. [55] proposed an IPAAS method to detect the web vulnerabilities.

Yunhui& Zhang [56] describes another use of static analysis. His approach to finding vulnerabilities in remote code execution (RCE) using the inter-procedural path and setting delicate investigation. RCE assaults require as a rule the difference in the string and non-string portions of the customer side data sources; hence, they propose an investigation that handles these parts in a composed and productive manner with the number of PHP contents and demands. They built up a calculation that comprehends these obliges in an iterative and elective style, so endeavors can be made from this arrangement. In one more study, Doup´e et al. [57] created deDacota, an automated tool that gives a clear partition between code and data in web pages. Amira et. al.,[58] proposes another static examination of web applications affirming the program's protection from meeting fixing ambushes called SAWFIX, a PHP static analyzer that outputs web applications for vulnerabilities in meeting fixation. To the best of our understanding, SAWFIX is the principle analyzer that checks extensively for this kind of powerlessness, while exchange strategies simply ensure half-precision that is limited to a modest quantity of plausible executions.

Khalid et al. [59] proposed and built up a WUM defenselessness examining apparatus (web interesting technique) to recognize and forestall every single significant weakness and tells the best way to distinguish unapproved access by recognizing vulnerabilities. The designers can discover possibly defenseless web applications with the assistance of wum Tool. WUM has created an elevated level of accuracy and similarity that is created underneath. Test's outcome shows proposed WUM helplessness scanner apparatus that gives less false positive and more vulnerabilities are identified. Another study is conctducted by Viega et al [60] on static vulnerability scanner for C and C++ code.

They developed Deepa et. al. [61] for recognizing various kinds of rationale vulnerabilities, for example, parameter control, get to control, and work process sidestep vulnerabilities. DetLogic utilizes the discovery approach and models the planned conduct of the application as a clarified limited state machine, which is thusly utilized for determining requirements identified with input parameters, get to control, and work processes. The recent study Nunes and Medeiros [62] the problem of consolidating various ASATs to improve the general identification of vulnerabilities in web applications, considering four advancement situations with various criticality objectives and limitations. These situations run from low spending plans to top of the line (e.g., business basic) web applications. The study of Long et. al, [64] has described some of the major widespread web-based vulnerabilities. These include SQLI, XSS, FI, SI etc. This study proposes an algorithm and improvements that are aimed at increasing efficiency of detecting these web-based vulnerabilities. The algorithm used to develop scanning tool use several software including UTLWebScanner. The algorithm can be compared with software providing similar functionality such as Nesus. The recent study in 2020 conducted by Aliero et al [65] to detect and minimize the occurrence of false positive and false negatives, they focus on enhancing the effectiveness of SQLIVS. They propose an object-based approach for developing SQLIVS. Three different web applications were used to test the accuracy of this approach. Each application had different types of vulnerabilities. The validity of proposed scanner was established using an experimental approach. Analytical evaluation was also used to compare the proposed scanner with other available scanners developed by various academicians. The results of experiments showed significant improvement as evidenced by high level of accuracy. The detailed research review on a secure programming method to find XSS, SQLi, CSRF, LFI/RFI, and some other vulnerabilities presented in Table III.

TABLE III. STATIC ANALYSIS OF EXISTING STUDIES

			Area of Focus					**Web Vulnerabilities**			
Research Article	***Language/Framework/Tool***	***Year***	***Vulnerability Detection***	***Vulnerability Prevention***	***Attack Detection***	***Attack Prevention***	***Vulnerability Predication***	***Query Manipulation/ injection***	***Client side injection***	***File and Path injection***	***Session Management***
Huang et al. [45]	WebSSARI	2004	✓					✓	✓		
Minamide [47]	analysis	2005	✓						✓		
Halfond and Orso [49]	AMNESIA	2005			✓	✓		✓			
Xie and Aiken [48]	method	2006	✓					✓			
Jovanovic et al. [42]	Pixy	2006	✓					✓		✓	
Wassermann and Su [46]	analysis	2007	✓					✓			
Thomas and Williams [50]	model	2007		✓				✓			
Shahriar and Zulkernine [54]	method	2009				✓		✓			
Son and Shmatikov [51]	SAFERPHP	2011	✓					✓	✓		
Shar and Tan [54]	XSSafer	2012	✓				✓		✓		
Shahriar and Zulkernine [24]	Program	2012	✓					✓			
Scholte et al. [55]	IPAAS	2012		✓				✓	✓		
Yunhui & Zhang [56]	script	2013					✓		✓		
Doup´e et al. [57]	deDacota	2013		✓							
Fonseca & Vieira [52]	tool	2014		✓				✓	✓	✓	✓
NUNES et al. [63]	phpSAFE	2015					✓	✓	✓		
Khalid et al. [59]	WUM	2017	✓					✓	✓	✓	
Amira et al. [58]	SAWFIX	2017					✓				✓
Deepa et al [61]	DetLogic	2018	✓					✓	✓		
Nunes and Medeiros ss[62]	ASAT Study	2019	✓					✓	✓	✓	✓
Long, et al. [64]	UTLWebScanner	2020	✓					✓	✓	✓	
Aliero et al [65]	Scanner	2020		✓			✓	✓			

1) Discussion

Static analysis tools, either source, binary, or intermediate, mechanize code inspection. The objective of the static examination is to look for vulnerabilities in the source code without running it [43]. Because in the development process, static application security testing tools are used early. Before software is deployed, they can identify vulnerabilities. These tools test line by line the source code, prevent flaws and provide the opportunity to fix them before becoming a true vulnerability on the web. It requires access to source code or binaries that certain organizations or individuals may not want to abandon application testers. In order to detect vulnerabilities before deployment into the live environment, it usually needs to be integrated into the system development lifecycle, which can make implementation difficult.

Each SAST tool tends to focus on a subset of possible weaknesses. The advantages are the ability to detect vulnerabilities that are not visible without access to the source code.The capacity to reveal to you the specific area of any source code shortcomings, including the line number. Probably the greatest test to choose the correct instrument when utilizing SAST is the number of false positives produced. From the current, valuable methodologies in the static examination are NUNES et al [63] Khalid et al. [59] and Nunes, P et al. [62]. The methods of secure programming are summarized year-wise as shown in the Fig. 4.

D. Fuzzing and Dynamic Analysis

Fuzzing and dynamic analysis is another method to identify web vulnerabilities In this method does not analyze web application code for vulnerability detection from static analysis but verifies in runtime whether injected data triggers some vulnerability in application specifications [38]. In this way, it is viewed as a testing procedure that finds bugs in programming by taking care of a program with s unexpected inputs (Evron&Rathaus, [66]; Sutton et al., [67]. In their study, NguyenTuong et al. [68] altered PHP transcriber to exactly infected data of the user on the character's granularity and it traces tainted user data at runtime. In another examination, Haldar et al. [69] formulated the arrangement of Java bytecode that can grow the Java framework with inadequate following assistance. These systems will in general be easy to apply in light of the fact that it doesn't require information about the program to test. Its cooperation with the program is constrained to the program's entrance focuses Jimenez et al. [70]. Mill operator et al. [71] that depicted how they took care of UNIX program utilities with irregular data sources, such as SPIKE [72], improve this thought by giving to the applications distorted sources of info, utilizing a conventional information structure to speak to various information types [67].

In an examination done by Huang et al. [73] utilized a system named as WAVES distinguish both SQLI and XSS vulnerabilities in web applications. Another utilization of Dynamic investigation Huang et al. [45] created white-box instrument b same group called WebSSARI, though WAVES.

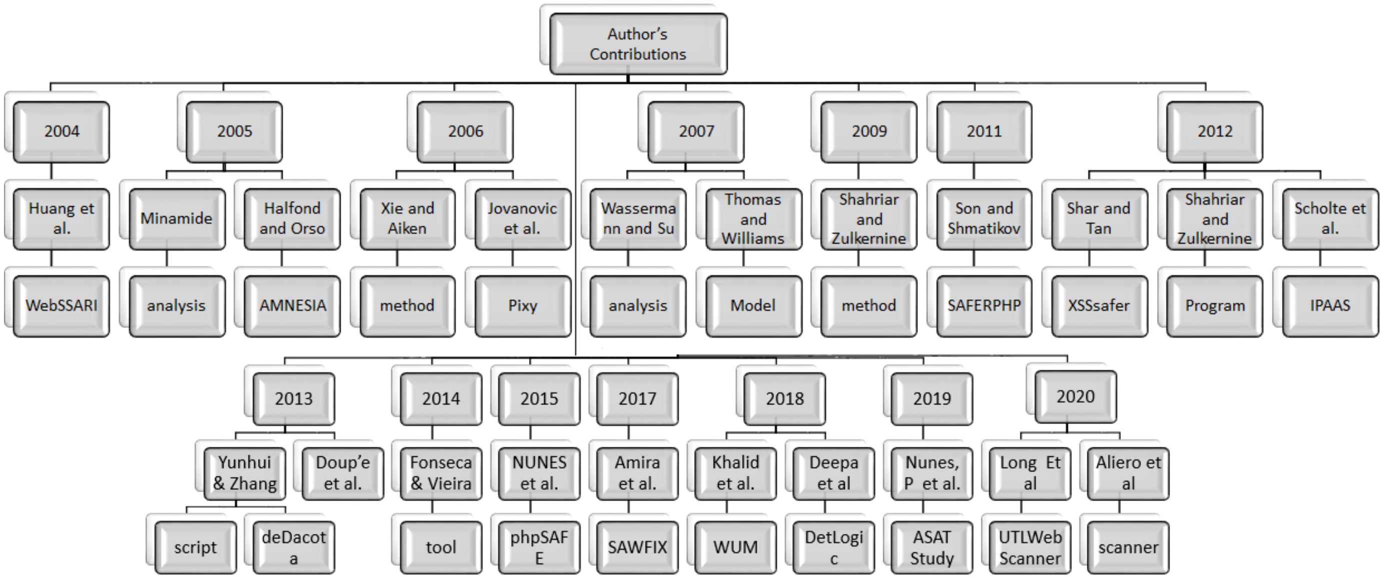


Fig. 4. Methods of Static Analysis.

perceive vulnerabilities with the assistance of a discovery approach In 2006, Kals et al. [74] built up a discovery defenselessness scanner, Secubat, to perceive vulnerabilities. The tool utilizes a crawler to perceive the site pages of the application, possess the structure fields on pages with assault vectors, and afterward break down vulnerabilities to distinguish them.

It is possible to classify fuzzers into two categories: blackbox and whitebox [67]. A blackbox fuzzer executes the method portrayed as yet. As the blackbox approach is generally free of the application and doesn't require setting up the application. Regardless of black-box fuzzers being helpful, they will, in general, find just shallow (bugs that are anything but difficult to discover) and as a rule have low code inclusion (don't practice every conceivable incentive for a given variable), missing numerous pertinent code ways and in this way numerous bugs. KameleonFuzz is a blackbox fuzzer that scans for cross-site scripting. It creates malignant contributions to control cross site sciprting XSS [75].

Another technique of dynamic analysis proposed by Antunes et al., [76] is related to the black box fuzzing in a certain way that is attack injection. A tool that implements this technique intends to mimic the behavior of an attacker, and continuously inject malformed inputs while monitoring the application. The procedure is rehashed to assemble all conceivable execution ways and checking a few properties in runtime [66, 67]. This is a form of white-box fuzzing that is actualized in the SAGE utilizing emblematic execution to practice all conceivable execution ways of the program. Since representative execution is moderate, in any case, it doesn't reach out to huge projects, it is difficult to find profound and complex bugs [84].

In a different study, Ciampa et al. [77] chose the result of the different advance tests on pattern matching the valid and error messages. Data stored in the form of tables and fields are tested by an empirical tactic to evaluate the gathered information. After all the computation, the compiled data is utilized to check attack inputs that are useful to recognize the vulnerabilities. In another study by Lekies et al. [78] used a taint-aware JavaScript engine to sense DOM-based XSS. Whereas, it is out of bound the other available methods to perceive these vulnerabilities. Whitebox fuzzers apply symbolic execution and imperative tackling to the source code Duchène et al., [75]. The working rule of some white-box fuzzers is to produce and to perform dynamic representative execution in an occurrence. It accumulates information stream ways and requirements on contributions from contingent branches that are experienced along with the execution. Then, the collected constraints are negated (constraint solving) and new inputs are injected to collect new execution paths. To deal with this difficulty, Dowser is a combination of symbolic execution with dynamic taint analysis to identify vulnerabilities in buffer overflow buried deep within Haller et al. [79] Programs implemented in their study.

In one, more study Duchene et al. [5] In order to get the auto production of unwanted inputs to access XSS vulnerabilities, the author used a genetic algorithm. Whereas, most of the available techniques do not have such an ability to reach the cause of DOM-based XSS vulnerabilities. In their study, Dohse, and Holz [80] purposed a very first known automatic testing method that uses static code to notice second-order vulnerabilities and correlate more than one step attacks in web applications. The flow of unattended data can be detected by checking the incomings and outgoings from the webserver. It has been a successful identification of unsensitized data streams by linking input and output points of data in databases. Another dynamic Analysis study by Weissbacher.et al., [81] gave a system to strengthen the JavaScript-based web application to protect them from the used side attacks named ZigZag. It is a tool of client-side code. It produces a model that tells how and with whom the client-side section is in the network. It is efficient enough to perform dynamic security code invariant detection by the respective models as well as it has the ability to handling templated JavaScript bypassing overall re instrumentation in cases where the JavaScript programs are structurally identical.

RanWang et al. [82] propose a unique recognition structure (TT-XSS) for DOM-XSS by methods for the pollute following at the customer side. They modify all JavaScript highlights and DOM APIs to corrupt the rendering procedure of programs and vectors are inferred to check the vulnerabilities naturally. In the recent study Park, et al. [83], a vulnerability detection technique is proposed that develops and manages safe applications and can resolve and analyze these problems. They developed a prototype analysis tool using our technique to test the application's vulnerability–detection ability, and show our proposed technique is superior to existing ones. The recent study in 2020 conducted by Falana [85] used Dynamic Analysis and Fuzzy Inference. The combination of these two techniques allowed them to come up with a hybrid mechanism that can be used for detection of XSS attacks. This approach used scans of website for possible SQL injections. Once this scan was done, they launched an attack vector using a HTTP request. The approach was used to test some active web applications. The results showed a large number of vulnerabilities were detected successfully. the detailed research review on dynamic analysis to prevent web vulnerabilities shown in Table IV.

1) Discussion

Dynamic analysis is a useful technique to prevent web vulnerabilities and does not analyze the source code of the website but verifies in runtime whether injected data triggers some vulnerability in the application. With this strategy, DAST tools offer risk examination and aids in the remediation endeavors, engineers do not generally know where precisely the vulnerabilities are found, nor do they generally know what countermeasures to execute. DAST approach detailing is not as much as agreeable in various examples. From the existing study RanWang et al.[82], Park et al. [83] are useful approaches in dynamic Analysis. The methods of dynamic analysis are summarized year-wise shows in Fig. 5.

TABLE IV. DYNAMIC AND FUZZING ANALYSIS EXISTING STUDIES

			Area of Focus					**Web Vulnerabilities**			
Research Article	***Language/Framework/Tool***	***Year***	***Vulnerability Detection***	***Vulnerability Prevention***	***Attack Detection***	***Attack Prevention***	***Vulnerability Predication***	***Query Manipulation/injection***	***Client side injection***	***File and Path injection***	***Session Management***
Huang et al. [73]	WebSSARI	2005	✓					✓	✓		
Haldar et al.[69]	For java	2005	✓					✓	✓		
Kals et al. [74]	Secubat	2006	✓					✓	✓	✓	✓
Antunes et al. [84]	Tool	2010	✓					✓	✓		
Ciampa et al. [77]	Tool	2010			✓			✓	✓	✓	
Lekies et al. [78]	Approched	2013	✓						✓		
Duchene et al. [5]	black-box fuzzer	2014	✓						✓		
Dahse and Holz [80]	Tool	2014			✓				✓		
Weissbacher et al. [81]	ZigZag	2014		✓					✓		
RanWang et al.[82]	TT-XSS	2018	✓					✓	✓		
Park et al. [83]	TOOL	2019	✓				✓				
Falana et al.[85]	approach	2020		✓				✓			

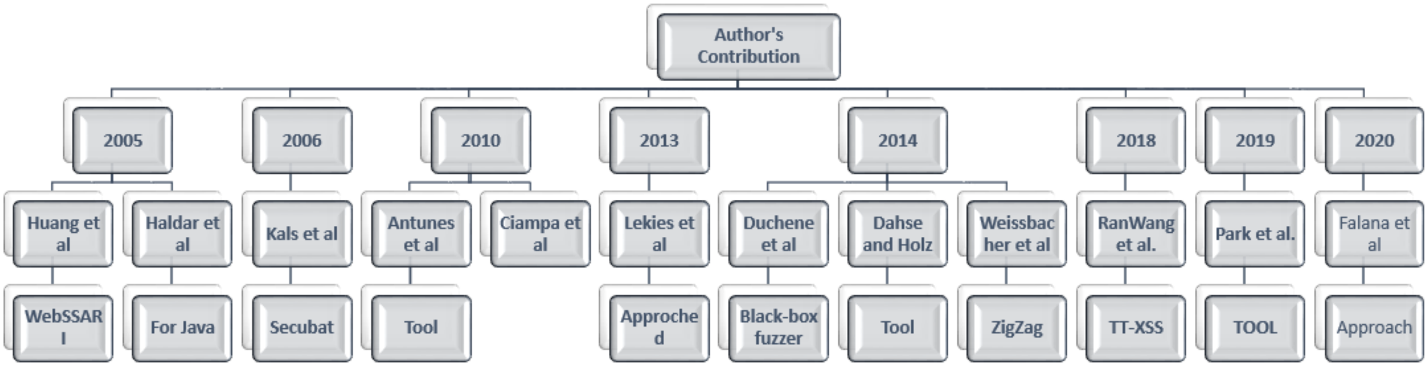


Fig. 5. Methods of Dynamic Analysis.

E. Hybrid Method (Static + Analysis)

Extracts of static and dynamic analysis are mixed to be named as hybrid analysis and provide a path toward precision analysis. In a study by Di Lucca et al. [86] identified vulnerable web pages by studying the application's source code by arrangement control flow graphs. XXS attacks are chiefly because of wrong input sanitization functions. Some web applications are also not successful to separate the suspected entries in the inputs. In another study by Balzarotti et al. [87] claimed various elusive defects could be introduced in the web application due to defective sanitization. These subtle flaws cannot be detected by the static and dynamic practices. The hybrid analysis is utilized by Saner to identify the validity of built-in and customized sanitization procedures. Saner is the first to implement the conventional static string investigation to model the working of user input sanitization. Saner first applies conventional static string analysis to model the sanitization of user input. In order to mark frail or wrong sanitization, a big series of malicious inputs are introduced into the test sanitization procedure.

Livshits et al [88] and Lam et al., [89] studied model checking, static and dynamic inspection, and runtime detection to purpose a holistic method. Which enhances the precision of static analysis by specifically using model checking. Model-checking can analytically search the space of a limited state system. It confirmed the authenticity of the system in reference to the provided conditions or characteristics. As well as this method of checking is capable to automatically produce tangible attacks, produce no false positives in vectors, and exploit path. Another technique of hybrid analysis study by Van Acker et al. [90] found the XSS vulnerabilities in flash applications by setting up Flashover. Whereas, the previous works up until focused on the discovery of conventional XSS web application vulnerabilities, Flashover identifies vulnerabilities in RIAs (Rich Internet Applications).

Another research is led in 2012 Lee et al. [91] struggles for finding SQLIAs by adopting both the static and dynamic methods. He examined the source code then the model of query is deduced from it. It removes the characters involved in SQL queries. After identifying and removing miscellaneous values, the obtained syntax is stored. The syntactic structure of quires is analyzed and compared with the already saved structure, this is how attacks are perceived in the runtime. The pros of using this scheme are that it can identify attach during the process. Another research Lee et[92], also applied both static and dynamic analysis methods for vulnerabilities of web applications. Along with the combination of these other techniques that are also being utilized for the specific application, dynamic black-box testing based on a fuzzing method is included in it. Vogt et al. [93], and Stock et al. [94] deploy the method to prevent the client-side browser from scripting XSS cross-site.

They propose He, X et al [95] a crossover examination strategy consolidating static and dynamic investigation for recognizing noxious JavaScript code that works by first directing grammar examination and dynamic instrumentation to remove inward highlights that are identified with malignant code and afterward performing characterization based identification to recognize assaults. In particular, MJDetector can distinguish JavaScipt assaults in current website pages with high precision 94.76% and de-jumble muddle code of explicit sorts with exactness 100% though the gauge strategy can just identify with exactness 81.16% and has no limit of de-obscurity. The recent study proposes Le et al. [96] E-THAPS which actualizes a novel discovery component, an improved SQL infusion, Cross-site Scripting, and helplessness identification capacities. For vindictive web shell identification, pollute examination, and example coordinating techniques are picked to be actualized in GuruWS. the detailed research review on hybrid analysis to find XSS, SQLi, CSRF, LFI/RFI, and some other vulnerabilities shown in Table V.

TABLE V. HYBRID (STATIC AND DYNAMIC) ANALYSIS EXISTING STUDIES

			Area of Focus					**Web Vulnerabilities**			
Research Article	***Language/Framework/Tool***	***Year***	***Vulnerability Detection***	***Vulnerability Prevention***	***Attack Detection***	***Attack Prevention***	***Vulnerability Predication***	***Query Manipulation/injection***	***Client-side injection***	***File and Path injection***	***Session Management***
Di Lucca et al. [86]	WA	2004	✓						✓		
Livshits et al. [88]	TOOL	2005	✓					✓	✓		
Vogt et al. [93]	novel	2007		✓		✓			✓		
Balzarotti et al. [87]	Saner	2008	✓					✓	✓		
Lam et al. [89]	*	2008	✓					✓	✓	✓	
Van Acker et al. [90]	Flashover	2012	✓						✓		
Lee et al. [91]	*	2012			✓			✓			
Stock et al. [94]	*	2014				✓			✓		
He, X et al [95]	MJDetector	2018	✓					✓	✓		
Le et al. [96]	GuruWS	2019		✓				✓	✓		

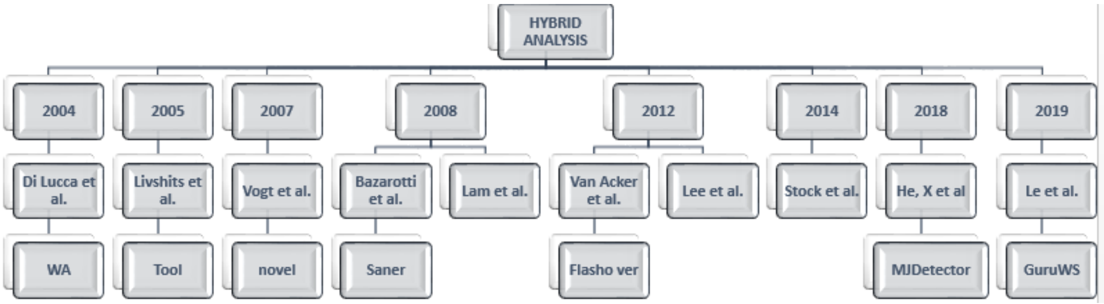


Fig. 6. Methods of Hybrid Analysis.

1) Discussion

Extracts of static and dynamic analysis are mixed to be named as hybrid analysis; it provides a path toward precision analysis. Hybrid Analysis (also called correlation) combines DAST and SAST to correlate and verify the results. Issues identified using dynamic analysis that will be traced to the offending line of code. SAST issues can be automatically prioritized using DAST information. The challenge with hybrid analysis is that DAST relies on data being reflected in the browser, so if a SAST data flow is not reflected in the browser as a DAST issue. From the existing study, Le et al. [96] and Stock et al. [94] are useful approaches in hybrid Analysis. The methods of hybrid analysis are summarized year-wise shows in Fig. 6.

F. Machine Learning Technique

This Technique is utilized in a few application zones (e.g., computer games and robotics). Application security on the web is based on a diverse package of techniques as presented in Fig. 7. It empowers PCs to learn information without programming (coding) it, and afterward to utilize the obtained information to take activities/choices. PCs must be guided to learn before taking activities. They need an informational index of models – preparing informational collection – from which to remove information, gaining from that point. An undertaking is called arrangement on the off chance that it expects to appoint input objects into classes. A classifier is a programmed technique that does classification. A classifier proceeds the dataset to collect the features and classify the dataset and provide the result based on machine learning. Email spamming is a basic example to filter the emails [97]. Machine learning is a different method to prevent web vulnerabilities.

1) Vulnerability Prediction Models in light of Software Metrics

Characterization is a type of information investigation wherein models predict the result. Model is used to predict input data class labels because each training instance's class label is referred to as supervised learning [2].

2) Anomaly Detection Approaches

To extract a program source code model & recognize vulnerabilities as separate from the usual dominant parts and principles, this work class uses unsupervised learning. This technique model isn't utilized to the class in the dataset to prevent web vulnerabilities [2].

3) Vulnerable Code Pattern Recognition

This is another machine learning method that selects the specific patterns of vulnerable code from the data set and utilizes the pattern matching to prevent web vulnerabilities on web applications [2].

4) Miscellaneous Paths

This method is used in the area of AI and data science for programming weakness in software programming and disclosure, which are not suitable other previously mentioned classes

The dataset has some attributes that the set of all instance forms of a training dataset.Attributes are divided into two categories first is numerical and the second type is categorical. Illustrating the first category, it is either discrete of continuous and named as numerical attributes. Whereas, categorical attributes possess non-numerical and distinct values. Categorical attributes have a special kind of binary attributes. Binary attributes have two expected values that are either true or false [98]. Therefore, dividing and arrangement in the form of classes is a type of data examination. It includes extracting models that specify data categorically discrete or unordered class labels, these models are known as classifiers. Classification of data involves two phases: (1) learning, where the classification model is made; (2) classification, where class labels for given input data are predicted by the respective model. Supervised learning is a class in which each training instance is labeled. For example, the input of the classifier is managed in the sense that it is programmed to identify each training instance belongs to which class. An alternative type of machine learning is where each class is unidentified to any attributive vector such a type is known as unsupervised learning. Moreover, the process does not know the set of learned classes prior. [99]. each classifier utilizes an AI calculation that relies upon the learning type (supervised or unsupervised). Furthermore, the training data set is used to classify correctly about the input data. the selection of classifiers depends on the data set factors [100].

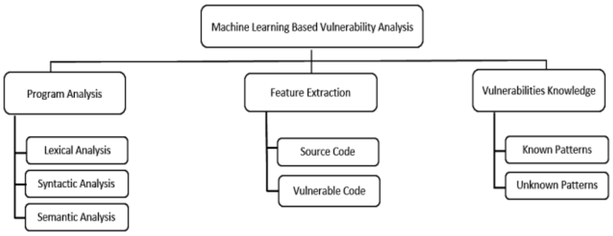


Fig. 7. Machine Learning-based Vulnerability Analysis.

Many researched have focused their studies to enhance the efficiency and precision of different techniques to detect web vulnerabilities. Support Vector Machines (SVM), J48, Artificial Neural Network, and many other classifications of techniques such as C5.0, Naïve Bayes, and linear regression are tested to train different datasets in order to detect web vulnerabilities. They are majorly grouped in two categories: probabilistic and machine learning. These techniques provided algorithms that are proved fruitful to cope with web vulnerability issues. Selected algorithms are analyzed by four metrics of, precision, recall, F1-score, and accuracy.

Suggested vulture tool in a study by Neuhaus et al. [101]. This unique tool will automatically explore existing vulnerabilities in archives for databases and versions. Vulture uses mine information to identify past component vulnerabilities. In addition, the components identified are classified by the most vulnerable to at least one type. This ranking of these components serves as a ground for investigations of the factors, which causes vulnerability to the targeted component. For instance, the study on the history of Mozilla vulnerability reveals an unexpected outcome that the components have one past vulnerability are mostly not affected by more vulnerabilities. In addition to it, those components that have the same functions calls are prone to vulnerabilities.

Machine learning has been utilized in certain attempts to quantify programming quality by gathering traits that uncover the nearness of software defects. Code type, counts of code lines, code metrics complication, and objected-oriented topographies are attributed in various studies. Some studies move ahead to consider the same type of metrics to guess the presence of vulnerabilities in source code. Moreover, other factors like past vulnerability events, called function and complication in codes are also used to conduct various other studies. These studies are not focused to identify bugs and mark their respective location but aim to examine the software codes according to the frequency of defect and vulnerabilities code[102].

Wang Yanya et al. [103] studied rapid density clustering called DSVRDC and intended methods to identify vulnerabilities in software using DCVRDC. Density-dependent clustering of vulnerability orders detected. The classifications examined are determined by the s-order difference. The density clustering methodology based on Rd-entropy is used to create vulnerability sequences in the first stage. Secondly, the respective vulnerabilities of the software are compared by s-order variation. Each order is dedicated to every cluster to calculate the difference in s-order as well as clusters comprises of under investigation software vulnerabilities are also computed.

In their study, Yamaguchi, Lindner, and Rieck [104] proposed a method to examine source code to detect vulnerabilities. This method schematically recognizes the API symbols of each function using lexical analysis. Then by the principal component analysis technique API symbols are introduced in vector space, and in dimension data in order to calculate the usage of API mode. Later on, the API usage mapping is created along with the estimated functions, supervisory code evaluation to classify likely vulnerabilities by utilizing known vulnerability functions as a standpoint. Another research is led in 2012. In order to protect SQLI and XSS vulnerabilities, Scholte et al [55]. combined static analysis and machine-learning and established IPAAS. This collection of information is utilized to deduce authentication policies about input fields that are helpful to save in-process attacks.

In another study by Wijayasekara et al. [105], a text mining technique was studied to remove potential vulnerabilities in the public bug database. This method creates a matrix for the term document. The mentioned process uses a text-mining technique to reach to the final task of classification of feature vectors into normal bug or vulnerability. The author has also purposed the increasing proportion of concealed vulnerabilities influence occurred during the past two years which is 53% for 53% for Linux and 10% for MySQL.

Another research is led in 2012 Nunan et al. [106], [107]. Likewise, recuperate web record and URL based highlights from an enormous box of an assortment of XSS assaults to examine how to depict the assaults and sort new potential XSS vector assaults as vindictive. Due to this enormous assortment, they perceived a lot of highlights (obscurity based, far fetched examples, and HTML/JavaScript plans) that license the specific arrangement of XSS in pages. At that point, they investigate consequently website pages to distinguish XSS assaults. These are the three stages process: one is identification and extraction of muddled highlights, the second one is unraveling of the website pages and includes, and the last one is the arrangement of pages by methods for an AI calculation.

Standard classifiers and other normal information-digging methods just search for the nearness of qualities, without relating them or thinking about their request. This can startthe wrong order and forecast. In earlier years, this perspective has been contemplated for improving exactness. Khosronejad [108] also aim to reduce the time of training during the construction of the HMM. They propose to assemble a model dependent on separated regular normal examples in follow occasions as opposed to taking each follow all alone. The follows are standard calls, since they can recognize the likelihood of deformities, by abnormal capacity call or by illicit utilization of assets as a result of assaults.. Bhole et al. [109] contrast the aftereffects of HMM and standard classifiers for the identification of oddities performed by an IDS. They infer that the HMM performs superior to the others. Another significant investigation shar and Tan [110]set forward their endeavors to recognize web vulnerabilities and to order different info sterilization methods in various classes as a lot of static code and a device called PHPMINER-1. In an investigation by Shar, Tan and Briand [111], they evaluated dynamic ascribe helplessness to supplement static traits. What's more, they utilized directed learning and estimation maps that are made together on the course of action and bunching to figure vulnerabilities. Both of these can perform exclusively in the nearness or nonattendance of marked preparing information. Creator presumed that they are appropriate without marked preparing information also.

In another examination by Soska and Christin [112]. The purpose of the study is to foresee the status of the site that will get vindictive later on or not before it is truly undermined. This is extremely useful by utilizing AI since they are effectively recovered includes about the server of site and the facilitating subtleties of sites. The highlights removed about the sites are, for example, the structure of record framework (e.g., catalog names that show that the site is made by CMS), the structure of the page (e.g., if the site page is made by a CMS format), and the catchphrases (e.g., presence of some HTML labels). It depends on the event of these highlights; they perceive whether a site will be undermined. In another study using the machine learning technique, Howard et al. [25] proposed the Psigene system to retrieve features from a large SQL injection attack collection box to study how to describe them.

Another study led in 2014 [113], Fabian et al. purposed a technique for efficient big source code data analysis to find the vulnerabilities. The author presented a code property graph for illustration of source code in a new way. These graphs combined the idea of standard program analysis that includes abstract syntax trees, managed flow graphs, and graph of programs into a collective data assembly. We can characterize integer overflows, buffer overflows, vulnerabilities in format strings, or memory disclosures. The purposed collective informative structural model for vulnerabilities in addition to their graphs representation makes a person aware of all the above-mentioned factors. The creditability of this technique is identified by real-time application in some well-known graph database, it is successful in the Linux kernel to find eighteen unfamiliar vulnerabilities in the source code. A technique for detecting RCE, XSS, LFI/RFI, and SQLi was developed by Singh et al. [114]. This study proposed a work to improve the accuracy of the current vulnerability finding scheme. Grieco et al.'s [115] recent study, suggested a method for estimating a vulnerability by blurring. This approach deduces topologies that negate memory by analysis of a binary program. In the consequence of this analysis, all the extracted results are classified to assist machine learning. VDISCOVER is used to check if the test category has vulnerabilities. 1039 program are observed using bug hunter to extract 138308 performance sets in order to statistically investigating 76083 different function calls. Methods are proven effective as the test results have been detected and certain memory leaks have been confirmed.

In another study, Medeiros et al. [116] proposed a new approach to deduce by extraction algorithm the basic and context structure of source code to identify vulnerabilities in web applications. The author also stated context-sensitive security flaws in the prevailing most distinguished XSS (cross-site scripting) technique to find the vulnerability. It is found that the XSS methodology is unable to include user input in the output statements. In Walden, Stuckman, and Scandariato, [117], compared two important feature software metrics and text mining of web vulnerabilities. The author tried to establish a prediction model comprising for PHP. Both the techniques are cross-validated. Application with a version named as Drupal 6.0, PHP My Admin 3.3, and Moodle 2.0. are selected for cross-validation test. Validity test is performed twice; software metrics and term frequency parameters are used respectively to guess vulnerability. After this, results are compared and eminence of guess parameters is analyzed.

In their study YUN et al., [118], gave a new technology VULPREDICTOR that investigates metrics and text mining to guess vulnerable files. At last, it purposes a compound prediction model. First VULPREDICTOR builds 6 basic classifiers on a file under observation in order to produce constructs a Meta classifier. These files are classified according to their text parameters and software algorithms. This method run in two stages firstly it constructs a model then comes prediction stages. In the model construction stage, VULPREDICTOR constructs a composite structure from training source code files with (vulnerable or not) known labels. While in assuming point, this model works as to guess, whether a new source code file is vulnerable or not. In another study, Abunadi, Ibrahim, and mamdouh [119] developed an empirical study method that examines the effectiveness of cross-project prediction to guess vulnerabilities in software. The open-source datasets are incorporated and five famous classifiers are tested. The results of these classifiers are compared to check them in cross-project vulnerability prediction situations.

A study Anbiya et al. [120], focused on using PHP native token and Abstract Syntax Tree (AST) as features then manipulate them to get the best feature. We pruned the AST to dump some unusable nodes or subtrees and then extracted the node type token with Breadth First Search (BFS) algorithm. They were able to get the highest recall score at 92% with PHP token as features and Gaussian Naïve Bayes as a machine learning classification method. Another study in 2018, Kronjee et al [121] built a tool called WIRECAML a contrasted instrument with different devices for powerlessness identification in PHP code. The apparatus performed best for web vulnerabilities. They likewise gave approach a shot various open-source programming applications.

In this study Smitha et al. [122], work investigates the exhibition of calculations like choice woodland, neural systems, bolster vector machine, and strategic relapse. Their exhibition has been assessed utilizing standard execution measurements. HTTP CSIC 2010, a web interruption identification dataset is utilized right now. Test results demonstrate that SVM and LR have been predominant in their exhibition than their partners. Prescient work processes have been made utilizing Microsoft Azure Machine Learning Studio (MAMLS), a versatile AI stage that encourages an incorporated improvement condition to information researchers.

The study conducted by Noman et al. [123], fabricates 6 classifiers on a preparation set of named records spoke to by their product measurements and content highlights. Furthermore, they manufacture a Meta classifier, which consolidates the six hidden classifiers. NMPREDICTOR is assessed on datasets of three web applications, which offer 223 prevalent quality vulnerabilities found in PHPMyAdmin, Moodle, and Drupal. In their study Kudjo et al. [124], directed an observational examination on three open-source helplessness datasets, to be specific Drupal, Moodle, and PHPMyAdmin utilizing five AI calculations. Shockingly, they found that in all instances of the 3 datasets considered, models gave a critical increment in accuracy and precision against the benchmark study. Zhou et al. [125] study presents an improved algorithm that generates test cases. This algorithm uses a new mutation method to divide test cases into various functional units to preserve their semantic structure. The results showed their algorithm not only generated better cases as compared with standard genetic algorithm and the adaptive genetic algorithm but also detected web vulnerabilities with high accuracy. Another study in machine learning is conducted by Tang et al. [126] that The statistical analysis of normal and SQL injection data was used to design eight feature types and train a machine-learning model. The accuracy of this model was 99%. The study proposed by Williams et al. [127] an integrated framework of data mining. This framework was capable of detecting evolution of web vulnerabilities. This framework three specific techniques i.e. Topically Supervised Evolution Model and Diffusion-Based storytelling technique, and prediction models. Through a series of experiments, it was shown this proposed framework not only discovered the evolution of web vulnerabilities and predict them with high level of accuracy. The methodology proposed by Calzavara et. al., [128] utilized machine learning to detect web application vulnerabilities. They used this methodology in Mitch. Mitch was the first machine-learning based solution to detect cross-site request forgeries.the detailed research review on machine learning to prevent web vulnerabilities shown in Table VI.

TABLE VI. MACHINE LEARNING EXISTING STUDIES

Research article	**Language /Framework**	**Metrics / Feature**	**Year**	**Dataset**	**Classifier**	**ML Method**	**Web Vulnerabilities**	**Performance Parameters**	**Application**
Neuhaus et al. [101]	vulture tool	14 components importing nsNodeUtils.h	2007	Mozilla with 134 vulnerabilities	SVM	A1	Security vulnerabilities	Precision, Recall	Mozilla Firefox
Wang Yanya et al. [103]	DSVRDC	67 series software Apache	2011	open-source web server software Apache httpd 2.2.8	Rd-entropy.	A2	Security vulnerabilities	Accuracy	C++ programming language
Yamaguchi et al. [104]	*	Extracting AST with a parser	2011	*	*	A2	*	*	C++ programming language
Wijayasekara et al. [105]	open bug database	Feature Vector	2012	Linux kernel vulnerabilities (Redhat Bugzilla)	Bayesian	A3	SQLi	Accuracy	hidden impact bugs
Nunan et al. [107]	Experimental study	obfuscation-based, doubtful patterns and html/JavaScript schemes	2012	15.366 websites XSSed database, dmoz.org and (2) 158.847	NB And SVM	A1	XSS	Detection rate, Accuracy rate and False alarm rate	HTML/JavaScript and PHP
Shar and Tan [110]	PHPMINER-1.	bytecode rewriting	2012	Java-based open source applications, Events, Classifieds, Roomba, PersonalBlog, and JGossip	*	A2	XSS, SQL	*	HTML/JavaScript and PHP
Soska and Christin [112]	complementary approach	structure of the file system, the structure of the webpage and the keyword	2014	PhishTank, search redirection attacks,	C4.5	A1	all	accuracy	all
Fabian et al. [113]	Code Property Graph	Extracting AST with a parser	2014	central vulnerability database by MITRE of CVE	*	A2	BO, Memory Mapping, Zero-byte Allocation	*	all
Howard et al. [25]	Psigene system Framework	SQL reserved words, SQLi signatures from the Bro, Snort IDS and the ModSecurity, web of WAF and SQLi reference documents	2014	the Exploit Database, PacketStorm Security, and the Open Source Vulnerability Database	Logistic regression	A2	SQL injection	Accuracy, Precision	PHP

Grieco et al. [115]	VDISCOVER method	N-gram analysis on the function call sequences	2016	bag-of-words, word2vec	Logistic Regression MLP	A2	vulnerabilities in operating systems	Accuracy	*
Medeiros et al. [116]	Method	Aggregate function, Numeric entry point, Complex query, Extract substring, String concatenation, Add char, Replace string, Error/exit, Remove whitespaces, Type checking, An entry point is set, Pattern control,	2016	Custome Dataset	ID3, C4.5/J48, RF, K-NN NB, MLP, SVM ,Logistic, Forest Tree, Bayes Net	*	SQLi, XSS,	Accuracy, Precession, Recall	PHP
Walden, Stuckman, and Scandariato [117]	Dataset Created	Software metrics, text mining	2014	Drupal, PHPMyAdmin , and Moodle with 223 web vulnerabilities	Random Forest	*	SQL injection, XSS, CSRF, and others	Accuracy, Precession, Recall	PHP
YUN et al.., [118]	VULPREDICTOR	Software metrics, text mining	2016	Drupal, PHPMyAdmin , and Moodle with 223 web vulnerabilities	Random Forest, Naïve Bayes, J48	*	SQL injection, XSS, CSRF and others	Accuracy, Precession, Recall	PHP
Abunadi, ibrahim, and mamdouh [119]	Proposed Method	Software metrics, text mining	2016	Drupal, PHPMyAdmin , and Moodle with 223 web vulnerabilities	RF, LR, SVM, J48, and NB	*	SQL injection, XSS, CSRF and others	Accuracy, Precession, Recall	PHP
Anbiya et al[120]	Proposed Method	PHP Tokens	2018	NVD	SVM, DT	A1	SQL injection, XSS,and others	Accuracy, Precession, Recall	PHP
Kronjee et al[121]	WIRECAML	All Features	2018	National Vulnerability Database and the SAMATE	probabilistic classifiers	A1	SQL injection, XSS,	Accuracy,	PHP
Smitha et al [122]	Comparative study ss	All Features	2019	HTTP CSIC 2010	SVM and LR	A1	QLI, XSS, LDAP, and Buffer overflow	Accuracy, Precession, Recall	PHP
Noman et al[123]	NMPREDICTOR	Software metrics, text mining	2019	Drupal, PHPMyAdmin , and Moodle with 223 web vulnerabilities	. J48, Naive Bayes and Random forest	A2	SQL injection, XSS, CSRF and others	Accuracy, Precession, Recall	PHP
Kudjo et al[124]	Model	Software metrics, text mining	2019	Drupal, PHPMyAdmin , and Moodle with 223 web vulnerabilities	RF, SVM, KNN, MLP, C4,5	A2	SQL injection, XSS, CSRF and others	Accuracy, Precession, Recall	PHP
Zhou et al [125]	Proposed Method	All Features	2020	Test cases	genetic algorithm	A3	SQL injection, XSS,	Comparative	PHP
Tang et al [126]	Model	eight Features	2020	Normal Dataset	MLP model	A1	SQL injection	accuracy	*
Williams et al [127]	framework	All Features	2020	*	*	A1 & A2	Vulnerabilties	*	*

* Missing in Paper
A3 Miscellaneous Approaches
A1:Vulnerability Prediction Models based on Software Metrics: A2: Vulnerable Code Pattern Recognition
A4: n-grams extraction algorithms.

1) Discussion Machine learning is considered very different approach with a wide range of web applications. However, it can also use to find out web vulnerabilities in source code. It is a very important area in today's collaborative work environment to detect 0day web vulnerabilities and new approaches are always desirable including the current and existing once. Many researched have focused their studies to enhance the efficiency and precision of different techniques to detect web vulnerabilities. Support Vector Machines (SVM), J48, Artificial Neural Network, and many other classifications of techniques such as K-Nearest Neighbor, C5.0, Naïve Bayes, and linear regression are tested to train different datasets in order to detect web vulnerabilities. Furthermore, mostly researcher evaluate their result with machine learning parameters such as precision, recall, F1-score, and accuracy. From the existing, study noman et al. [127], Medeiros et al. [12] are useful approaches in Machine learning.

V. Conclusion

This study provides a comprehensive survey of existing methods in the research area of web applications vulnerabilities. We highlighted several open issues that still needs to be addressed. In this paper, we reviewed classification and detection of web vulnerabilities with different approaches like static analysis, dynamic analysis, hybrid analysis, combined three analyses for scanners and machine learning. We also reviewed various types of web vulnerabilities with different classification. The input validation vulnerabilities and improper session management and methods to perceive web vulnerabilities. There are lot of works that have been performed to cater to such issues. The best approach we identified to secure a web application is concluded such as for secure programming is Agrawal et al (2019), Kang and Park (2017), Zhu et al. (2014), in case of Static analysis is Nunes, P et al. (2019) Khalid et al. (2018) and NUNES et al (2015). Furthermore, in case of Dynamic analysis is Park et al. (2019), Kang and Park (2017) and Zhu et al. (2014), in case of Hybrid analysis is Le et al. 2019 and Stock et al. (2014), however, in case of machine learning is noman et al (2019), Medeiros et al. (2016).

References

[1] Pop, D. P., & Altar, A. (2014). Designing an MVC model for rapid web application development. Procedia Engineering, 69, 1172-1179.

[2] Deepa, G., & Thilagam, P. S. (2016). Securing web applications from injection and logic vulnerabilities: Approaches and challenges. Information and Software Technology, 74, 160-180.

[3] Awoleye, O. M., Ojuloge, B., & Ilori, M. O. (2014). Web application vulnerability assessment and policy direction towards a secure smart government. Government Information Quarterly, 31, S118-S125.

[4] Bozic, J., & Wotawa, F. (2015, August). PURITY: a Planning-based secURITY testing tool. In Software Quality, Reliability and Security-Companion (QRS-C), 2015 IEEE International Conference on (pp. 46-55). IEEE.

[5] Duchene, F., Rawat, S., Richier, J. L., & Groz, R. (2014, March). KameleonFuzz: evolutionary fuzzing for black-box XSS detection. In Proceedings of the 4th ACM conference on Data and application security and privacy (pp. 37-48). ACM.

[6] Medeiros, I., Neves, N. F., & Correia, M. (2014, April). Automatic detection and correction of web application vulnerabilities using data mining to predict false positives. In Proceedings of the 23rd international conference on World Wide Web (pp. 63-74). ACM.

[7] Tsipenyuk, K., Chess, B., & McGraw, G. (2005). Seven pernicious kingdoms: A taxonomy of software security errors. IEEE Security & Privacy, 3(6), 81-84.

[8] Mitropoulos, D., Louridas, P., Polychronakis, M., & Keromytis, A. D. (2017). Defending Against Web Application Attacks: Approaches, Challenges and Implications. IEEE Transactions on Dependable and Secure Computing.

[9] Li, Xiaowei, and Yuan Xue. "A survey on server-side approaches to securing web applications." ACM Computing Surveys (CSUR) 46, no. 4 (2014): 54.

[10] Wedman, S., Tetmeyer, A., & Saiedian, H. (2013). An analytical study of web application session management mechanisms and HTTP session hijacking attacks. Information Security Journal: A Global Perspective, 22(2), 55-67.

[11] Shahriar, H., & Zulkernine, M. (2010, November). Client-side detection of cross-site request forgery attacks. In Software Reliability Engineering (ISSRE), 2010 IEEE 21st International Symposium on (pp. 358-367). IEEE.

[12] Delgado, Nelly, Ann Q. Gates, and Steve Roach. "A taxonomy and catalog of runtime software-fault monitoring tools." IEEE Transactions on software Engineering 30, no. 12 (2004): 859-872.

[13] Igure, Vinay M., and Ronald D. Williams. "Taxonomies of attacks and vulnerabilities in computer systems." IEEE Communications Surveys & Tutorials 10, no. 1 (2008)

[14] Krsul, Ivan Victor. Software vulnerability analysis. West Lafayette, IN: Purdue University, 1998.

[15] Halfond, William G., Jeremy Viegas, and Alessandro Orso. "A classification of SQL-injection attacks and countermeasures." In Proceedings of the IEEE International Symposium on Secure Software Engineering, vol. 1, pp. 13-15. IEEE, 2006.

[16] Chandrashekhar, Roshni, Manoj Mardithaya, Santhi Thilagam, and Dipankar Saha. "SQL injection attack mechanisms and prevention techniques." In International Conference on Advanced Computing, Networking and Security, pp. 524-533. Springer, Berlin, Heidelberg, 2011.

[17] Garcia-Alfaro, Joaquin, and Guillermo Navarro-Arribas. "A survey on cross-site scripting attacks." arXiv preprint arXiv: 0905.4850 (2009).

[18] M. Cova, V. Felmetsger, G. Vigna, Vulnerability analysis of web-based applications, in: Test and Analysis of Web Services, Springer Berlin Heidelberg, 2007, pp. 363–394.

[19] NUNES, P., FONSECA, J. & VIEIRA, M. (2015). phpSAFE: A security analysis tool for OOP web application plugins. In Proceedings of the 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks.

[20] Shahriar, Hossain, and Mohammad Zulkernine. "Taxonomy and classification of automatic monitoring of program security vulnerability exploitations." Journal of Systems and Software84, no. 2 (2011): 250-269.

[21] Hydara, Isatou, Abu Bakar Md Sultan, Hazura Zulzalil, and Novia Admodisastro. "Current state of research on cross-site scripting (XSS)–A systematic literature review." Information and Software Technology 58 (2015): 170-186.

[22] Buczak, Anna L., and Erhan Guven. "A survey of data mining and machine learning methods for cyber security intrusion detection." IEEE Communications Surveys & Tutorials 18, no. 2 (2016): 1153-1176.

[23] Ghaffarian, Seyed Mohammad, and Hamid Reza Shahriari. "Software Vulnerability Analysis and Discovery Using Machine-Learning and Data-Mining Techniques: A Survey." ACM Computing Surveys (CSUR) 50, no. 4 (2017): 56.

[24] Shahriar, H., & Zulkernine, M. (2012, October). Information-theoretic detection of sql injection attacks. In High-Assurance Systems Engineering (HASE), 2012 IEEE 14th International Symposium on (pp. 40-47). IEEE.

[25] Howard, G. M., Gutierrez, C. N., Arshad, F. A., Bagchi, S., & Qi, Y. (2014, June). psigene: Webcrawling to generalize sql injection signatures. In Dependable Systems and Networks (DSN), 2014 44th Annual IEEE/IFIP International Conference on (pp. 45-56). IEEE.

[26] Juillerat, N. (2007, October). Enforcing code security in database web applications using libraries and object models. In Proceedings of the 2007 Symposium on Library-Centric Software Design (pp. 31-41). ACM.

[27] Johns, M., Beyerlein, C., Giesecke, R., & Posegga, J. (2010). Secure Code Generation for Web Applications. ESSoS, 5965, 96-113. Chicago

[28] Grabowski, R., Hofmann, M., & Li, K. (2011). Type-Based Enforcement of Secure Programming Guidelines-Code Injection Prevention at SAP. Formal Aspects in Security and Trust, 7140, 182-197.

[29] Chong, S., Vikram, K., & Myers, A. C. (2007, August). SIF: Enforcing Confidentiality and Integrity in Web Applications. In USENIX Security Symposium (pp. 1-16).

[30] Vikram, K., Prateek, A., & Livshits, B. (2009, November). Ripley: automatically securing web 2.0 applications through replicated execution. In Proceedings of the 16th ACM conference on Computer and communications security (pp. 173-186). ACM.

[31] Yip, A., Wang, X., Zeldovich, N., & Kaashoek, M. F. (2009, October). Improving application security with data flow assertions. In Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles (pp. 291-304). ACM.

[32] Jia, L., Vaughan, J. A., Mazurak, K., Zhao, J., Zarko, L., Schorr, J., & Zdancewic, S. (2008, September). Aura: A programming language for authorization and audit. In ACM Sigplan Notices (Vol. 43, No. 9, pp. 27-38). ACM.

[33] Swamy, N., Corcoran, B. J., & Hicks, M. (2008, May). Fable: A language for enforcing user-defined security policies. In Security and Privacy, 2008. SP 2008. IEEE Symposium on (pp. 369-383). IEEE.

[34] Corcoran, B. J., Swamy, N., & Hicks, M. (2009, June). Cross-tier, label-based security enforcement for web applications. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (pp. 269-282). ACM.

[35] Swamy, N., Chen, J., & Chugh, R. (2010, March). Enforcing Stateful Authorization and Information Flow Policies in Fine. In ESOP (pp. 529-549).

[36] Krishnamurthy, A., Mettler, A., & Wagner, D. (2010, April). Fine-grained privilege separation for web applications. In Proceedings of the 19th international conference on World wide web (pp. 551-560). ACM.

[37] Zhu, J., Xie, J., Lipford, H. R., & Chu, B. (2014). Supporting secure programming in web applications through interactive static analysis. Journal of advanced research, 5(4), 449-462

[38] Kang, J., & Park, J. H. (2017). A secure-coding and vulnerability check system based on smart-fuzzing and exploit. Neurocomputing.

[39] Na Meng, Stefan Nagy, Danfeng Yao, Wenjie Zhuang, and Gustavo Arango-Argoty. Secure coding practices in java: Challenges and vulnerabilities. In 2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE). IEEE, 372--383, 2018.

[40] Bangani, S., Futcher, L., & van Niekerk, J. (2019, July). An Approach to Teaching Secure Programming in the. NET Environment. In Annual Conference of the Southern African Computer Lecturers' Association (pp. 35-49). Springer, Cham.

[41] Agrawal, A., Alenezi, M., Kumar, R., & Khan, R. A. (2019). A source code perspective framework to produce secure web applications. Computer Fraud & Security, 2019(10), 11-18

[42] Jovanovic, N., Kruegel, C., & Kirda, E. (2006, May). Pixy: A static analysis tool for detecting web application vulnerabilities. In Security and Privacy, 2006 IEEE Symposium on (pp. 6-pp). IEEE.

[43] Chess, B., & McGraw, G. (2004). Static analysis for security. IEEE Security & Privacy, 2(6), 76-79.

[44] Durães, João, and Henrique Madeira. "A methodology for the automated identification of buffer overflow vulnerabilities in executable software without source-code." In Latin-American Symposium on Dependable Computing, pp. 20-34. Springer, Berlin, Heidelberg, 2005.

[45] Huang, Y. W., Yu, F., Hang, C., Tsai, C. H., Lee, D. T., & Kuo, S. Y. (2004, May). Securing web application code by static analysis and runtime protection. In Proceedings of the 13th international conference on World Wide Web (pp. 40-52). ACM.

[46] Wassermann, G., & Su, Z. (2008, May). Static detection of cross-site scripting vulnerabilities. In Proceedings of the 30th international conference on Software engineering (pp. 171-180). ACM.

[47] Minamide, Y. (2005, May). Static approximation of dynamically generated web pages. In Proceedings of the 14th international conference on World Wide Web (pp. 432-441). ACM.

[48] Xie, Y., & Aiken, A. (2006, July). Static Detection of Security Vulnerabilities in Scripting Languages. In USENIX Security Symposium (Vol. 15, pp. 179-192).

[49] Halfond, W. G., & Orso, A. (2005, November). AMNESIA: analysis and monitoring for NEutralizing SQL-injection attacks. In Proceedings of the 20th IEEE/ACM international Conference on Automated software engineering (pp. 174-183). ACM.

[50] Thomas, S., & Williams, L. (2007, May). Using automated fix generation to secure SQL statements. In Proceedings of the Third International Workshop on Software Engineering for Secure Systems (p. 9). IEEE Computer Society.

[51] SON, S. & SHMATIKOV, V. (2011). SAFERPHP: Finding semantic vulnerabilities in PHP applications. In Proceedings of the ACM SIGPLAN 6th Workshop on Programming Languages and Analysis for Security.

[52] J. Fonseca, N. Seixas, M. Vieira, H. Madeira, Analysis of field data on web security vulnerabilities, IEEE Transactions on Dependable and Secure Computing 11 (2) (2014) 89–100.

[53] Shahriar, H., & Zulkernine, M. (2009, May). Mutec: Mutation-based testing of cross site scripting. In Proceedings of the 2009 ICSE Workshop on Software Engineering for Secure Systems (pp. 47-53). IEEE Computer Society.

[54] Shar, Lwin Khin, and Hee Beng Kuan Tan. "Automated removal of cross site scripting vulnerabilities in web applications." Information and Software Technology 54, no. 5 (2012): 467-478.

[55] Scholte, T., Robertson, W., Balzarotti, D., & Kirda, E. (2012, July). Preventing input validation vulnerabilities in web applications through automated type analysis. In Computer Software and Applications Conference (COMPSAC), 2012 IEEE 36th Annual (pp. 233-243). IEEE.

[56] Zheng, Y., & Zhang, X. (2013, May). Path sensitive static analysis of web applications for remote code execution vulnerability detection. In Proceedings of the 2013 International Conference on Software Engineering (pp. 652-661). IEEE Press.

[57] Doupé, A., Cui, W., Jakubowski, M. H., Peinado, M., Kruegel, C., & Vigna, G. (2013, November). deDacota: toward preventing server-side XSS via automatic code and data separation. In Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security (pp. 1205-1216). ACM.

[58] Amira, Abdelouahab, Abdelraouf Ouadjaout, Abdelouahid Derhab, and Nadjib Badache. "Sound and Static Analysis of Session Fixation Vulnerabilities in PHP Web Applications." In Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy, pp. 139-141. ACM, 2017.

[59] Muhammad Noman khalid, Muhammad Iqbal, Muhammad Talha Alam, Vishal Jain, Hira Mirza and Kamran Rasheed, "Web Unique Method (WUM): An Open Source Blackbox Scanner for Detecting Web Vulnerabilities" International Journal of Advanced Computer Science and Applications(IJACSA), 8(12), 2017.

[60] Viega, John, Jon-Thomas Bloch, Yoshi Kohno, and Gary McGraw. "ITS4: A static vulnerability scanner for C and C++ code." In Computer Security Applications, 2000. ACSAC'00. 16th Annual Conference, pp. 257-267. IEEE, 2000.

[61] Deepa, G., Thilagam, P. S., Praseed, A., & Pais, A. R. (2018). DetLogic: A black-box approach for detecting logic vulnerabilities in web applications. Journal of Network and Computer Applications, 109, 89-109.

[62] Nunes, P., Medeiros, I., Fonseca, J., Neves, N., Correia, M., & Vieira, M. (2019). An empirical study on combining diverse static analysis tools for web security vulnerabilities based on development scenarios. Computing, 101(2), 161-185.

[63] NUNES, P., FONSECA, J. & VIEIRA, M. (2015). phpSAFE: A security analysis tool for OOP web application plugins. In Proceedings of the

45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks

[64] Long, H. V., Tuan, T. A., Taniar, D., Can, N. V., Hue, H. M., & Son, N. T. K. (2020). An efficient algorithm and tool for detecting dangerous website vulnerabilities. International Journal of Web and Grid Services, 16(1), 81-104.

[65] Aliero, M. S., Ghani, I., Qureshi, K. N., & Rohani, M. F. A. (2020). An algorithm for detecting SQL injection vulnerability using black-box testing. Journal of Ambient Intelligence and Humanized Computing, 11(1), 249-266.

[66] EVRON, G. & RATHAUS, N. (2007). Open Source Fuzzing Tools. Elsevier Inc., 1st edn.

[67] SUTTON, M., GREENE, A. & AMINI, P. (2007). Fuzzing: Brute Force Vulnerability Discovery. Addison-Wesley, 1st edn.

[68] Nguyen-Tuong, A., Guarnieri, S., Greene, D., Shirley, J., & Evans, D. (2005). Automatically hardening web applications using precise tainting. Security and Privacy in the Age of Ubiquitous Computing, 295-307.

[69] Haldar, V., Chandra, D., & Franz, M. (2005, December). Dynamic taint propagation for Java. In Computer Security Applications Conference, 21st Annual (pp. 9-pp). IEEE.

[70] Jimenez, W., Mammar, A., & Cavalli, A. (2009). Software vulnerabilities, prevention and detection methods: A review1. Security in Model-Driven Architecture, 6.

[71] Miller, B. P., Fredriksen, L., & So, B. (1990). An empirical study of the reliability of UNIX utilities. Communications of the ACM, 33(12), 32-44.

[72] BRADSHAW, S. (2010). An introduction to fuzzing: Using fuzzers (spike) to find vulnerabilities. http://resources.infosecinstitute.com/intro-to-fuzzing/.bufferoverflow vulnerabilities in executable software without source-code. In Proceedings of the 2nd Latin-American Conference on Dependable Computing, 20–34.

[73] Huang, Y. W., Tsai, C. H., Lin, T. P., Huang, S. K., Lee, D. T., & Kuo, S. Y. (2005). A testing framework for Web application security assessment. Computer Networks, 48(5), 739-761.

[74] Kals, S., Kirda, E., Kruegel, C., & Jovanovic, N. (2006, May). Secubat: a web vulnerability scanner. In Proceedings of the 15th international conference on World Wide Web (pp. 247-256). ACM.

[75] DUCHÈNE, F., RAWAT, S., RICHIER, J. & GROZ, R. (2014). Kameleonfuzz: Evolutionary fuzzing for black-box XSS detection. In Proceedings of the 4th ACM Conference on Data and Application Security and Privacy, 37–48.

[76] Antunes, N., & Vieira, M. (2010, July). Benchmarking vulnerability detection tools for web services. In Web Services (ICWS), 2010 IEEE International Conference on (pp. 203-210). IEEE.

[77] Ciampa, A., Visaggio, C. A., & Di Penta, M. (2010, May). A heuristic-based approach for detecting SQL-injection vulnerabilities in Web applications. In Proceedings of the 2010 ICSE Workshop on Software Engineering for Secure Systems (pp. 43-49). ACM.

[78] Stock, B., Lekies, S., & Johns, M. (2013). 25 Million Flows Later-Large-scale Detection of DOM-based XSS. 20th CCS. ACM.

[79] HALLER, I., SLOWINSKA, A., NEUGSCHWANDTNER, M. & BOS, H. (2013). Dowsing for overflows: A guided fuzzer to find buffer boundary violations. In Proceedings of the 22nd USENIX Security Symposium, 49–64.

[80] Dahse, J., & Holz, T. (2014, August). Static Detection of Second-Order Vulnerabilities in Web Applications. In USENIX Security Symposium (pp. 989-1003).

[81] Weissbacher, M., Robertson, W. K., Kirda, E., Kruegel, C., & Vigna, G. (2015, August). ZigZag: Automatically Hardening Web Applications against Client-side Validation Vulnerabilities. In USENIX Security Symposium (pp. 737-752).

[82] RanWang, GuangquanXu, XianjiaoZeng, XiaohongLi, ZhiyongFeng. (2018). TT-XSS: A novel taint tracking based dynamic detection framework for DOM Cross-Site Scripting.Journal of Parallel and Distributed Computing Volume 118, Part 1, August 2018, Pages 100-106

[83] Park, J., Choo, Y., & Lee, J. (2019). A hybrid vulnerability analysis tool using a risk evaluation technique. Wireless Personal Communications, 105(2), 443-459.

[84] CHIPOUNOV, V., KUZNETSOV, V. & CANDEA, G. (2011). S2e: A platform for in-vivo multi-path analysis of software systems. In Proceedings of the 16th International Conference on Architectural Support for Programming Languages and Operating Systems, 265–278.

[85] Falana, O. J., Ebo, I. O., Tinubu, C. O., Adejimi, O. A., & Ntuk, A. (2020, March). Detection of Cross-Site Scripting Attacks using Dynamic Analysis and Fuzzy Inference System. In 2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS) (pp. 1-6). IEEE.

[86] Di Lucca, G. A., Fasolino, A. R., Mastoianni, M., & Tramontana, P. (2004, September). Identifying cross site scripting vulnerabilities in web applications. In Web Site Evolution, Sixth IEEE International Workshop on (WSE'04) (pp. 71-80). IEEE.

[87] Balzarotti, D., Cova, M., Felmetsger, V., Jovanovic, N., Kirda, E., Kruegel, C., & Vigna, G. (2008, May). Saner: Composing static and dynamic analysis to validate sanitization in web applications. In Security and Privacy, 2008. SP 2008. IEEE Symposium on (pp. 387-401). IEEE.

[88] Livshits, V. B., & Lam, M. S. (2005, August). Finding Security Vulnerabilities in Java Applications with Static Analysis. In USENIX Security Symposium (Vol. 14, pp. 18-18).

[89] Lam, M. S., Martin, M., Livshits, B., & Whaley, J. (2008, January). Securing web applications with static and dynamic information flow tracking. In Proceedings of the 2008 ACM SIGPLAN symposium on Partial evaluation and semantics-based program manipulation (pp. 3-12). ACM.

[90] Van Acker, S., Nikiforakis, N., Desmet, L., Joosen, W., & Piessens, F. (2012, May). FlashOver: Automated discovery of cross-site scripting vulnerabilities in rich internet applications. In Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security (pp. 12-13). ACM.

[91] Lee, I., Jeong, S., Yeo, S., & Moon, J. (2012). A novel method for SQL injection attack detection based on removing SQL query attribute values. Mathematical and Computer Modelling, 55(1), 58-68.

[92] Lee, T., Won, G., Cho, S., Park, N., & Won, D. (2012). Experimentation and Validation of Web Application's Vulnerability Using Security Testing Method. In Computer Science and its Applications (pp. 723-731). Springer, Dordrecht.

[93] Vogt, Philipp, Florian Nentwich, Nenad Jovanovic, Engin Kirda, Christopher Kruegel, and Giovanni Vigna. "Cross Site Scripting Prevention with Dynamic Data Tainting and Static Analysis." In NDSS, vol. 2007, p. 12. 2007.

[94] Stock, Ben, and Martin Johns. "Protecting users against XSS-based password manager abuse." In Proceedings of the 9th ACM symposium on Information, computer and communications security, pp. 183-194. ACM, 2014.

[95] He, X., Xu, L., & Cha, C. (2018, December). Malicious JavaScript Code Detection Based on Hybrid Analysis. In 2018 25th Asia-Pacific Software Engineering Conference (APSEC) (pp. 365-374). IEEE.

[96] Le, V. G., Nguyen, H. T., Pham, D. P., Phung, V. O., & Nguyen, N. H. (2019). GuruWS: A Hybrid Platform for Detecting Malicious Web Shells and Web Application Vulnerabilities. In Transactions on Computational Collective Intelligence XXXII (pp. 184-208). Springer, Berlin, Heidelberg.

[97] Hladka, B., & Holub, M. (2015). A Gentle Introduction to Machine Learning for Natural Language Processing: How to Start in 16 Practical Steps. Language and Linguistics Compass, 9(2), 55-76. Chicago

[98] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[99] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[100] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15.

[101] Neuhaus, S., Zimmermann, T., Holler, C., & Zeller, A. (2007, October). Predicting vulnerable software components. In Proceedings of the 14th ACM conference on Computer and communications security (pp. 529-540). ACM.

[102]SHIN, Y., MENEELY, A., WILLIAMS, L. & OSBORNE, J.A. (2011). Evaluating complexity, code churn, and developer activity metrics as indicators of software vulnerabilities. IEEE Transactions on Software Engineering, 37, 772–787.

[103]Wang, Y., Wang, Y., & Ren, J. (2011). Software Vulnerabilities Detection Using Rapid Density-based Clustering. JOURNAL OF INFORMATION &COMPUTATIONAL SCIENCE, 8(14), 3295-3302.

[104]Yamaguchi, F., Lindner, F., & Rieck, K. (2011, August). Vulnerability extrapolation: assisted discovery of vulnerabilities using machine learning. In Proceedings of the 5th USENIX conference on Offensive technologies (pp. 13-13). USENIX Association.

[105]Wijayasekara, D., Manic, M., Wright, J. L., & McQueen, M. (2012, June). Mining bug databases for unidentified software vulnerabilities. In Human System Interactions (HSI), 2012 5th International Conference on (pp. 89-96). IEEE.

[106]Nunan, A. E., Souto, E., dos Santos, E. M., & Feitosa, E. (2012, July). Automatic classification of cross-site scripting in web pages using document-based and URL-based features. In Computers and Communications (ISCC), 2012 IEEE Symposium on (pp. 000702-000707). IEEE.

[107]Sultana, A., Hamou-Lhadj, A., & Couture, M. (2012, June). An improved hidden markov model for anomaly detection using frequent common patterns. In Communications (ICC), 2012 IEEE International Conference on (pp. 1113-1117). IEEE.

[108]Khosronejad, M., Sharififar, E., Torshizi, H. A., & Jalali, M. (2013). Developing a hybrid method of Hidden Markov Models and C5. 0 as a Intrusion Detection System. International Journal of Database Theory and Application, 6(5), 165-174.

[109]Bhole, A. T., & Patil, A. I. (2014). Intrusion Detection with Hidden Markov Model and WEKA Tool. International Journal of Computer Applications, 85(13).

[110]Shar, L. K., & Tan, H. B. K. (2012, June). Mining input sanitization patterns for predicting SQL injection and cross site scripting vulnerabilities. In Proceedings of the 34th International Conference on Software Engineering (pp. 1293-1296). IEEE Press.

[111]Shar, L. K., Briand, L. C., & Tan, H. B. K. (2015). Web application vulnerability prediction using hybrid program analysis and machine learning. IEEE Transactions on Dependable and Secure Computing, 12(6), 688-707.

[112]Soska, K., & Christin, N. (2014, August). Automatically Detecting Vulnerable Websites Before They Turn Malicious. In USENIX Security Symposium (pp. 625-640).

[113]Yamaguchi, Fabian. Golde, N., Arp, D., & Rieck, K. (2014, May). Modeling and discovering vulnerabilities with code property graphs. In Security and Privacy (SP), 2014 IEEE Symposium on (pp. 590-604). IEEE.

[114]Singh, N., Dayal, M., Raw, R. S., & Kumar, S. (2016, March). SQL injection: Types, methodology, attack queries and prevention. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 2872-2876). IEEE.

[115]Grieco, G., Grinblat, G. L., Uzal, L., Rawat, S., Feist, J., & Mounier, L. (2016, March). Toward large-scale vulnerability discovery using Machine Learning. In Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy (pp. 85-96). ACM.

[116]Medeiros, I., Neves, N., & Correia, M. (2016). Detecting and removing web application vulnerabilities with static analysis and data mining. IEEE Transactions on Reliability, 65(1), 54-69.

[117]Walden, James, Jeff Stuckman, and Riccardo Scandariato. "Predicting vulnerable components: Software metrics vs text mining." In Software Reliability Engineering (ISSRE), 2014 IEEE 25th International Symposium on, pp. 23-33. IEEE, 2014.

[118]Zhang, Yun, David Lo, Xin Xia, Bowen Xu, Jianling Sun, and Shanping Li. "Combining software metrics and text features for vulnerable file prediction." In Engineering of Complex Computer Systems (ICECCS), 2015 20th International Conference on, pp. 40-49. IEEE, 2015.

[119]Abunadi, Ibrahim, and Mamdouh Alenezi. "An Empirical Investigation of Security Vulnerabilities within Web Applications." J. UCS 22, no. 4 (2016): 537-551.

[120]Anbiya, D. R., Purwarianti, A., & Asnar, Y. (2018, November). Vulnerability Detection in PHP Web Application Using Lexical Analysis Approach with Machine Learning. In 2018 5th International Conference on Data and Software Engineering (ICoDSE) (pp. 1-6). IEEE.

[121]Kronjee, J., Hommersom, A., & Vranken, H. (2018, August). Discovering software vulnerabilities using data-flow analysis and machine learning. In Proceedings of the 13th International Conference on Availability, Reliability and Security (pp. 1-10).

[122]Smitha, R., Hareesha, K. S., & Kundapur, P. P. (2019). A Machine Learning Approach for Web Intrusion Detection: MAMLS Perspective. In Soft Computing and Signal Processing (pp. 119-133). Springer, Singapore.

[123]Khalid, M. N., Farooq, H., Iqbal, M., Alam, M. T., & Rasheed, K. (2018, October). Predicting Web vulnerabilities in Web applications based on machine learning. In International Conference on Intelligent Technologies and Applications (pp. 473-484). Springer, Singapore.

[124]Kudjo, P. K., Aformaley, S. B., Mensah, S., & Chen, J. (2019, July). The Significant Effect of Parameter Tuning on Software Vulnerability Prediction Models. In 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C) (pp. 526-527). IEEE.

[125]Zhou, X., & Wu, B. (2020, June). Web Application Vulnerability Fuzzing Based On Improved Genetic Algorithm. In 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (Vol. 1, pp. 977-981). IEEE.

[126]Tang, P., Qiu, W., Huang, Z., Lian, H., & Liu, G. (2020). Detection of SQL injection based on artificial neural network. Knowledge-Based Systems, 190, 105528.

[127] Williams, M. A., Barranco, R. C., Naim, S. M., Dey, S., Hossain, M. S., & Akbar, M. (2020). A vulnerability analysis and prediction framework. Computers & Security, 92, 101751.

[128]Calzavara, S., Conti, M., Focardi, R., Rabitti, A., & Tolomei, G. (2020). Machine Learning for Web Vulnerability Detection: The Case of Cross-Site Request Forgery. IEEE Security & Privacy, 18(3), 8-16.

Data Warehouse System for Multidimensional Analysis of Tuition Fee Level in Higher Education Institutions in Indonesia

Ardhian Agung Yulianto[1], Yoshiya Kasahara[2]
Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa, Japan[1, 2]
Industrial Engineering Department, Universitas Andalas, Padang, Indonesia[1]

Abstract—**In this study, we developed a data warehouse (DW) system for tuition-fee-level management for higher education institutions (HEI) in Indonesia. The system was developed to provide sufficient information to the administrators for decision making of tuition fees of applicants by integrating multisource data. A simple but sufficient method was introduced using the open-source following the business requirements of HEI's administrator. As a business intelligence (BI) approach, four procedures are applied e.g., preparation, integration, analysis, and visualization to construct a tuition-fee-level management system. The DW demonstrate four basic dimensions (faculty, year, entrant type, and tuition fee level) in all seven dimensions and three data regarding applicants, tuition fee level, and payment status. Analytical results were tuition fee level trends, top five faculty by applicants, and fees collected from the student trends. Those analysis results were presented in various charts and graphics contained at a dashboard of tuition fee level, which has many functions to provide insight relative to the business performance. The DW system described in this paper can be used as a guideline for tuition-fee-level management for HEIs in Indonesia.**

Keywords—Data warehouse; higher education institution; multidimensional analysis; Indonesia; tuition-fee-level management

I. Introduction

The United Nations' Agenda for Sustainable Development Goals have identified higher education as an integral part of the lifelong learning vision to ensure high-quality education. As the instrument of higher education, higher education institutions (HEI) play an essential role in developing the national capacities of a country by educating students, publicizing research activities, and participating in the development of civil society. However, different countries and HEIs require different strategies to satisfy issues related to the access, affordability, and quality of higher education. Therefore, as in [1] differentiating tuition fees is one way for countries to adjust tuition fees.

In Indonesia, approximately three million high school graduates are competing to obtain one of approximately four hundred thousand regular seats in public HEIs [2]. In 2013, the Integrated Academic Fee (IAF) was introduced by the government to provide more opportunities to qualified citizens from low-income families. This policy implemented variable tuition fees for undergraduate students based on the financial ability of their parents or guardians [3]. To determine the level of tuition fees, HEIs are permitted to formulate parameters, e.g., family information (number of family members and number of children in college), income information (parent income), asset information (home or real estate equity), and monthly family expenses are commonly applied as general parameters. In addition, some techniques and methods to determine tuition fee level have been presented previously [4],[5].

HEIs face problems in managing the financial data after each applicant got their tuition fee level. For example, at Universitas Andalas, a public HEI in Indonesia, even though a web-based application was introduced to determine tuition fees for each applicant, different reports are submitted by university and faculty staff regarding the number of applicants and tuition fee collected at each level. Time is required to generate manual such reports. It is difficult for university staff to distribute students equitably among faculties and study programs when the data are inconsistent. The IAF states that the higher the level, the higher the amount of money paid by the applicants. The faculty is expected to receive as many applicants as possible who have been placed at a higher level of the tuition fee. Without consistent tuition fee data, HEI administrators do not have sufficient information to support financial decision making.

Another problem arises from the fact that the tuition-fee-level decision system functions independently of other information systems, e.g., student registration, payment, and teaching systems. Following student admission, the result of the tuition-fee-level decision system, i.e., the tuition-fee-level data for each applicant, must be migrated to these payment, registration, and teaching systems. IT staff must manage these data migration processes in consideration of the data structure of each faculty's database system. To provide reports, IT staff must raise an individual query for each database of faculty. As a result, the databases are frequently slow or even crash when retrieving reports.

To address these issues, the authors have investigated using open-source software to construct a data warehouse (DW) system, which is considered the backbone tool of a decision support system (DSS). The DW system consolidates various

This work was supported by the Indonesia Endowment Fund of Education (LPDP), Ministry of Research, Technology and Higher Education, Ministry of Finance of the Republic of Indonesia (BUDI-LN Scholarship) with reference No: PRJ-3678/LPDP.3/2016.

data sources from many transactional systems or files, and then stores them in an integrate information data store. The DW system also maintains historical data and provides analytical functionality to realize the users about the situation of their business [6], [7]. In addition, DW systems are considered a core component of business intelligence (BI), which is a general term that describes the analysis of information to improve and optimize business decisions and performance. In the remainder of this paper, as in [8], we use the term DW/BI to reflect the shift of emphasis from the DW being an end in itself to BI. The development of DW/BI in the education sector is very limited compared to other major sectors, e.g., financial services and the medical industry [9]. In [10] confirmed that many areas in the academic institution (e.g., enrollment data, course data, and alumni data) could identify data warehousing efforts and as in [11] described the importance of maintaining institutional strategy that accepts information systems as critical to decision making. In [12] was the first development of a DW in an HEI at Arizona State University in 1992. DW/BI in HEI has primarily been implemented in didactics and research fields [9], [13-15].

This study focuses on multidimensional analytics of the HEI tuition fee level in Indonesia under the IAF policy. Multidimensionally modeled data were designed to facilitate complex analysis and effortless visualization [13]. A data-driven decision approach was applied to enhance the DW system. The major features of a data-driven DSS are accessing and manipulating raw data and creating data displays [16]. These roles are performed by IT experts who know the metadata of the database systems and the tuition-fee-level decision system's workflow.

The goal of this study was to construct a DW system for tuition-fee-level management. A simple but sufficient method was used in this study compared to DW architecture and best practices for DW implementation in education. The proposed DW system was developed using the open-source Pentaho BI software suite, which includes a complete toolset for DW development. Note that the Community Edition of Pentaho was used to reduce the development cost, even if it employs IT experts to manage it.

The results are promising and demonstrate four basic dimensions (faculty, year, entrant type, and tuition fee level) and three data regarding applicants, tuition fee level, and payment status. Seven tuition-fee-level data were analyzed and visualized in charts and tuition-fee-level dashboard. The visualizations are realized via user interaction in the DW system. In addition, the visualization of the results is fast, accurate, and easy to understand.

The remainder of this paper is organized as follows. Section 2 describes work related to DSSs in HEI admission systems and the IAF policy in Indonesia. Section 3 describes the study methodology. Section 4 presents the development of the proposed DW system. Section 5 provides the system output and Section 6 prepares the discussion of multidimensional analysis of tuition fee level. Section 7 present result of this study, Section 8 presents conclusions, and Section 9 suggestions for future work.

II. Related Work

A. DSS in HE Admission System

DSSs for admission or enrollment systems in HEIs have been studied in recent years. As in [17], the knapsack problem approach was introduced to optimize admission exercise in Nigerian institutions. Reference [18] points to the trends in 21st century education and trends in transferring student's curriculum of Thai students in HEI. In [19], the admission process discussed as an academic business process in the SADIA System of a Portuguese university. Reference [20] described a web-based DSS application to improve the efficiency of admission to universities in Saudi Arabia.

Author in [21] studied an intelligent DSS for developing student admission policies based on an enterprise resource planning system. As in [22], a DW for the marketing process in Indonesia's HEIs has been discussed relative to support management in marketing decision making. Here, the primary marketing process was identified to analyze needs in a private university. In addition, DW models and data mining techniques were employed to design a higher education star scheme for analytic tools in 19 subsystems [23].

We found that the available current works do not address the admission system to manage the applicants' classification following the family background paying the student tuition fees.

B. IAF Policy for HEI in Indonesia

Under the Law on Higher Education 12/2012, higher education must set reasonable fees according to the financial qualifications of the students, their parents, or guardians. The IAF (or *Uang Kuliah Tunggal*) is the current admission policy for undergraduate students in public HEIs in Indonesia. Here, integrated means the students pay a fixed amount for education expenses each semester rather than several unit costs, e.g., development cost, number of credit units, and laboratory costs. The fixed tuition fee is derived from the calculation of all education costs in a year for a given study program. An annual regulation of the Ministry of Research and The Higher Education (MRTHE) is issued for the exact number of tuition fee levels at each HEI across the country. In addition to this regulation, the amount of money following on each tuition fee level for a given study program at each HEI was also provided.

The IAF policy determines tuition fees for only two regulars of all three programs in undergraduate entrant type: SNMPTN, and SBMPTN. SNMPTN is national selection based on high school academic reports, national exam scores, and another academic achievement, e.g., finalist at reputable science/sports/arts competitions, to rank qualified applicants relative to their HEI and study program of choice. SBMPTN is a second round of entrance by examination. Applicants who pass the tests of these regular entrance types must undertake a decision system to get their tuition fee level. The final round of HEI entrance is an independent program, which is automatically set to the highest tuition fee level for each study program.

The IAF policy targets 122 HEIs and 6725 study programs [2]. The IAF policy also covers public HEIs managed by the

MRTHE and other ministries, e.g., 58 HEIs are managed by the Ministry of Religious Affairs. The number of designated HEIs typically increases each year following government regulation to take over private HEIs and by extending the scope of HEIs under other ministries. After six years of IAP policy implementation, the national gross enrollment ratio (GER) increased from 29.15% (2014) to 34.58% (2018) [2]. The GER is a measure that compares the number of undergraduate students (diploma and bachelor) with the population aged 19–23 years.

III. Proposed Method

To determine what dimensional data could be obtained from tuition-fee-level data, we examined two approaches, i.e., the three-tier DW architecture [24] and the five-step DW implementation in education [9]. The three-tier approach is the broadest DW conceptual architecture is encapsulated for a development environment and divided into three levels or layers, i.e., bottom, middle, and top tier. We then associated the three-tier approach to the practical method of DW implementation in education. In [9] surveyed data warehousing in education and found five steps as a best practice implementation method are as follows: (1) information needs analysis and requirements analysis, (2) data source and data supply analysis, (3) DW design and multidimensional modeling, (4) extract, transform, and load (ETL) processes, and (5) system, application, reporting, dashboard, and online analytical processing (OLAP) development.

In this study, we formulated four procedures, i.e., preparation, integration, analytics, and visualization. The main motivation for choosing the proposed method is to merge the longer step of best practice and modest conceptual DW architecture, then following the software tools capability. A designated software utilized by a software suite for each procedure, except in the preparation procedure because it is such kind of examination. Fig. 1 shows our method used and how it correlates with three-tier DW architecture and five-step DW implementation in education.

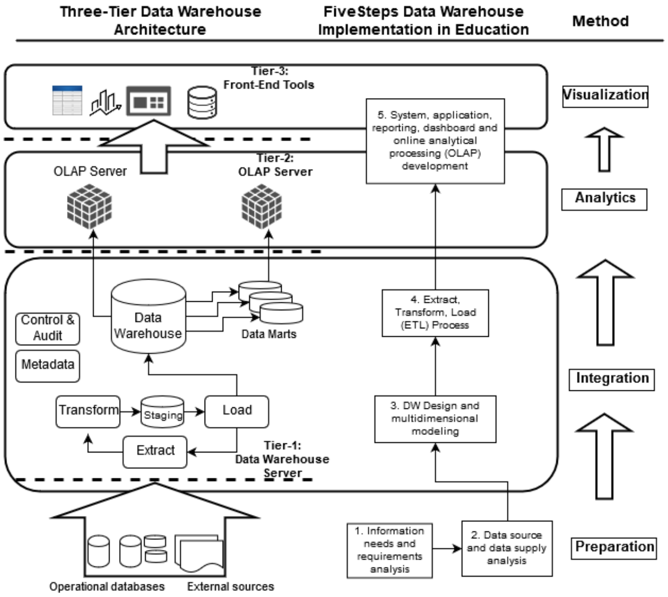

Three-Tier Data Warehouse Architecture
FiveSteps Data Warehouse Implementation in Education
Method
Tier-3:
Front-End Tools
Visualization
5. System, application, reporting, dashboard and online analytical processing (OLAP) development
OLAP Server
Tier-2:
OLAP Server
Analytics
Control & Audit
Metadata
Data Warehouse
Data Marts
4. Extract, Transform, Load (ETL) Process
Transform
Staging
Load
Extract
Tier-1:
Data Warehouse Server
3. DW Design and multidimensional modeling
Integration
Operational databases
External sources
1. Information needs and requirements analysis
2. Data source and data supply analysis
Preparation

Fig. 1. Proposed Method.

A. Preparation

The preparation procedures capture the first of two steps in the best practice DW implementation in education. Information requirements, sources analysis, DW system stakeholder identification, and goalsetting for each decision can be obtained by field observations, document checking, and regulation. The data-driven approach gives a realistic view of IT experts relative to determining the dimensions of analytics by analyzing data sources (operational databases or external sources) that must be incorporated into a single data repository.

B. Integration

The integration procedure occupies bottom tier of the three-tier DW architecture, which has access to the databases or data storage systems. The integration procedure also involved Steps 3 and 4 of DW implementation in education to design a multidimensional model and ETL processes. Existing data calls extracted from sources are transformed in the staging area and loaded into the DW. As a logical design of the DW, the dimension and fact tables are designed using a star or fact constellation schema. This model allows the DW system to observe the data in n-dimensional aspects.

C. Analytics

In the analytics procedure, an online analytical processing (OLAP) server functions in the middle tier of the DW architecture. OLAP is a common approach to analyzing and differentiating multidimensional data [8]. The aggregation of data is conceptualized in cubes by assigning which dimension tables apply to what fact table and how the fact table is measured. Multidimensional of analytics provided in the drilling-down or drilling-across ways is obtained by querying fact tables. Relative to the five steps of the DW implementation method in education, the analytics procedure corresponds to a part of Step 5 in OLAP development. The OLAP development satisfies the need for user visualization of reports, graphics, and a dashboard.

D. Visualization

Analytics data are automatically provided in many reports and graphics in dashboard. It is clear that the visualization procedure matches the top tier of the DW architecture as frontend tools. This procedure shares the same activities as part of step 5 of the DW implementation method in education on building system, application, reporting, and dashboard. The visualization data are displayed based on user privilege in the tuition-fee-level support system. Here, several actions are provided to further process the visualized data, e.g., conversion to Excel files or saving as images.

IV. System Development

A. Preparation

1) Data and Tools Specifications

The data used for the proposed DW system are for an Indonesian public HEI located in Sumatera Island. As discussed in Section 2.B, the IAF policy is complex and is applied to all public HEI in Indonesia. This HEI has experienced changes in tuition fee levels (five to seven levels) because the MRHTE regulated it five years ago. Data were taken from 15 faculties and 51 study programs over five years of policy implementation.

Furthermore, the data type was categorized as structured data because they data are highly organized and fit in fields and columns. In this study, we considered the software, which offers many toolsets and components that accommodate the four procedures discussed in Section 3. Here, we utilized Pentaho BI Suite Community Edition (CE), which has features in data integration, reporting, OLAP pivot table, and dashboarding [25]. Gartner places Pentaho in the visionary quadrant due to its mature data access, deep data transformation (provided by Pentaho Data Integration [PDI]), and advanced analytic capabilities (through the Data Science Pack). As in [26], Pentaho can integrate structured data from enterprise DWs with unstructured data from social media or IoT sources.

2) Information Requirements and Sources Analysis

When investigating the relevant regulations of the IAF policy, we focused on the probability of attributes changing. For example, the number of faculty (comprising numerous hierarchical study programs), the study program identifier, the number of tuition fee levels, and entrant types have a high possibility to change.

In the stakeholder analysis, we defined three groups engaged in tuition-fee-level management, i.e., HEI administrators, IT staff, and financial staff. The HEI administrators comprise administrators at the university, faculty, and study program levels concerned about tuition fee data of applicants by level and admission type. Note that IT staff engage in all operations of DW management, and the primarily interest of financial staff is an applicant's payment status.

After implementation of the IAF policy, the HEIs' financial decisions are determined based on such a multidimensional model. The primary problems targeted by this study occur at the institution management level (between the university and faculty levels). The following summarizes several example decisions.

- How should tuition fee levels be distributed among faculty?
- Which faculty obtains the highest number of students in each tuition fee level?
- Percentage of each tuition fee level in university or certain faculty.
- Trends of fees collected from students.
- Which one has a significant portion among high-level groups of tuition fee levels (level 5–7) and low-level groups (level 1–4)?

For tuition-fee-level management, the source databases are UKT (*Uang Kuliah Tunggal* – IAF), SIREG (Student Registration), and SIA (Academic/Didactics System). Database selection depends on the information requirements and target decision. Table I shows a data source analysis with correlation to information needs, decision category, and loading frequency.

TABLE I. DATA SOURCE ANALYSIS

Decision Category	Information needs	Data Source	Loading frequency
Accepted applicant data	Applicant distribution among faculty	UKT, SIREG	Twice per year
Tuition Fee Level	Tuition Fee Classification among faculties	UKT, SIREG, SIA	Twice per year
Payment	Tuition Fee Level Distribution of successful payment	SIREG, SIA	Daily in designated period

The data sources for retrieving accepted applicant analysis are integrated from the UKT and SIREG databases. This analysis is performed twice per year following the regular undergraduate admission program, i.e., SNMPTN and SBMPTN. With the accepted applicant analysis, the level of tuition fee analysis can be performed. However, all three databases (i.e., UKT, SIREG, and SIA) must be incorporated to analyze tuition fee level. The SIREG and SIA databases are involved in student payment analysis because SIREG database records the payment with detailed data for each tuition fee level and the SIA database stores the data of student registration in certain semesters. Completing or canceling tuition fee payment indicates the status of student registration in certain semester. Note that student payment analysis is performed daily during the payment period, and the data are recorded in the SIREG database.

After examining all data sources, we decided to integrate all data sources in consideration of tuition-fee-level management and enterprise-scale education analysis, e.g., course systems and student performance.

3) Improvement System

Data-driven approaches to decision making improve information requirement analysis. The preparation procedure is initiated by understanding the relevant regulation, reviewing the HEI business strategy, and gathering experience from IT experts engaged in the implementation of this policy. In this study, we acted as IT experts with knowledge about the workflow for the tuition-fee-level decision system. Thus, as explained in [16], we obtained the benefits of a data-driven approach because we had full access to current database structures.

The goal of this study is identical to BI solutions. The proposed DW system provides not only the integration of multiple information systems into a single repository but also data analysis to facilitate better-informed decision making to achieve an institution's goals. Moreover, the analytics is displayed as a simple graphical user interface that is easy to understand. All features in the proposed DW system are deployed using single open-source software suites that have both technical and financial advantages.

4) System Configuration

The system configuration is shown in Fig. 2. The PDI tool is used to build ETL function from all MySQL-based data sources to a PostgreSQL-based DW. First, data are fed into staging area and then loaded onto warehouse in the form of dimension table and fact table. In the application server, the Mondrian analytics engine uses OLAP schema and *Multidimensional Expressions* (MDX) query to handle requests from client that is performed using a tool called *Pentaho Schema Workbench* (PSW). Pentaho *Business Analytics* (BA) Server, a web container that interacts with Java servlets, responds to all requests from the client that accessed via a web browser. On the top of Pentaho BA server, Saiku and CTools, a set of community-driven plugins are installed to create dashboard, chart, and graphics. Note that users only access the system using a web browser.

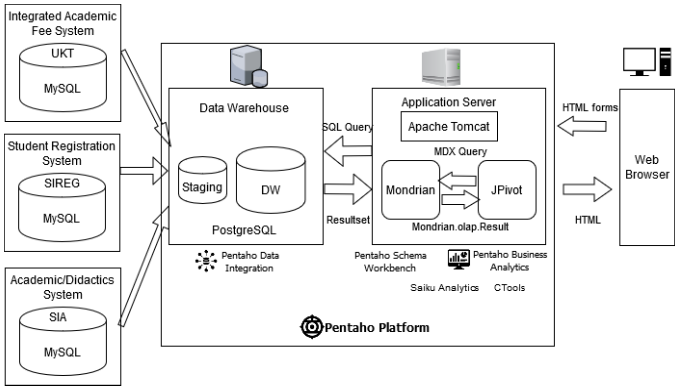


Fig. 2. System Configuration Diagram.

B. *Integration*

1) DW Schema

The DW system is designed to support all stakeholders in their roles. As a decision support instrument, the DW system is configured according to a top-down approach according to user information requirements. The core of the DW technology is a dimensional design comprising fact and dimensional data [27]. Fact data represent a set of business measurements to analyze, e.g., tuition-fee-level distribution and tuition fee percentage. In contrast, dimensional data represent the context descriptors of the measurements.

We define the basic dimensions used in the tuition fee level DW system as follows.

(1) *Faculty*. This dimension is the structure representing the level of management in HEI. This dimension is organized into a hierarchy of three levels, i.e., university, faculty, and study program. Each level is permitted to aggregate data at a desired level of abstraction. The attributes of the dimension are surrogate key ID (system-generated identifier to distinguish the dimension), study program ID, name of program study, faculty ID, and faculty name. In this dimension table, the faculty ID column is related via parent-child relation with the study program ID column, where the same faculty ID could has several study program IDs.

(2) *PreRegistrationPeriod*. This dimension is the structure for the period of the HEI's entrance type. This period occurs twice per year as two types of HEI's entrant that must determine the tuition fee level. The attributes of this dimension are the surrogate key ID, period ID (existing primary key from source data), year, name (description of the context), and entrant type.

(3) *GroupEntranceType.* This dimension is the structure of the HEI entrant type. As discussed in Section 2.B, only two types of admission (i.e., SNMPTN and SBMPTN) must participate in determining the tuition fee level.

(4) *TuitionFeeClass.* This dimension structures the level of tuition fees in reference to government regulation for each institution. For the case examined in this study, there were data for only seven levels of tuition fee during the five-year implementation of the IAF policy.

These basic dimensions can be used by any cube to define measurements and data aggregation. These dimensions are essential elements in the solution to the defined problem. Note that other supporting dimensions were designed, i.e., *Applicant*, *Date*, and *PaymentStatus* dimensions.

The information requirements shown in Table I have different processes and occur independently. Here, the fact data include *Applicant* data, *Tuition-fee-level* data, and *Payment* data. A denormalized facts constellation is used to relate dimension tables and multiple fact tables. In [24] explained that the facts constellation (also referred to as the galaxy schema) can serve multiple processes and has several shared dimensions. The schema applied in this study is illustrated in Fig. 3.

In this schema, the *Faculty* dimension is shared across all three fact tables, and the *Tuition-fee-level* fact table has six dimensions and two measures, i.e., *paymentAmount* and distinct count of *fk_applicant* derived from the *Applicant* dimension table. Note that all fact tables were designed as fact-less fact tables that tracked the tuition fee level of each applicant because each applicant has only a single tuition fee level. Reference [28] described many activities in educational institution admissions as the condition of events probability that might happen; thus, we employed the fact-less fact table design. The fact constellation schema comprises multiple star schemas; thus, a denormalized table was formed. A primary objective of the dimensional model is simplicity relative to reducing the number of tables and reducing disk consumption. The denormalized facts constellation schema was designed to optimize query efficiency and improve the DW processing speed.

2) Functionality in Integration Process

The integration procedure employs the PDI tool, which includes several functions required to construct the DW system.

a) Database Connection Pool Management

The connection to another database is critical for extracting data from sources and loading data to the target DW. Note that many types of databases with many access types can be utilized in connection pool management. In this study, as the Java-based technology, PDI required a Java Database Connectivity (JDBC) for connection to the MySQL-based data sources and PostgreSQL-based target DW.

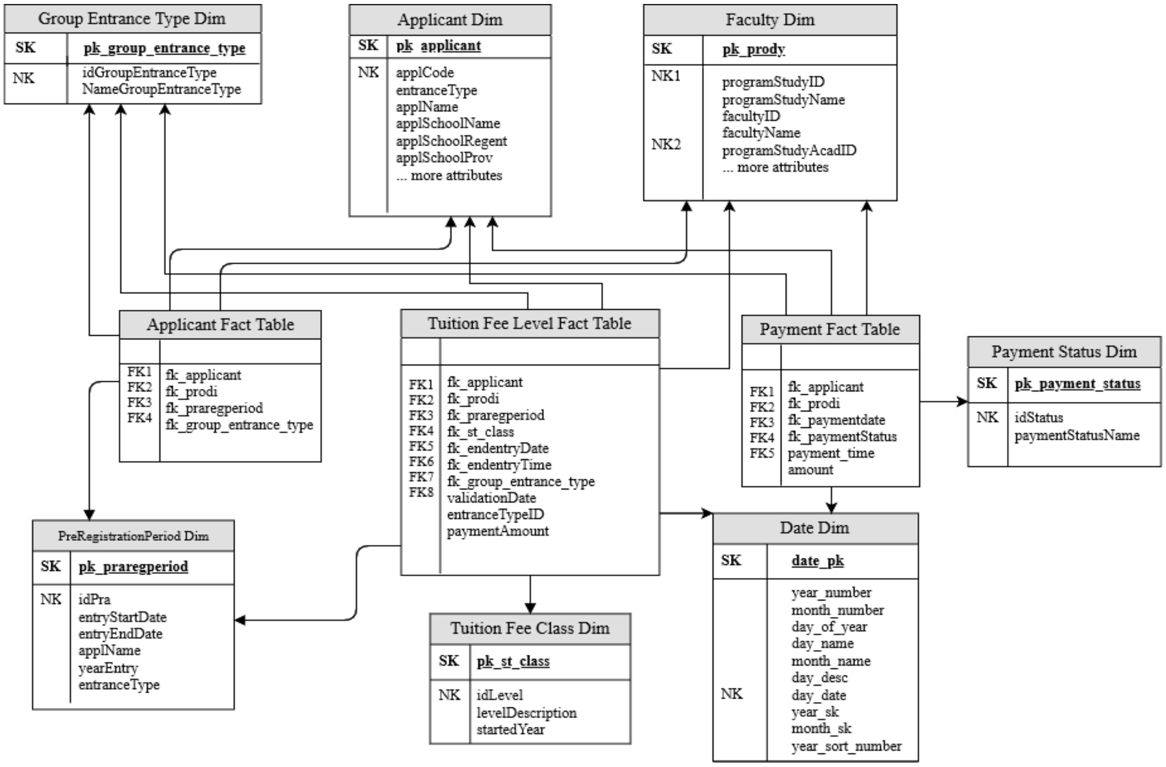


Fig. 3. Denormalized Multidimensional Facts Constellation Schema.

b) ETL Processes

Based on the PDI perspective, several ETL transformation files and a single job file for unifying processes were created. Transformations describe the ETL data flow, e.g., source connection, transforming data, and loading data into the target location. Jobs are used to coordinate ETL activities, e.g., flow definition, dependencies, and query execution preparation [29]. In the proposed DW system, ETL transformation is divided into four processes, i.e., loading data source to the staging area, performing dimension table creation, pre-fact table creation, and fact table creation. Table II shows the results of the ETL processes.

c) Automating ETL Processes

ETL jobs and transformation processes can be scheduled to run automatically at specific times. The Pentaho CE only provides a scheduler method by scripting an executor file using *cron* on a Linux-based server and a task scheduler or *at command* on a Windows-based server. This method needs to call the PDI command as the executer. In this study, the proposed DW system ran on a Windows-based server; therefore, a task scheduler application service was employed.

TABLE II. RESULTS OF ETL PROCESSES

Phase	Type of file	No. of files
Loading data source to the staging area	Transformation	11
Dimension table creation from staging area	Transformation	8
Pre-fact table creation	Transformation	3
Fact table creation	Transformation	3

C. Analysis of Tuition Fee Level

In a DW, the data analysis techniques refer to OLAP. OLAP is represented as a cube that stores a summary of corresponding dimension values in multidimensional space. Author in [30] described the data cube can be indexed in various ways, e.g., roll-up, drill down, slice, dice, and pivot. These OLAP operations are efficient ways to access the data cube for multidimensional analysis.

The OLAP cube of tuition-fee-level data is shown in Fig. 4. As a measurement, the number of applicants in the year 2017 is displayed and surrounded by a set of dimensions, i.e., the *Faculty*, *PreRegistrationPeriod*, *GroupEntranceType*, and *Tuition-FeeClass* dimensions. This multidimensional structure stores and distinct intersection values for the tuition fee level.

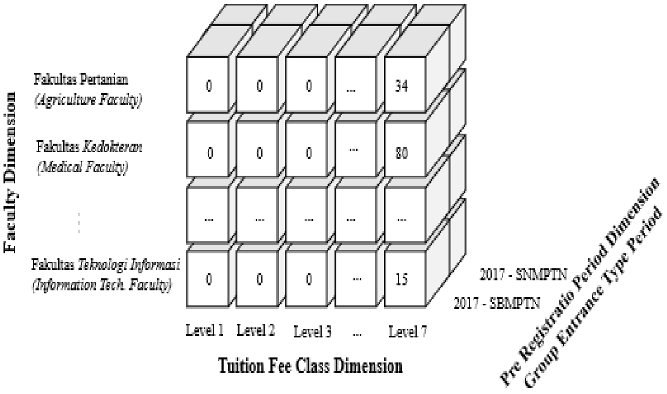


Fig. 4. OLAP Cube of Tuition Fee Level.

TABLE III. PIVOTED TUITION-FEE-LEVEL DATA

	PreRegistrationPeriod	2017	
	Group Entrant Type	*SBMPTN*	*SNMPTN*
Faculty	***Level Tuition Fee***	***No. of Participant***	
Agriculture Faculty	Level 1	0	0
	Level 2	0	0
	Level 3	2	3
	…		
	Level 7	34	19
Medical Faculty	Level 1	0	0
	Level 2	0	0
	Level 3	1	0
	…		
	Level 7	80	54
…			
Information Tech. Faculty	Level 1	0	0
	Level 2	0	0
	Level 3	0	0
	…		
	Level 7	6	15

OLAP operations can be applied to view data from different perspectives. For example, roll-up of the year in the *PreRegistrationPeriod* dimension is performed to aggregate or generalize year without counting entrant type data. The drill-down operation shows deep and smaller parts of the dimension, e.g., showing the number of applicants in the industrial engineering program as the lower level of the Engineering Faculty hierarchy in the *Faculty* dimension. The slice operation selects a single dimension, e.g., showing only Level 7 of the *TuitionFeeClass* dimension. An example of the dice operation is the selection of two or more dimensions as a filter/examination of tuition fee data for the Agriculture faculty by entrant type of SBMPTN in the year 2017. In addition, the pivot operation allows us to rotate the data axes of tuition fee level (Table III). With OLAP, analysis can be performed quickly because the data can be pre-calculated and pre-aggregated.

Another analysis can be performed using the MDX language as a written query language that is appropriate for multidimensional databases. To show the rank of faculty with the highest number of applicants in a particular year, the MDX query uses the *Topcount* syntax, and then shows the measurement data and faculty data in columns and rows. As a result, the percentage and trends of tuition fee levels can be analyzed.

D. Visualization

The visualization procedure is available in a web-based client application. As shown in the system configuration diagram (Fig. 2), users can access the DW system using a web browser in HTML format. This allows an easy access to the DW for HEI administrator (university or faculty level) and financial staff. The user logins will show the user console and load individualized result applicable to stakeholder's information requirements.

1) Charts and Graphics

The result of tuition-fee-level analysis is illustrated using many different types of charts. These charts are immediately loaded by querying the script that was deployed from many functionalities of the OLAP cubes. Bar chart, line chart, and pie chart are chosen to represent the report and analysis. The bar chart is usually designed to represent percentages, totals, and count. The graphic of tuition fee by faculty, the top five faculty by applicant, and the trend of fees collected from students are presented in the bar chart. The line chart is used to show the tuition fee level trend by year. Another type of graphic, e.g., stacked area chart, *heatgrid* chart, and metric dot chart, can be chosen with relevance to the data for display.

2) Dashboard

The dashboard represents a user interface for the DW system. The dashboard operates as a graphic container that displays analysis data in a single view. Many analytics and charts can be displayed together on the dashboard as shown in Fig. 5. The dashboard enables convenient multidimensional identification of tuition-fee-level analytics for users.

3) Dashboard Functionality

Dashboard functionality and interactivity are selected as the features of the dashboard. As in [31], the filter or parameters, alert, drill down, and data conversion are included in these features. A parameter is assigned a value from the attributes of the dimensional data to narrow the search as well as to filter and classify the fact data [32]. In the dashboard of the tuition fee level, two parameters are used, i.e., year and entrant type. An alert delivers a quick note monitoring a single event within the dashboard. The design of these charts is also designed to be exported to Excel or PDF formats for further action.

V. SYSTEM OUTPUT

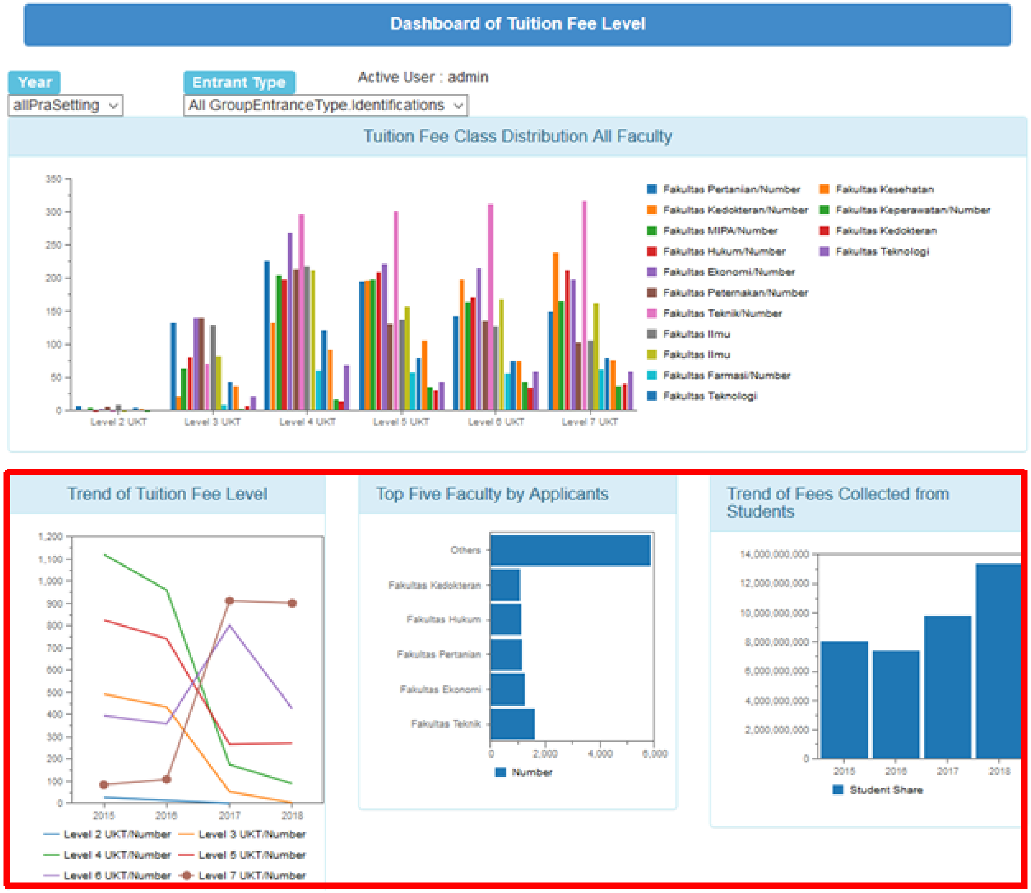
The detailed analytical results are presented in three graphs outlined in red in Fig. 5. The analysis of tuition fees by faculty provides insightful and valid data for HEI administrators.

A. Tuition Fee Level Trends

Tuition Fee Level Trends illustrates the trend in the level of tuition fees from 2015 to 2018. As can be seen, Level 7 increased significantly in the last two years. Level 6 fluctuated over the entire period and declined significantly in 2018 compared to the level in 2017. In this period, Levels 3, 4, and 5 decreased every year, with Level 4 showing the most significant reduction.

B. Top Five Faculty by Applicants

The top five faculties by applicants are presented in bottom middle of Fig. 5. As can be seen, the Engineering Faculty (*Fakultas Teknik*) attracted the greatest number of applicants, i.e., approximately 1600 applicants, over the last five years. The Economics Faculty (*Fakultas Ekonomi*) ranked second with approximately 1250 applicants, followed by Agriculture (*Pertanian*), Law (*Hukum*), and Medical (*Kedokteran*) faculties. The others bar represents an aggregate number of applicants from 10 other faculties.



Dashboard of Tuition Fee Level

Year: allPraSetting

Entrant Type: All GroupEntranceType.Identifications

Active User : admin

Fig. 5. Dashboard for Tuition Fee Level Management.

C. *Fees Collected From The Student Trends*

Fees collected from the student trends provides information about the funds collected from the students in HEIs for the years 2016–2018. Over this period, the lowest amount was collected in 2016, i.e., approximately 7 billion Indonesian Rupiah (IDR). The student share increased steadily over the next two years, reaching approximately 13 billion IDR in 2018.

VI. Discussion

The proposed DW system to manage the disparity of tuition fees and decision making based on tuition fee data was evaluated through a systematic analysis of three-tier DW architecture and a five-step method for an educational DW project implementation. In this study, we simplified the five-step method into four procedures, preparation, integration, analytics, and visualization. Based on a case study of a public HEI in Indonesia, the results indicate that DW technology makes multidimensional analysis of tuition fees possible. Multiple queries representing different data perspectives were processed against the same unified data repository. A visual representation of the level of tuition fees among all faculties allows university administrators to better understand the economic characteristics of the applicants to each faculty. The top faculties by applicants reveal the extent to which particular faculties are able to attract applicants and the extent of their financial contribution to the institution (Section 5.B). The trend of student fees over a period of four years was shown in Section 5.C. Financially, HEIs want to obtain higher levels of tuition fees for their financial stability on educational cost.

The configuration of the proposed DW system (Fig. 2) appears to ensure fast and accurate analysis results. First, dividing database functionality into the source (operational database) and the target (staging area and DW) resulted in stable performance. The multidimensional queries were only connected to the DW, were formulated at different aggregation levels, and executed automatically. The system of transactional processes worked for this DW system merely at the designated schedule (Table I). Second, the designed to the hierarchy of study program, faculty, and university level in Faculty dimension. The use of a hierarchical dimension in the DW enabled data to be measured at the desired level.

VII. Result

The analysis resulted in many charts that can be assembled on a dashboard. This feature should enhance a manager's ability to process information and act [33]. The dashboard for tuition-fee-level management (Fig. 5) displays different analyses in various charts and provides insight into tuition-fee-level performance relative to a particular faculties' revenue target. When analysis results reveal positive trends in the previous year for both student share and the number of applicants, administrators should make decisions designed to sustain such trends. Furthermore, when the analysis indicates that many faculties have attracted comparatively fewer applicants, the administration can take various actions, e.g., develop an advertising campaign or special promotion to attract potential candidates.

An improved data-driven decision approach is expected to enhance the development of the proposed DW system. Determining business requirements and performance indicators requires the engagement of IT experts who understand the workflow of tuition-fee-level decisions. IT experts are also required to develop DW functions using the open-source BI tools. In our opinion, the Pentaho CE utilized in this study has a sufficiently comprehensive toolkit as the open-source BI Suite. HEIs must provide an IT expert to manage this software; however, as noted in [15], we believe that this investment in human resources is essential to effectively implement BI tools.

VIII. Conclusion

The multidimensional tuition fee management presented in this paper is part of the admission DSS. For the case in Indonesia, the disparity of tuition fee level, independence of current HEI system, and time consumption in providing reports causes the lack of student-based income data credibility that affects the sustainability of HEIs. The development of DW system offers a way out for having a single source of truth by integrating multisource data following the business requirements of HEI's administrator.

As a BI approach, the DW system supports the aggregation of information at desired levels required by users. A predefined OLAP analysis improves the processing speed that enables safe operational database when retrieving electronic historical data. Analysis results were presented in various charts, graphic, and dashboard of tuition fee level, which has many functions to provide insight relative to the business performance.

IX. Future Work

The DW system described in this paper can be used as a guideline for tuition-fee-level management for HEIs in Indonesia. The government of Indonesia has a strategy to increase the number of public HEIs, and they apply the IAF policy to all public HEIs under all ministries. The proposed DW system does not require an unreasonable amount of effort to implement.

Future research should involve monitoring HEI capacity, the actual paid tuition fees to capture the gap between paid and unpaid payment status, and the gap between collected tuition fees by level and the standard education cost. This study should also be extended to other academic units, such as the teaching process, staff data, and research areas.

Acknowledgment

This work was supported by the Indonesia Endowment Fund of Education (LPDP), Ministry of Research, Technology and Higher Education, Ministry of Finance of the Republic of Indonesia awarded to A.A.Y, with reference No: PRJ-3678/LPDP.3/2016. We would also like to thank the reviewers for their insightful feedbacks on the paper.

References

[1] Organization for Economic Cooperation and Development. Education at a Glance 2019: OECD Indicators. Paris: OECD Publishing, 2019.

[2] Indonesian Ministry of Righer Education. Higher Educational Statistical Year Book 2018. Jakarta: Pusdatin Iptek Dikti, Setjen, Kemenristekdikti, 2018.

[3] M. N. Y. Utomo, A.E. Permanasari, E. Tungadi, and I. Syamsuddin, "Determining single tuition fee of higher education in Indonesia: a comparative analysis of data mining classification algorithms," 4th International Conference on New Media Studies, Yogyakarta, Indonesia, 2017.

[4] B. Karim, S. Sentinuwo, A. Sambul, "Decision of integrated academic fee for new entrant at Sam Ratulangi University using data mining" (Penentuan besaran uang kuliah tunggal untuk mahasiswa baru di Universitas Sam Ratulangi menggunakan data mining), E-Journal Teknik Informatika Vol 11, No.1, Sam Ratulangi University, Manado, Indonesia, 2017.

[5] Indonesian Ministry of Education and Culture Regulation (No 55. Integrated Academic Cost and Integrated Academic Fee for State High Education Institutions, 2013.

[6] M. Golfarelli and S. Rizzi, "From star schemas to big data: 20 + years of data warehouse research". In: Flesca S., Greco S., Masciari E., Saccà D. (eds) A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years. Studies in Big Data, vol 31. Springer, Cham. pp 93-107, 2017.

[7] V. Rainardi, Building a data warehouse with examples in SQL Server. Berkeley, CA: Apress, New York, 2008.

[8] R. Kimball and M. Ross, The Kimball Group Reader: Relentlessly practical tools for data warehousing and business intelligence. Indianapolis, IN: John Wiley & Sons, 2016.

[9] O. Moscoso-Zea, J. Paredes-Gualtor, and S. Luján-Mora, "A holistic view of data warehousing in education," IEEE Access Volume 6. pp. 64659-64673, 2018.

[10] D. Wierschem, J. McMillen, and R. McBroom, "What academia can gain from building a data warehouse," Educause Quarterly, Vol. 26 No. 1. pp 41-46, 2003.

[11] J. Guan, W. Nunez, and J. F. Welsh, "Institutional strategy and information support: the role of data warehousing in higher education" Campus-Wide Information Systems, Vol. 19 Issue: 5. pp.168-174, 2002.

[12] J. D. Porter and J. J. Rome, "Lessons from a successful data warehouse implementation," Cause/Effect (Winter) . pp. 43-50, 1995.

[13] F. D. Tria, E. Lefons, F. Tangorra, "Academic data warehouse design using a hybrid methodology," Computer Science and Information Systems, Vol. 12, No. 1, pp. 135–160, 2015.

[14] I.M. Aljawarneh, "Design of a data warehouse model for decision support at higher education: A case study," Information Development, Vol. 32, No. 5, pp. 1691–1706, 2016.

[15] B. Scholtz, A. Calitz, and R. Haupt, "A business intelligence framework for sustainability information management in higher education," International Journal of Sustainability in Higher Education, Vol. 19 No. 2, pp 266-290, 2018.

[16] D.J. Power, "Understanding data-driven decision support systems," Information Systems Management, 25:2, pp 149-154, 2007.

[17] H. Bello and A. M. Jingi ,"Admission Decision Support System for Nigerian Universities," International Journal of Computer Application, Vol 181, No.42, pp 35-44, 2019.

[18] L. loywatanawong ,"Decision Support Systems Model for Admission to Higher Education," 2nd IEEE International Conference on Computer and Communication, Chengdu, China, Oct 14-17, 2016.

[19] E. Cardoso, H. Galhardas, and R. Silva, "A decision support system for IST academic information," Informatica (Slovenia) Vol. 27 No. 3, pp 313-323, 2003.

[20] A. Alotaibi, A. Ayesh, and R. Hall, "Managing admission in Saudi universities: a system approach," International Journal of Information and Education Technology, Vol. 6, No. 4, pp 314-321, 2016.

[21] R. Vohra1 and N.N. Das, "Intelligent decision support systems for admission management in higher education institutes,“ International Journal of Artificial Intelligence And Applications, Vol.2, No.4, pp 63-70, 2011.

[22] Rudy and E. Miranda, "Management report for marketing in higher education based on data warehouse and data mining," International

Journal of Multimedia and Ubiquitous Engineering, Vol. 10, No. 4, pp. 291–302, 2015.

[23] Rudy, E. Miranda, and E. Suryani, “Implementation of datawarehouse, datamining and dashboard for higher education,” Journal of Theoretical and Applied Information Technology, Vol. 64, No. 3, pp. 710–717, 2014.

[24] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques (3rd. ed.). San Francisco, CA: Morgan Kaufmann Publishers Inc, 2012.

[25] R. Bouman and J.V. Dongen, Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL. Indianapolis, IN: Wiley Publishing, 2009.

[26] H. Yoshida; https://community.hitachivantara.com/s/article/bridging-the-gap-in-bimodal-it; Oct. 2, 2019.

[27] C. Adamson, Star Schema: The Complete Reference. New York: McGraw-Hill Company, 2010.

[28] R. Kimball and M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd edition. Indianapolis, IN: John Wiley & Sons Inc, 2013.

[29] Pentaho Doc; https://help.pentaho.com/Documentation/6.0/0J0/0C0/030; Oct. 2, 2019.

[30] S. Chaudhuri and U. Dayal, “An overview of data warehousing and OLAP technology. ACM SIGMOD Record, Vol 26 No.1, pp 65-74, 1997.

[31] N. Rasmussen, C.Y. Chen, and M. Bansal, Business Dashboards; A Visual Catalog for Design and Deployment. Hoboken, NJ: John Wiley & Sons, 2009.

[32] P. Rob and C. Coronel, Database System: Design, Implementation, and Management, 8th ed. Boston: Thomson Learning, 2009.

[33] W. Bremser and W.P. Wagner, “Developing dashboards for performance management. CPA Journal 83, pp 62-67, 2013.

Signature based Network Intrusion Detection System using Feature Selection on Android

Onyedeke obinna Cyril[1], Taoufik Elmissaoui[2], Okoronkwo M.C[,3], Ihedioha Uchechi .M[4], Chikodili H.Ugwuishiwu[5], Okwume .B. Onyebuchi[6]

Department of Computer Science, University of Kairouan, Tunisia[1]

Innov'Com , SUP'COM,University of Chartage & Higher Institute of Applied Mathematics and Computer Science, University of Kairouan, Tunisia[2]

Department of Computer Science, University of Nigeria, Nsukka (UNN), MNCS, MCPN[3]

Department of Computer Science, University of Nigeria, Nsukka (UNN)[4, 5]

Department of Computer Science, University of Nigeria, Nsukka (UNN)[6]

***Abstract*—This paper Smart Intrusion Detection System (IDS), is a contribution to efforts towards detecting intrusion and malicious activities on Android phone. The goal of this paper is to raise user's awareness of the high rate of intrusions or malicious activities on Android phones and to provide counter measure system for more secured operations. The proposed system (SIDS) detects any intrusion or illegal activities on android and also takes a selfie of the intruder unknown to him/her and keep in the log for the view of the user. The object oriented analysis and design method (OOADM), was adopted in the development. This approach was used to model and develop the system using real intrusion features and processes to detect intrusions more flexibly and efficiently. Signature detection was also used to detect attacks by looking for specific patterns. The system detects intrusions and immediately sends an alert to the user to notify of an illegal or malicious attempt and the location of the intruder.**

Keywords—Signature Detection; Feature Selection; android phone; Smart Intrusion Detection System (SIDS)

I. Introduction

Smartphones, tablets, and other mobile platforms are rapidly emerging as popular appliances with progressively amazing computing, networking, and detecting abilities. Smartphones are currently the overwhelming individualized computing devices with so many features, and strength comparable to mini computers. Some of the attractive features of these smartphones include calls, short messages, multimedia, email, video calling, voice dictation, eservices, file exchange, internet browsing, services, etc. According to Pew Research Centre in 2015, about 43% of the global population uses a smartphone device [1]. Also, there were 5.11 billion interesting portable clients worldwide in 2019, and 2.71 billion of them use smart phones, it evaluated that there will be 2.87 billion smartphone clients worldwide in 2020, and 2.5 billion dynamic Android gadgets around the world, this value was based on Google's Play store Statistics, and this implies that the number is higher. These numbers of Android devices and users additionally underscore the size of the fracture challenge and Google hopes to apply essential updates and security principles to all Android gadgets across various renditions, districts, and producers. Android was launched by Google and Open Handset Alliance in September 23, 2008. Android has experience a vast growth since its inception because of its user friendliness, open source, ease of developing and publishing applications.

The ubiquitous usage of Android OS has induced the burst of mobile application market. Google Play is the largest app store followed by Apples App store. According to [2], The Android Applications are available for download through Google Play Store and third party agents. Though intrusion is not specific to android phones; most smart devices are used for e-businesses; which expose both private and financial data to public domain. Several techniques have been proposed and implemented to detect, prevent and reduce malicious intrusions on smartphones.

Notwithstanding, intrusion is any unapproved action on a computer network. Much of the time, such undesirable action retains network assets expected for different utilizations, and about consistently compromises the security of the system as well as its information. Appropriately structuring and sending a system intrusion detection system will help obstruct the interlopers. Recognizing an intrusion relies upon the protectors having a clear understanding of how assaults work [3]. Intrusion activities seek to unsettle the confidentially, availability or integrity of a resource or the controlling applications. As a result of high prevalence, intrusion detection systems (IDS) are provided to checkmate intrusions. IDS is a sort of security measures use to alert the right owner of a device when a person or thing is attempting to bargain data framework through vindictive activities or through security approach encroachment. The Proposed system (SIDS) is focused at developing a model that will identify malicious intrusions on smart phones, through finger print and password validity and also takes a selfie of the intruder unknown to him/her and keep in the log for the view of the user.

The techniques used for detecting intrusion can be arranged into Signature based location and Anomaly based recognition. Signature based detection is termed as misuse detection which helps in the detection of attacks by looking for specific patterns. Here, the dataset has number of occasions and each data must be named as typical or malevolent. In [4], AI calculations are utilized to prepare the informational collection as indicated by their name, and abuse identification strategy is made naturally. Contingent upon the vigor and earnestness of a mark that is initiated inside the framework, alert reaction or warning is sent to the correct authorities. Anomaly detection strategy is intended to reveal the examples that are a long way from the ordinary and others are hailed as an interruption. Irregularity discovery is helpful for discovering assaults like abuse of convention and administration ports, DoS dependent on made payloads, DoS dependent on volume, cradle flood and other app payload inconsistency [5].

II. Android Malware Detection

Intrusion detection system (IDS) is an instrument for finding attempts to bargain a framework [6]. Possibly, such endeavors can be forestalled; in such case, the framework is called an interruption avoidance framework. Interruption recognition components applied in Android phones depend on indistinguishable standards from instruments utilized in different frameworks (for example PCs and computer networks). In spite of the reality the frameworks are distinctive in their sort and design; the establishments of assurance against assaults continue as before. This takes into consideration the appropriation of existing procedures and their use in the Android security zone. Interruption recognition frameworks can be arranged by the discovery approach and based on the sort of dissected information. Another characterization approach distinguishes the area of the IDS. [6].These classifications are described in this section below.

A. *Detection Approach*

Intrusion detection systems are ordered by the location approach utilized to distinguish meddling exercises [7]. The most generally discovery strategies are irregularity and abuse location.

Anomaly detection is intended to distinguish malevolent activities through recognizing deviations from an ordinary profile conduct. Despite the fact that this sort of IDSs performs better in distinguishing novel assaults, they ordinarily experience the ill effects of high FP rate. [4]. Signature recognition, is the place the location procedure depends on known marks or patterns, and plans to recognize authentic occurrences from the malignant ones. Without the downside of inconsistency detection, it is solid for recognizing known assaults with low FP rate. However, this sort of IDSs can't recognize obscure assaults or varieties of known ones [4].

B. *Android Architecture*

Android Stack is based on Linux kernel and it consists of four layers that manage the whole system starting from hardware sensors to the user's high-level apps. It consists of different layers running on one another, the lower ones offering types of assistance to the upper level layers [8]. This architecture explains the functions of each layers on android phones.

The first layer; the Linux Kernel is the most important represents the heart of Android system. It provides the OS services and manages the hardware's functions such as memory, power, drivers, network stack, security settings, shared libraries and hardware abstraction.

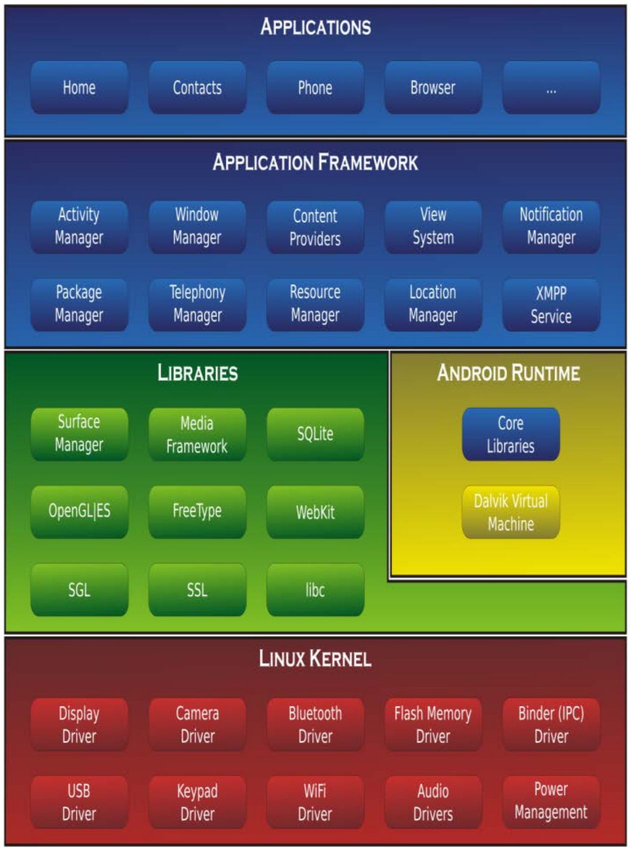


Fig 1. Android's Stack Structure (Adapted from [8]).

The second layer; the library, provides native libraries which are a set of instructions that manage data processing. It provides the open source libraries and the android runtime.

The third layer; the Application Framework, includes the Android APIs. The APIs are classes and interfaces for Android apps' development. This layer interacts with the running apps and manages the basic functions on the device.

The fourth layer; the Application provides the phone's functions to the end-user such as making calls, managing contacts, sending messages, and browsing web. Also this layer provides a set of core applications, such as email client, calendar, browser, maps, contacts, SMS program, gallery, etc. Fig. 1 illustrate the architecture of an android.

C. *Feature Selection*

The general methodologies for feature elites are classified into three: filter method, wrapper technique, and embedded technique, each component decisions to calculate and make effective use of any of the three element choice systems [9].

The point of feature selection is to discover a subset of the qualities from the first set that are more delegate for the information, and for which the inherent part in the subset are applicable to the expectation; it improves the forecast presentation of Artificial Intelligence models by wiping out noisy factors. It also provide less difficult models that gives better explanations of the complex random procedure, reduced expense of huge trial estimations and subset of factors that can be analyzed for causal induction [10].

D. The Filter Approach

This Approach survey the importance of the highlights from the dataset; the choice of the features depends on the measurements, the arrangement execution is utilized in wrapper techniques as a piece of the component subsets assessment and determination forms. As opposed to wrapper draws near, installed approaches are process of mathematical calculations less serious than wrappers since they consolidate a collaboration between feature determination and learning process. Albeit inserted approaches incorporate a regularized hazard capacity to upgrade the features assigning limit and indicator parameters, it is hard to roll out an improvement in the arrangement model to get better [11].

E. Review of Related Literature

This section, examines a portion of the past methodologies used by researchers for recognizing intrusions. Various approaches have been used to detect intrusions and they can be generally assembled into filter, wrapper and embedded. Below, we give a concise survey of research studies that have been conducted using these approaches.Filters approach don't depend on the classifier calculation, yet utilize other criteria dependent on relationship ideas [9], while Wrappers consider include subsets by the nature of the presentation on a demonstrating calculation, which is taken as a discovery evaluator. Implanted techniques perform include features during the demonstrating calculation's execution [9].The venture embraces the filter techniques for IDS. Because of the continuous development of information dimensionality, include determination as a pre-handling step is turning into a fundamentals part in structure intrusion detection frameworks.

In [12], proposed a novel stage free conduct based oddity discovery system for smartphone. It can distinguish vindictive exercises on smartphone progressively by utilizing solo AI methods called K-implies grouping. The procedure utilized is constrained in light of the fact that it depends on static examination of use consent and system calls.

A host based IDS model for advanced mobile phones and make evidence of idea app for android stage was proposed [13]. The framework arrangements depend upon customers' current system, diverse approach stage is included and discovery instrument is on higher alarm in broad daylight systems. The significant constraint is on client experience dangers, cost creating danger and protection encroaching dangers isn't comprehended.

According to [6], presents a novel AI based IDS to expand the exactness and proficiency of arrangement. The framework diminishes the preparation and testing time from 113.53 and 2.93 to 44.78 and 2.06 on the CIC – IDS 2017, it additionally accomplishes the most elevated F-proportions of 0.998 and least bogus alert rate and dispose of insignificant highlights.

In [9], develop a system that detects any illegal/malicious intrusions in android phones using filter based feature selection algorithm. It evaluates the dependence between features and output classes, also scan to ascertain between legal/illegal users through pin validity. However the authentication level is not strong enough using pin and it does not track the location of the user.

In [14], proposed the utilization of an orderly depiction plot for managing the portrayals used to portray IDS capacities. This methodology ought to take into account an assessment of IDSs dependent on their depictions, without requiring experimentation. The weakness of this methodology is the prerequisite of exact depictions. Right now, such a methodology doesn't exist so executing it is beyond the realm of imagination. This methodology holds a specific guarantee for what's to come.

According to [15], manages the importance of each component in KDD 99 intrusion recognition dataset to the discovery of each class. Their exact outcomes uncovered that a few features (hot Login, number of Compromised situations, number of record creation assignments, visitor login) have no pertinence in intrusion detection. Harsh set level of reliance and reliance proportion of each class were utilized to decide the most isolating features for each class.

In [16], proposed a novel method to deal with break down factually the system traffic unrefined information. The enormous measure of rough information of real system traffic from the IDS is investigated to decide whether traffic is an ordinary or hurtful one. The issue is currently transformed into the sensor system to build the exact recognition rate, on the grounds that no hunt spaces are diminished.

In [17], present the different structures of IDS, measures that help to characterize the level of adequacy of IDS and the continuous work of institutionalization and homogenization of IDS. The system enables us to update the analyzer to find conceivable new assaults or varieties of assaults. Their limitations don't guarantee 100% security, ridiculous and the disservice of this arrangement is the rate of FP because of strange or unordinary conduct of clients, who are not really hurtful.

In [18], proposed a framework so as to improve the security of the portable applications which will assess the versatile applications security dependent on the distributed computing stage and information mining. The assessment results shows that it is reasonable to use appropriated computing stage and information mining to confirm all put away applications routinely to filter through malware applications from versatile application markets. The weakness is the moving of the security usefulness into the cloud could likewise be perilous, if not all pieces of the phone can be imitated into the cloud.

In [19], evaluated data in regard to classifiers configuration, utilized dataset, feature extraction, clustering strategies, exactness location measures and so on. The work of numerous and cross breed classifiers, improves the precision of the grouping and encourages understanding troublesome issues. The shortcoming is that binomial or typical (measurable circulations) can't delineate example acknowledgment conduct, which implies that standard systems of parametric techniques may not work.

In [20], proposed another solid half breed technique for an oddity system based IDS utilizing artificial bee colony (ABC) and Adaptive Boosting calculations (ADA Boost) so as to pick up a high recognition rate with low FP rate. The exactness and identification rate of this technique has been improved in correlation with unbelievable strategies. The shortcoming is the bogus alert report of intrusion to the system and intrusion detection precision that occurs because of the high volume of system information.

According to [21] proposed a mutual data based calculation that logically chooses the ideal element for grouping. The evaluation results shows that the feature selection calculation contributes progressively basic features(Logs records, hot logins, number of compromised condition) for least square help vector machine based interruption discovery framework for a better precision and lower computational expense. The deficiency is that "huge information" thwarts the entire detection process and may prompt inadmissible grouping precision because of the computational challenges.

In [22], used both static and dynamic investigation to recognize malware in android applications. They consolidated the static investigation (consent) and dynamic examination (System call following) with AI. They performed static investigation by removing authorizations from the Android's manifest.xml record and analysed the complexity between the quantity of consents mentioned by favourable and vindictive applications. They understood that the quantity of authorizations mentioned by charitable and dangerous application is marginally the equivalent. This strategy was tried on different benevolent and threatening applications.

III. Description of the Existing System

Based on the literature reviewed, the previous work done on the existing system of IDS specifically those that use anomaly based approached is described as follows:

1) Most of the system of IDS authentication access is through pin and emails.

2) The system barely tracks the location of the phones.

3) They don't have a reliable accountability system (i.e. keeping a records of all activities carried out on the phone – like a shot of the intruder face unknown to him and send to as MMS to the user's phone and also kept on the app for record purpose.

4) One noteworthy issue of the current framework is the false alert that is brought about by ICMP (web control message convention). This is a mistake announcing convention arrange gadget like router/host use to create blunder messages and operational data showing that a mentioned administration isn't accessible or that a host/router couldn't be come to.

A. Analysis of the Proposed System

The proposed system seeks to address all the problems identified in the existing system by effectively detecting intrusions in Android phones. The following are the features of the proposed system:

1) The proposed system authentication access is through finger print and password.

2) The system has a GPS Tracker to help in the location of the phone.

3) The system has a feature that helps take a selfie of the intruders face during attempt on the phone and sent the intruders face to the MMS of the user other phone and also keeps all facial logs attempts for record purpose .

4) The problem of false alarm is avoided because the proposed system major alert agent is through SMS, MMS and not e-mail that requires ICMP.

5) The proposed system is design in a format that makes installation very simple and easy for the user thereby making navigation accessible.

The role of each actor representing the system flow and activities carried out:

1) Smart intrusion detection system (SIDS): This is the proposed application; its role is to detect, Filter and authenticate Intrusion.

2) User: He/she will download/install the app, configure settings and also check intruders' selfie records.

3) Sensor Agent: The agent audit, selfie of intruder, log and mail alert of an attempt to the user. Fig. 2 present a Use Case diagram of the proposed system (SIDS).

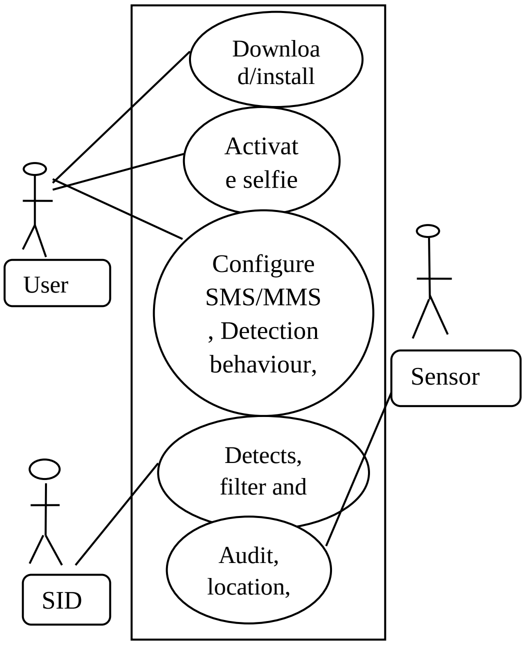


Fig 2. Use-Case Diagrams of the System.

B. System Architecture

The architectural design of the Proposed System (SIDS) is of 4 (four) tiers as shown in Fig. 3. SIDS was designed based on four layers that manage the activities on the system starting from;

Data Collection are sensors in charge of information accumulation and are in this manner the data wellsprings of IDS. This data is drawn from different sources, for example, enlisted information and log documents. **Data Pro-Processing** in this stage information gets changed or encoded to carry it to such an express, that the machine can without much of a stretch parse it and are processed to generate the basic features.

Attack Recognition, here the system compare information's in the dataset, after analysing the data it makes decision if it's a normal flow or an intrusion.

Result is the outcome that tells if an intrusion is recognized. It takes the information and contrast and the prepared dataset, and match on the off chance that the information is assaulted or typical, on the off chance that the information is assault, at that point an alarm will be sent to the phone number of the client (showing intrusion and location). Fig. 3 presents the structure of the proposed system (SIDS).

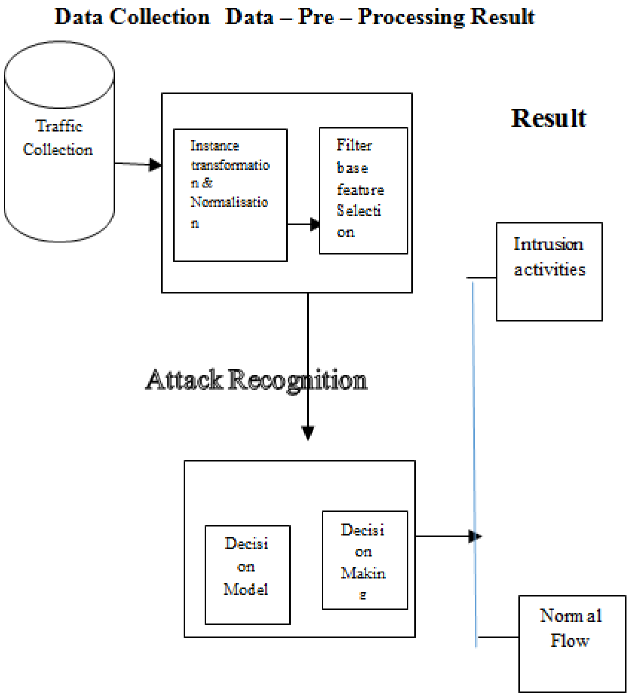


Fig 3. The Structure of the Proposed System (SIDS).

IV. System Implementation

The system implementation is the development of the new system or application following the laid plans from analysis and design stage. This chapter depicts how the plan from the previous section is executed with the aim of providing a proficient system to detection of intrusion on Android phone. Apparatuses and techniques used to actualize are presented in this section.

A. Choice of Development Environment

The integrated development environment (IDE) used in the development of this work is the Android studio 3.5.3, JRE 1.8.0_202-release-1483-b03 amd64, JVM: OpenJDK 64-Bit Server VM by JetBrains s.r.o on which the source codes are written, compiled and uploaded on Google Play Store. Android Studio offers numerous features that improves profitability when building Android applications, for example, Gradle-based system which is use to manage all dependencies (to build, test, run and package your app), Android Virtual Device (Emulator) also helps run and debug apps in the Android studio. The programming languages employed in this project are Java while Shared Preferences integrated database management system was used.

B. Implementation Architecture

The implementation architecture of the SIDS is represented in Fig. 4 below. It is made up of the various components of the software modules and their linkages. Fig. 4 illustrates the Implementation Architecture of the system.

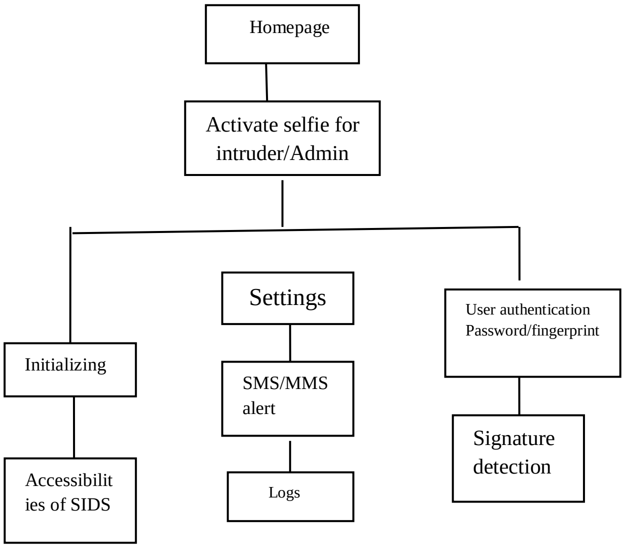


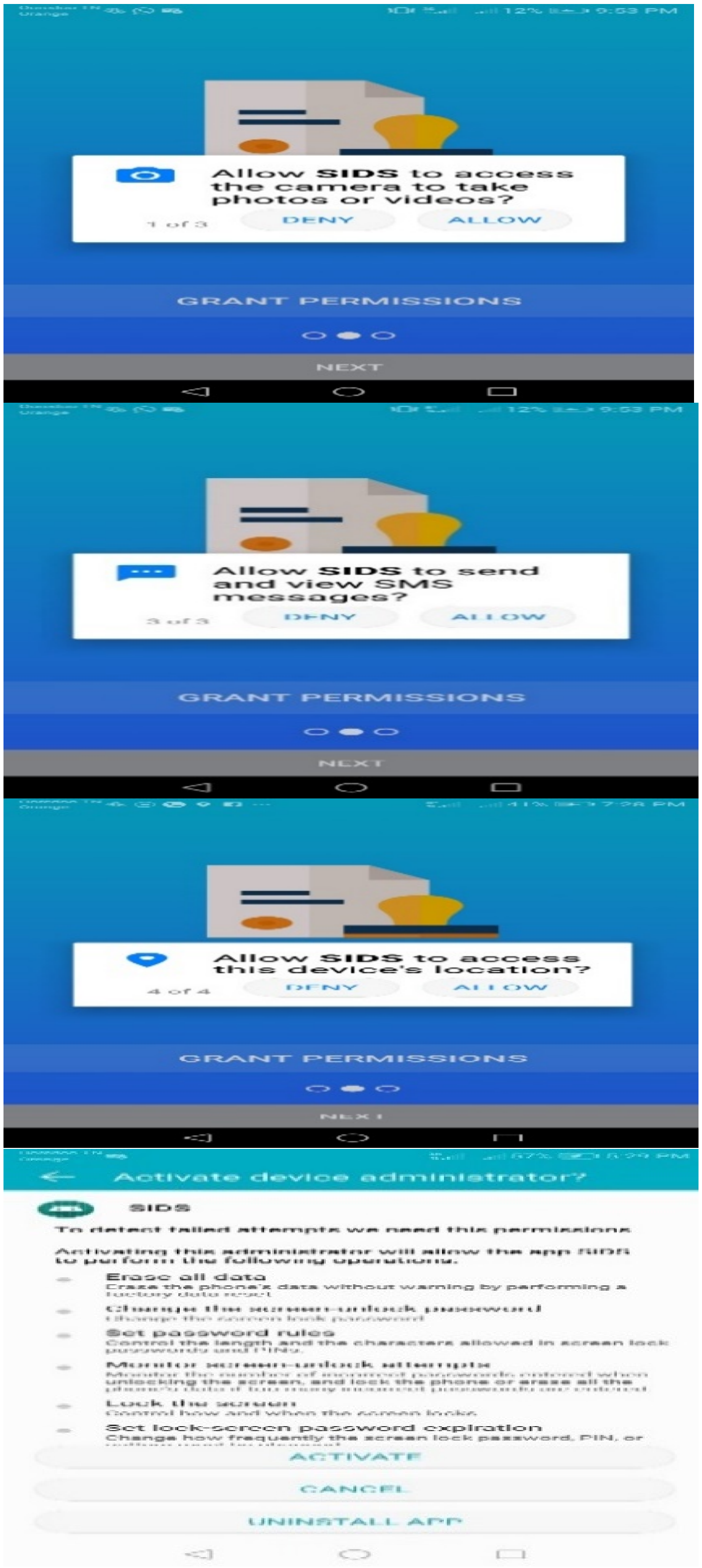
Fig 4. The Implementation Architecture.

V. Result and Discussions

Smart intrusion detection system (SIDS) is a mobile application developed using Java. After the application has been downloaded and installed, then activation of selfie for intruder and the Admin, also the user will configure SMS and location alert, number of attempts, SMS number. If an intrusion is detected, immediately the alert agent sends an SMS and MMS (that contains a statement indicating an intrusion and also the location of the phone), while a selfie of the intruder will be kept in the app log for the users view. The problem of false alarm is avoided because the proposed system major alert agent is through SMS and not only email that requires ICMP (which sends error messages to email indicating service is not available or not reachable).

Fig 5. Home Page of User Phone of SIDS.

The figure's below is an illustration of the output (result) displayed of an intrusion attempt. Fig. 5, 6, 7, 8, 9 and 10 presents the screenshots of the output for the proposed system (SIDS).



Allow **SIDS** to access the camera to take photos or videos?

1 of 3

Allow **SIDS** to send and view SMS messages?

3 of 3

Allow **SIDS** to access this device's location?

4 of 4

Activate device administrator?

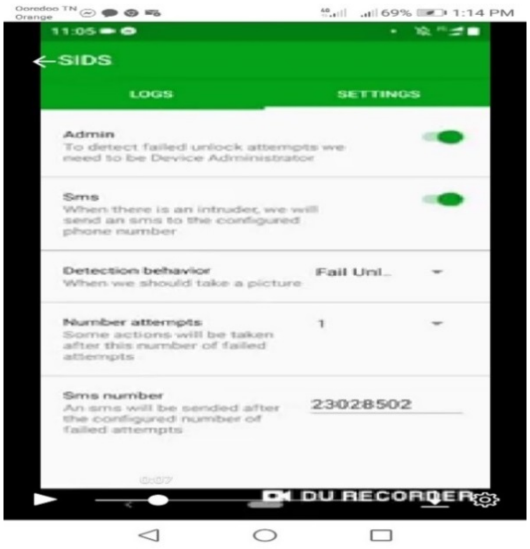
SIDS

To detect failed attempts we need this permissions

Activating this administrator will allow the app SIDS to perform the following operations:

- Erase all data
  Erase the phone's data without warning by performing a factory data reset
- Change the screen-unlock password
  Change the screen lock password
- Set password rules
  Control the length and the characters allowed in screen lock passwords and PINs.
- Monitor screen-unlock attempts
  Monitor the number of incorrect passwords entered when unlocking the screen, and lock the phone or erase all the phone's data if too many incorrect passwords are entered
- Lock the screen
  Control how and when the screen locks.
- Set lock-screen password expiration
  Change how frequently the screen lock password, PIN, or [illegible]

Fig 6. Activating Intruder Selfie and Admin of SIDS.



Admin
To detect failed unlock attempts we need to be Device Administrator

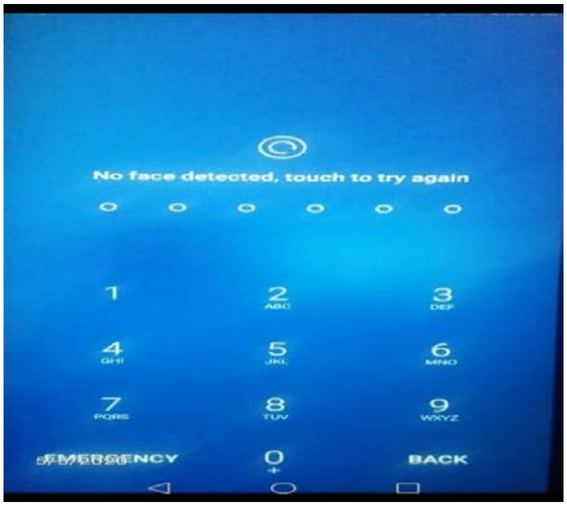
Sms
When there is an intruder, we will send an sms to the configured phone number

Detection behavior: Fail Unl...
When we should take a picture

Number attempts: 1
Some actions will be taken after this number of failed attempts

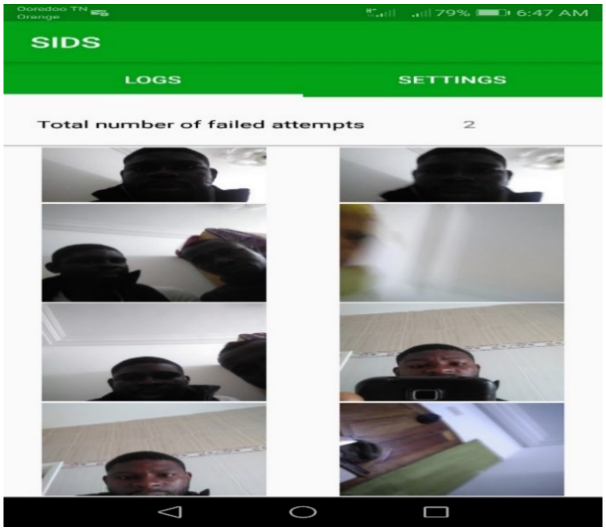
Sms number: 23028502
An sms will be sended after the configured number of failed attempts

Fig 7. User Configuration Settings of SIDS.



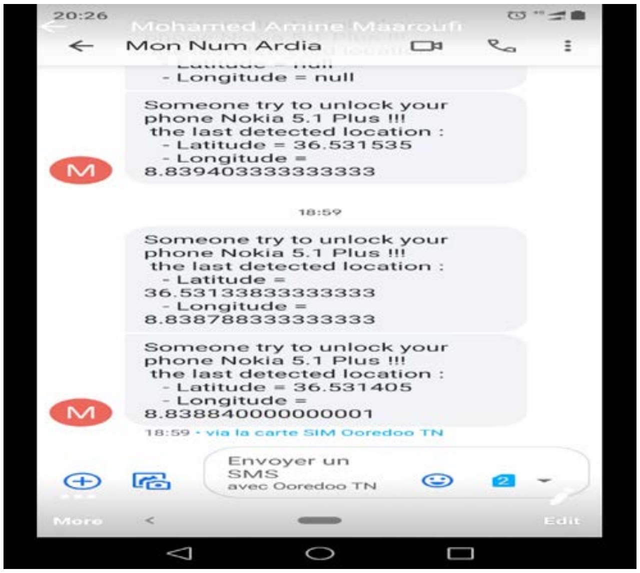
No face detected, touch to try again

Fig 8. User Login Password of SIDS.



Total number of failed attempts 2

Fig 9. Intruder Selfie of SIDS.



- Longitude = null

Someone try to unlock your phone Nokia 5.1 Plus !!!
the last detected location :
- Latitude = 36.531535
- Longitude = 8.839403333333333

Someone try to unlock your phone Nokia 5.1 Plus !!!
the last detected location :
- Latitude = 36.531338333333333
- Longitude = 8.838788333333333

Someone try to unlock your phone Nokia 5.1 Plus !!!
the last detected location :
- Latitude = 36.531405
- Longitude = 8.838840000000001

18:59 • via la carte SIM Ooredoo TN

Fig 10. SMS Alert of an Intrusion with Location of the Phone of SIDS.

A. Conclusion and Future Work

In conclusion, Smart IDS is introduced in order to detect intrusions when other defensive measures fall flat, by inactively observing system events and searching for security related issues. This paper gives a successful and productive procedure to detect noxious activities (attempt of authentication, selfie records of intruder) in the Phone. We have had the option to plan and build up an application named SIDS that can detect intrusion on Android Phone. SIDS was developed using Android Studio, Android SDK (software development kit) written with Java. The Object Oriented Analysis and Design Methodology (OOADM) were used for the analysis, design and development of the system and Unified Modeling Language (UML) to model the system.

The future work will involve the detection and screenshot of all activities of the intruder on the android phone. These activities will be sent to the email of the user and also kept on the log for the user's view with finger print required for access.

ACKNOWLEDGMENT

We thank all the authors for their huge contributions in this paper. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] Drake Bear. “How many people own smartphones around the world - Business Insider,” *Available: http://www.businessinsider.com/*how-many-people-own-smartphones-saround-the-world. 2017,

[2] Okoronkwo M.C and Onyedeke O.C. “An intrusion detection system(IDS) on android phones using a filter base feature selection algorithm,” *Int. journal of innovative research and developmenyt.*, vol.8, issue 11, pp. 101. 2019.

[3] Robert Moskowitz. “Network Intrusion: methods of Attack”. *RVS Conference*. 2020.

[4] Ansam Khraisat, Igbal Gondal, Peter Vamplew & Joarder Kamruzzaman. “Survey of intrusion detection systems: techniques, datasets and challenges”. *Scientific Data*. Vol 20, 2019.

[5] Dr. S. Vijayaram and Ms. Maria.“Intrusion Detection System – A Study”, *International Journal of security, Privacy and Trust Management (IJSPTM).* Vol 4, issue 1. 2015.

[6] Martin Borek. “Intrusion Detection System for Android: Linux Kernel System calls Analysis”. *School Of Information and Communication Technology, Sweden. 2017.*

[7] D. Ashok Kumar, S. T. Venugopalan. “Intrusion Detection Systems: A Review”. *International Journal of Advance Research in computer science.* Vol 8, issue 8. 2017.

[8] V. Grampurohit. “Android App Malware Detection,” *International Institute of Information Technology, India.* 2016.

[9] P. Garcı´a-Teodoroa, *, J. Dı´az-Verdejoa , G. Macia´-Ferna´ndeza , E. Va´zquezb. “Anomaly-based network intrusion detection: Techniques, systems and challenges”. Journal homepage: *www.elsevier.com/locate/cose,* computers & security. Vol 28, Pp 18–28. 2009.

[10] Yubin Kuang. “A Comparative Study on Feature Selection Methods and Their Applications in Causal Inference”. Department of Computer Science, Faculty of Science, *Lund University*, pp. 1-3. 2009.

[11] Yuyang Zhou, Guang Cheng, Shanqing Jiang, and Mian Dai. “An Efficient Intrusion Detection System Based on Feature Selection and Ensemble Classifier”. *School of cyber Science and Engineering*. 2019.

[12] Khurram Majeed1, DrYanguo Jing2, DrDusica Novakovic3, and Prof Karim Ouazzane4. “Behaviour based anomaly detection for smart phones using machine learning Algorithm”. *International conference on Computer Science and Information Systems (ICSIS)* Oct 17-18. 2014.

[13] Muhamed Halilovic, AbdulhamitSubasi. “Intrusion Detection on Smartphones”. *International Burch University Faculty of Engineering and Information Technologies,* Department of Information Technologies, Sarajevo, Bosnia and Herzegovina. 2014.

[14] Alessandri, D. “Using Rule-Based Activity Descriptions to Evaluate Intrusion Detection Systems”. *Recent Advances in Intrusion Detection, Third International Workshop.* 2017.

[15] Adetunmbi A.Olusola, AdeolaS. Oladele, DaramolaO,.Abosede. “Analysis of KDD ’99 Intrusion Detection Dataset for Selection of Relevance Features”. *Proceedings of the World Congress on Engineering and Computer Science*, Vol 1, Pp 2663-2664. 2010.

[16] A.A. Waskita, H. Suhartantoy, P.D. Persadhazy, L.T. Handoko. “A simple statistical analysis approach for Intrusion Detection System”. *Center for Development of Nuclear Informatics-National Nuclear Energy Agency*, Pp 1.2014.

[17] Yousef Farhaoui, Ahmed Asimi. “Creating a Complete Model of an Intrusion Detection System effective on the LAN”. *International Journal of Advanced Computer Science and Applications (IJACSA),* Vol. 3, No. 5, Pp 1-2. 2012.

[18] Mr. Akash J Wadate, Prof. N. R Chopde, Prof. D. R. Datar. “Malware Detection System for Android Mobile Applications“. *International Journal of Engineering Research and General Science,* Vol 4, Issue 1, Pp 21-22. 2016.

[19] Abdulla Amin Aburomman, Mamun Bin IbneReaz. “Evolution of Intrusion Detection Systems Based on Machine Learning Methods”**.** *Australian Journal of Basic and Applied Sciences,* Vol 7, no 7, Pp 46. 2013.

[20] Mehrnaz Mazini[a], Babak Shirazi[b], Iraj Mahdavi[b]. “Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and Ada Boost algorithms”. *Journal of King Saud University*, Pp 799-806, 2018.

[21] Mohammed A, Ambusaidi, Xiangjian He, Priyadarsi Nanda, Zhiyuan Tan. “Building an intrusion detection system using a filter-based feature selection algorithm”. *IEEE transactions on computers,* vol 1, no 1 Pp 1-3. 2014.

[22] P. Kaushik and A. Jain. “Malware Detection Techniques in Android,” *Internationnal Journal. Comput. Appl.*, vol. 122, no. 17. 2015.

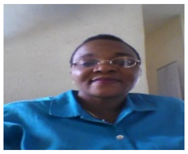
ONYEDEKE OBINNA CYRIL received his B.Sc. degree in computer science from Renaissance University, Nigeria (2011), M.Sc. degree in computer science from University of Nigeria, Nsukka (2020). His current address is Department of Computer Science, university of Kairouan, Tunisia. He has published over nine international journals and counting. He is a nominated Africom Scholar. He also has certification of award of achievement in project management from University of California, Berkeley. His current research interest includes: Network Security, Cryptology, information system. 0binna046@gmail.com

PROF. TAOUFIK ELMISSAOUI is an associate professor in the higher institute of Applied Mathematics and computer science, university of Kairouan, Tunisia since 2014. He is actually member of ELIVES project, point of contact of AFRICOM project and coordinator of the TNIS master in the higher institute of Applied mathematics and computer science. From 2006 to 2007 he was a staff engineer with Norkatech, Tunisia. From 2007 to 2013 he was an assistance professor in the Sousse University, Tunisia. He received his PhD, master's degree and engineering degree in telecommunications form National Engineering School of Tunisia ENIT). His research focuses on medical radar system and behind wall localization. elmissaoui.enit@gmail.com,

Dr. Okoronkwo, M. C (MSc, PhD), Senior lecturer, Department of Computer Science at University of Nigeria, Nsukka (UNN), MNCS, MCPN. Research interest is in ICT in Governance, Big Data, Artificial Intelligence and Networking. matthew.okoronkwo@unn.edu.ng.

Ihedioha Uchechi Michael is an Academic member of University of Nigeria Nsukka. My Current address is Department of computer Science, University of Kairuoan, Tunisia. I have participated in the publication of five journal articles till date and counting. I am also an Africom nominated Scholar. I hold a bachelor's degree in Computer science from the University of Nigeria Nsukka and an MSc in Information Technology from the National Open University of Nigeria and currently rounding off a second MSc in System Engineering in an Africom scholarship exchange program between the Universities of Nigeria Nsukka and the University of Kairuoan, Tunisia. mikeuche2002@gmail.com

Dr. Chikodili H.Ugwuishiwu is a lecturer and research fellow in the Department of Computer Science, University of Nigeria, Nsukka. She holds different degrees including BSc. (2004), MSc. (2009) and Ph.D (2018), all from Computer Science Department in University of Nigeria Nsukka. Her areas of research interest are on Information System (IS), Computer modeling and Simulation and Data mining. She has published many articles in journals and conferences, both local and international. She has also attended many local and international workshops. She is a member of professional bodies including Computer Professionals Registration Council of Nigeria (CPN), Nigeria Computer Society (NCS), Nigeria Women in Information Technology (NIWIIT), Organisation for Women in Science for the developing world (OWSD) and Association of Information System (AIS). chikodili.ugwuishiwu@unn.edu.ng



Okwume Benedette Onyebuchi, a staff of University of Nigeria Nsukka. She holds a BSc, MSc in computer science and currently running her PHD in the computer science department, University of Nigeria Nsukka. Areas of interest include Artificial Intelligence, Data Mining, and Queuing Theory and information system. okwumebenedette@gmail.com

Hybrid Machine Learning: A Tool to Detect Phishing Attacks in Communication Networks

Ademola Philip Abidoye[1], Boniface Kabaso[2]
Department of Information Technology
Cape Peninsula University of Technology
Cape Town, South Africa

***Abstract*—Phishing is a cyber-attack that uses disguised email as a weapon and has been on the rise in recent times. Innocent Internet users if peradventure clicking on a fraudulent link may cause him to fall victim to divulging his personal information such as credit card PIN, login credentials, banking information, and other sensitive information. There are many ways in which attackers can trick victims to reveal their personal information. In this article, we select important phishing URLs features that can be used by an attacker to trick Internet users into taking the attacker's desired action. We use two machine learning techniques to accurately classify our data sets. We compare the performance of other related techniques with our scheme. The results of the experiments show that the approach is highly effective in detecting phishing URLs and attained an accuracy of 97.8% with 1.06% false-positive rate, 0.5% false-negative rate, and an error rate of 0.3%. The proposed scheme performs better compared to other selected related work. This shows that our approach can be used for real-time applications in detecting phishing URLs.**

Keywords—Phishing attack; data sets; URL classification; phishing URL; attackers; machine learning; classifiers; Internet

I. Introduction

In the last decade, Internet usage has been increasing tremendously and makes our lives easy, simple, and transforms our daily lives. It plays a major role in the areas of communication, education, business activities, and commerce [11, 27]. A lot of useful data, information, and knowledge can be obtained from the Internet for personal, organizational, economic, and social development. Positive and productive use of the Internet will assist users to become successful in their careers and businesses. The Internet makes it easy to provide many services online and enables us to access various information at any time, from anywhere around the world. Online banking, including transferring money between accounts, online bills paying, and so on. These services have become very prevalent as more financial institutions start to provide almost free online services. Presently, about 40% of the world population are connected to the Internet [22]. The main purpose of the Internet is to provide worldwide access to various types of data for advancing research in engineering, science, design, and medicine as well as in maintaining global defense and surveillance [7]. However, as more people are using the Internet globally, different kinds of attacks have been identified including denial-of-service and distributed denial of service attacks, drive-by attack, man-in-the-middle attack, password attack, eavesdropping, and phishing attack [30]. Over the last decade, phishing has skyrocketed to staggering proportions and will continue to increase due to various phishing groups using different methods of attacks. Therefore, it is imperative to comprehensively study the mode of operation of attackers. The word phishing comes from the fact that cyber-attackers are fishing for sensitive data and information. The “ph” is coined from the advanced methods the phishers employ to distinguish their activities from the more simplistic fishing. The concept of phishing is a form of social engineering and can be traced back to the early 1990s via America Online (AOL) [8].

Phishing is the act of sending a fake email, messages, or malicious websites to trick the recipient/Internet users into divulging sensitive personal information such as personal identification number (PIN) and password of their bank account, credit card information, date of birth, or social security numbers. To perpetuate this type of attack, the attacker usually poses as a trustworthy organization. For instance, an attacker may send an email that looks like it is from a financial institution or a reliable credit card company requesting for their account information by tricking the target that there is a problem or a need to update his/her within a stipulated time. There were 112163 unique phishing attacks and 60889 unique phishing sites reported in the U.S. in June 2019 [3]. Phishing attacks affect hundreds of thousands of internet users across the globe. Individuals and organizations have lost a huge sum of money and private information through phishing attacks [12].

What differentiates phishing from other Internet attacks is the form the message takes: the attackers disguise as a real person, trusted entity of some kind, or an organization the target might transact business with. It is one of the fastest-growing types of cyber-attack and most widespread due to financial gain the attackers derive from any successful phishing. The attackers capitalize on some recipients’ desire to respond to urgent requests from their “financial institutions” by clicking a link or download an attachment provided in a spoofed email that looks "official", but it is linked to a fraudulent website(s) which may result in financial losses, identity theft, or other fraudulent activity.

A. Statistics of Phishing Attacks

The sudden attack of phishing against financial institutions was first known in July 2003. Since then, commercial banks, E-gold, and E-loan are the main target of the phishers. Among financial institutions that have been attacked in the U.S., commercial banks account for 91 percent of the attacks while insurance companies account for 7 percent. Similarly, about 39 percent of the total retail banking activities and 25 percent of the credit-card companies have been attacked in 2018 [6].

The number of global phishing attacks rose to 129.9 million during the second quarter of 2019; it increased by 21% more than the same quarter of 2018. Greece has the highest number of phishing attacks at 26.2%, followed by Venezuela, Brazil, Australia, and Portugal. In terms of financial institutions and establishments, commercial banks have the highest percentage of phishing emails at 30.7%, followed by payment systems at 20.1%, worldwide Internet portals at 18%, and social networks at 9% [15]. The act of phishing is not limited to a particular country; it occurs everywhere and every day. The reason is that phishers are using the Internet to phish unsuspecting Internet users for financial gain [9]. Phishing information flow is shown in Fig. 1.

Phishers are looking for more effective and advanced ways to launch phishing attacks. They are developing new techniques for attacks and improving on the old ones. With the advancement in technology, they have refined their attacks both in the usage of websites and emails. They can develop more innovative and effective methods of targeting innocent victims. It is essential to note that different phishers have various methods they use for phishing, but all have similar techniques and tools. These methods can be majorly grouped into three namely impersonation, forwarding, and popups [28].

In recent years, researchers and stakeholders have paid much attention to the problem of phishing and how it could be solved. They have developed different approaches in the literature for detecting malicious uniform resource locators (URLs) and emails. Some of these approaches are presented below.

B. Aim of Research

This work aims to develop a technique that can detect all forms of phishing strategies created by attackers in communication networks. We generate our set of rules which rely on our observations and hybrid machine learning techniques. We gather different methods and tricks used by attackers to entice unsuspecting victims to fabricated web pages and use those attributes to design our rule data sets.

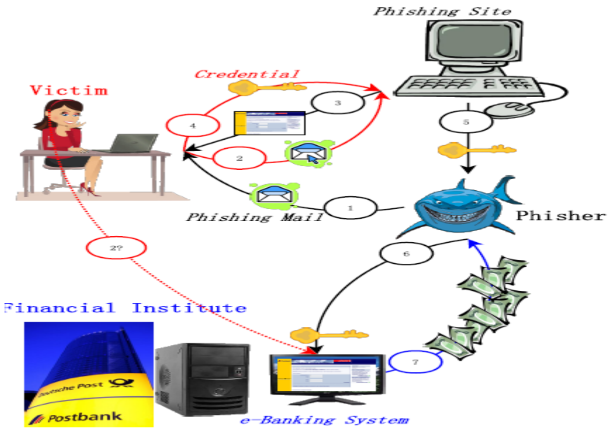


Fig. 1. Phishing Information Flow [20].

C. The Significance of the Study

In recent times, there is an increasing need to identify phishing URLs and emails because of the negative effect they have on their targets. Researchers have developed various methods and applications for exposing phishing websites and detecting malicious emails, but only a few scholars have used machine learning methods for detecting phishing websites. In this study, we are using a hybrid machine learning technique for detecting phishing URLs. A combination of Support Vector Machines (SVM) and Naïve Bayes techniques are used for accurate phishing URLs detection and to minimize false positive detection. This approach provides up to date protection against zero-day phishing attacks.

D. Problem Statement

Phishing detection methods do suffer from low detection accuracy and high positive false alarm, particularly when new phishing techniques are invented. Besides, a blacklist is a common method for detecting phishing URLs but it is ineffective in responding to new phishing attacks since it is now very easy to register a new domain, no comprehensive blacklists can ensure an adequate up-to-date database.

Researchers have developed various approaches to detect phishing websites using different learning algorithms, but this problem still needs more attention of the researchers because new phishing websites are being deployed every day and phishers are using different techniques to lunch their attacks. Consequently, most of the solutions provided for phishing attacks were based on small experimental data sets, the accuracy and effectiveness of these algorithms on real large data sets cannot be ascertained. Thus, the number of malicious websites increases very fast, how to detect phishing websites from a large number of legitimate websites in real-time with high accuracy must also be addressed. It is imperative to design intelligent anti-phishing algorithms that are capable of detecting ever-increasing phishing attacks. A hybrid machine learning technique is used for the detection of phishing URLs. We use both SVM and Naïve Bayes classifiers for the detection since no single classifier is perfect. SVM scales relatively well to high dimensional data, and error can be explicitly controlled. Also, it is very easy to implement. However, it does not scale very well for a large data set. Naïve Bayes classifier is used to overcome the weakness in SVM. This classifier is capable of handling large data sets and scales linearly with the number of predictors and data points.

E. Contributions

This research work uses hybrid machine learning techniques to accurately classify our data sets into either phishing or benign URLs in communication networks. These two classifiers are used together because strengths in one classifier complement the weaknesses in the other classifier. Besides, we use 13 important lexical features to model our classifiers to achieve high precision and to provide a better-accuracy trade-off. We observe that using important lexical features increases the overall classification across all the data sets and minimize the error rate. This shows that the proposed approach can be used for near real-time applications in detecting phishing URLs.

The rest of this article is organized as follows. In Section 2 related work is discussed. Section 3 discusses the proposed approach. Data used for the experiments, relevant features in predicting phishing URLs, and the classifiers used are discussed in this section. In Section 4, we present the various experiments conducted and also discuss the performance evaluation of the two machine learning techniques used. Finally, the conclusion is presented in Section 5.

II. Related Work

Blacklisting and whitelisting are the two widely methods that have been used to manage which entities get access to our system.

A blacklist is a list of suspicious or forbidden URLs that should be blocked or denied access on a network or system. This method is very simple to implement. It is just to deny any strange or suspected URLs access to the network. However, this method is too weak to detect the majority of phishing incidents since new threats are many and constantly emerge every day, such as a zero-day attack. This approach is incapable of detecting or stopping any new kind of attack. It requires keeping a comprehensive list of suspicious websites and their reports which consume a lot of system resources [18]. Phishers sometimes design URLs specifically to evade detection by tools that use a blacklist system. Finally, this approach fails to identify some types of attacks that target a profitable organization.

On the other hand, a whitelist allows several websites to be accessed and blocks other websites that are not on the list. It denies any new URL unless it is proven to be benign (legitimate). Whitelist applications can be used to identify websites by their file name, size, and directory path. Thus, whitelisting access control is higher than blacklisting, as the default is to block websites and allows only those websites that are proven to be legitimate to be accessed. However, its implementation is more complex and hard to assign because it requires more information on the application being used to create the whitelist. Also, it is infeasible to create a whitelist that contains all the list of legitimate sites due to their large number [19]. Another challenge of whitelisting is that a user must remember to check the interface each time he visits any website. Thus, there is a need to develop innovative methods that are capable to detect any recent methods the phishers are using for phishing.

A recent increase in suspicious URLs has attracted the attention of many researchers, and they have developed different techniques for website phishing detection. The definition of phishing constantly changes concerning the way phishing is performed. Email and website are the two major methods the phishers are using for phishing. These two methods have the same goal but there are some differences between the two.

Aburrous et al. [1] proposed an intelligent system for phishing webpage detection in e-banking. They developed a model that combines fuzzy logic with a data mining algorithm to detect phishing websites and categorize the phishing type using 10-fold cross-validation. This model achieved 86.38% grouping accuracy. However, this model has a high percentage of false positive.

Basnet et al. [5] proposed a heuristic-based approach to group phishing URLs by using the data available only on URLs. The authors used a binary classification method to detect phishing URLs and grouped URLs into phishing URLs and legitimate URLs. The results of the experiments show that the proposed approach is very effective in detecting phishing URLs compared to related work. However, this approach is only tested on a data set that is less than 300. It may not be effective on a large data set.

Jain and Richariya [13] developed a new method for detecting phishing emails using link-based features. A prototype web browser was used as a means to process each incoming email to detect a phishing attack. A combination of the prototype and their algorithm assist the system users to be notified of possible attacks and prevent them from clicking any malicious URLs.

Mahmood and Rajamani [21] proposed an anti-phishing detector (APD) technique based on association rule mining for detecting phishing websites. APD dynamically traces out any possible phishing attacks during message transmission between computer users. In addition, the authors developed an algorithm to extract frequently reoccurring words and forward the information to APD for further processing. The results of the approach shown to be effective.

Ajlouni et al. [2] proposed a method for detecting phishing websites based on associative classification algorithms. It is an improvement over [1]. The results of the experiment show that the method achieved 98.5% accuracy in detecting phishing webpages. However, there is no information about how many rules they used for the extraction.

Zhang et al. [32] proposed a new classification method based on a Sequential Minimal Optimization classifier algorithm that consists of features of websites. The results show that the algorithm performs better than the selected baseline. However, this approach can only detect phishing webpages with the Chinese language.

A new rule-based approach for detecting phishing attacks in internet banking is presented in [23]. The authors used two feature sets that have been developed to find webpage identity and support vector machine algorithm to classify webpages. The proposed features are independent web browser history or search engine results. The results of the experiments show that the method can detect phishing webpages with an accuracy of 99.14% true positive and only 0.86% false-negative alarm.

Ramesh et al. [25] developed a method for detecting phishing webpages. The webpage is scrutinized and classified as indirect and direct links associated with the page. Indirect link features are extracted from the search engine result while direct links are extracted from the page contents. Also, they used a third-party DNS lookup to match the domains of a malicious webpage and phishing target to the corresponding IP address. The results of this approach achieve 99.62% accuracy. However, the efficiency of this method depends on largely the speed of search engine and DNS lookup time which can affect its performance. A comparison of the related studies that have been used to detect phishing URLs in the literature with our work is presented in Table I.

TABLE I. EVALUATION OF RELATED WORK WITH PROPOSED APPROACH

Work	Approach	A	B	C	D
[1]	Fuzzy logic	No	No	Yes	Yes
[5]	Binary clarification	Yes	No	Yes	Yes
[13]	Web browser	No	No	Yes	Yes
[21]	Rule-based (APD)	No	No	Yes	Yes
[2]	Data mining	Yes	Yes	Yes	no
[32]	Sequential Minimal Optimization	No	Yes	No	Yes
[23]	Rule-based approach	Yes	No	No	No
[25]	Domain identification	Yes	Yes	Yes	No
	Proposed approach	Yes	Yes	Yes	Yes

where A = Zero-day phishing detection
B = 3rd-party services' Sovereignt
C = Search engines sovereignty
D = Language sovereignty

III. PROPOSED APPROACH

In this section, we present in detail our method for detecting malicious URLs. The approach is divided into two parts, and each part's output is an input to the next part as shown in the proposed framework in Fig. 2.

The first part is based on data collection, processing of data sets, and URLs feature extraction. We consider different heuristic features in the structure of URLs, ranging from a generic social engineering feature, lexical feature in the URL, multiple alphabets, and phishing target brand name. The feature vector is constructed with 13 important features to model our classifiers. The second part is based on the classification of data set using a hybrid of machine learning classifiers to evaluate our approach. –We performed different experiments. The results of the experiment show that our scheme achieves 97.8% accuracy on average. The description of each part is briefly discussed in the following subsections.

A. Processing of Data Sets and URLs Features Extraction

A large number of data sets (36,874), discussed in sub-Section 3.1, were collected and processed to make them suitable for the requirement of this study. The processing involved many stages, these include webpages feature extraction, data standardization, and attribute weighing. These steps are very important so that the classifiers would be able to understand the data sets and appropriately categorize them into their classes. The classifier is regularly trained with new phishing web pages to learn new trends in phishing. The outcome of this phase is used as input to the next part of the appropriate classifiers.

We propose a hybrid machine learning approach to effectively classify phishing URLs based on the information available to an individual URL. Phishing URLs are treated as a binary classification problem with the benign URLs belong to the negative class and phishing URLs belonging to the positive class. We collected our phishing and benign URLs from PhishTank, Yahoo directory, and the Google engine to form our data sets. Thereafter, we extract many features that have proved to be effective in predicting phishing URLs by employing different publicly available resources to classify the data sets into their respective classes. We apply both SVM and Naïve Bayes algorithms to create models from training data sets which consist of feature extractions and class labels. Fig. 2 shows the proposed framework for phishing URLs detection.

We use two types of data sets for this research. The first set is phishing data sets and the other one is benign data sets. The data sets are collected from different credible sources [10, 24],

The data sets contain 36874 URLs with their related features. We wrote Python scripts code to automatically download certified phishing URLs from PhishTank.

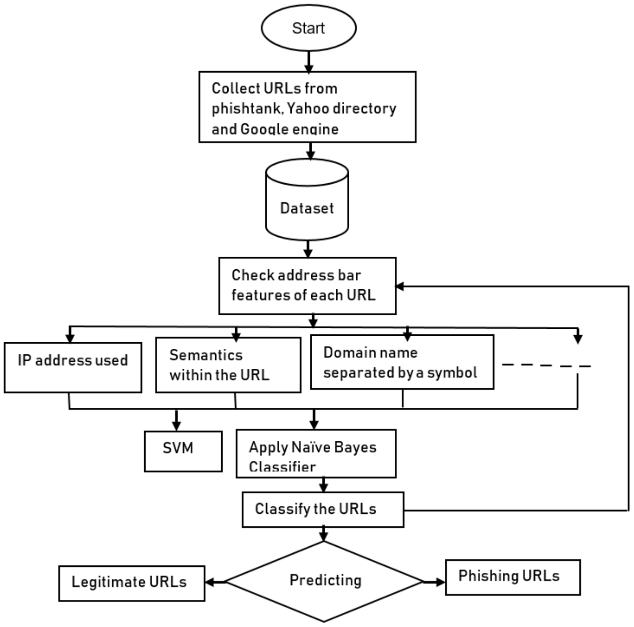

Start
Collect URLs from phishtank, Yahoo directory and Google engine
Dataset
Check address bar features of each URL
IP address used
Semantics within the URL
Domain name separated by a symbol
SVM
Apply Naïve Bayes Classifier
Classify the URLs
Legitimate URLs
Predicting
Phishing URLs

Fig. 2. The Proposed Framework for Detecting Phishing URLs.

B. Phishing Data Sets

PhishTank is a joint project to which people can submit suspicious phishing URLs for confirmation. It is a public clearinghouse for phishing URLs [4]. Suspicious URLs are further scrutinized by many people before being confirmed as phishing URLs and added to a blacklist. PhishTank provides a comprehensive list of current and active phishing URLs.

Researchers and developers can download phishing URLs from the Phishing Web site after signing up. They would be able to download the URLs from PhishTank in different file formats with an API key.

We downloaded two sets of phishing URLs. The first data-set is referred to as DTS1, contains 14,298 phishing URLs. They were collected from March 4, 2019, to April 19, 2019, based on the reports in [26] which shows that phishing attacks are usually higher during this period than the preceding months. Also, we observe that phishers constantly develop new tactics to get personal information from unsuspecting users, to explore various and recent methods the attackers are using motivated us to collect the second sets of data. The second set of data, referred to as DTS2, contains 7,350 phishing URLs. They were collected from November 1 to December 4, 2019. We chose this period because it has a special day “Black Friday” (November 29, 2019) in which many people have been waiting for to buy cheap goods from stores, online using their credit or debit cards. Phishers also use this period as an opportunity to display their tactics and launch different attacks on unsuspecting users. A total of 21,648 phishing URLs was collected from the PhishTank Web site.

C. Legitimate Data Sets

Our benign URLs were collected from the Yahoo directory. Yahoo provides a generator that arbitrarily produces an URL in its directory each time the Web page is visited. This service is used to randomly choose an URL and download the contents of the Web page with the server header information. This service is used to collect 9,045 random URLs from May 6, 2019, to June 10, 2019. Our list consists of URLs from financial institutions, e-commerce, online services, cloud storage, religious organizations to get different URL structures and Web page contents [16]. To provide more learning instances for legitimate URLs, we chose 6,181 legitimate URLs from the Open Directory Project (DMOZ) Web directory [29]. DMOZ is a multilingual open-content directory of World Wide Web links containing more than three million URLs.

We use a Google tool to analyze the list of benign URLs collected and crawled. These URLs are used as legitimate webpages based on the assumption that all the URLs extracted were benign since they were downloaded from legitimate Internet sources.

Python and Java scripts are used to parse the legitimate and phishing URLs and extract the features discussed in subsection 3.2. Web pages that we could not extract features from their contents were discarded to get only valid URLs for our data sets. The total number of our data sets is presented in Table II.

D. Data Authentication

Data sets collected need to be authenticated to ascertain the real status of the URLs, particularly in the case of phishing websites as it is known that the phishing website only lasts a few weeks [31]. Thus, every URL needs to be authenticated before processing.

In this section, we present relevant features that are effective in predicting phishing web sites. Each feature is discussed with its associated rules.

TABLE II. DATASETS FOR PHISHING URLS DETECTION

Data set	**Phishing**	**Non-phishing**	**Total data sets**
DTS1	14,298	9,045	23,343
DTS2	7,350	6,181	13,531
DTS1 + DTS2	21,648	15,226	36,874

A generic social engineering feature

Phishers use generic greetings in their messages such as “Sir”, “Dear Bank Customer”, “Dear Customer”, and “Dear Member” to address their target victims. The content of the message is always threatening such as “please update your bank account to prevent it from being blocked”, "Your account has been compromised!", "Urgent action required!", "Your account will be closed!" These intimidation strategies are becoming more common than the promise of "instant riches"; taking advantage of victims’ anxiety and concern to get them to provide their personal information.

$$Rule\begin{cases} \boldsymbol{if}\ the\ greeting\ is\ directed\ to\ account\ owner \\ \quad and\ do\ not\ require\ to\ supply \\ \text{a piece of } personal\ information\ via\ a\ link \\ \quad in\ the\ message \rightarrow Legitimate \\ \boldsymbol{else\ if} \text{ the } greeting\ is\ generic \rightarrow Suspicious \\ \boldsymbol{else}\ update\ your\ information \\ \quad via\ a\ given\ link \rightarrow Phishing \end{cases}$$

Lexical features explain lexical patterns of phishing URLs such as long IP addresses, special characters, number of dots, and so on.

IP-based URL

Internet Protocol (IP) address is one of the ways to hide the webpage address. If an IP address is used instead of a Domain Name System (DNS) address in the URL, it will be difficult for innocent users to ascertain where they are being directed to when they click the link or press the Enter key on their system to load the page. Another reason for using the IP address is that phishers would not like to spend money to buy a domain for their phony web pages.

$$Rule\text{:}\begin{cases} \boldsymbol{If}\ the\ domain\ name\ has\ an\ IP \\ \quad Address \rightarrow Phishing \\ \boldsymbol{else} \rightarrow Legitimate \end{cases}$$

Long URL to hide the fake part

Attackers can use lengthy URLs to mask the fake part in the address bar. For instance,

“http://prudentbank.com/2k/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&amp;dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.net.html”

We computed the length of URLs in our data sets and determined their average length to ensure the accuracy of our research. The findings showed that if the URL length is less than 52 characters, it is classified as legitimate; it is suspicious if the length is between 52 and 73 characters, and it is a phishing URL if the URL is more than 73 characters. A method based on frequency has been used to update this feature rule, which improves its accuracy.

$$\text{Rule: }\begin{cases} \boldsymbol{If}\ URL\ length < 52\ \ characters \rightarrow\ Legitimate \\ \boldsymbol{else\ if}\ URL\ length \geq 52\ and\ \leq 73 \\ \quad characters \rightarrow Suspicious \\ \boldsymbol{else} \rightarrow Phishing \end{cases}$$

Shortened URL "TinyURL"

Short URL enables to reduce long links from social networks and top sites on the Internet. This is achieved by the service provider through an "HTTP Redirect" on a domain name that is short and redirects to the corresponding long URL [17]. For instance, an URL for Wiki's article "http://en.wikipedia.org/wiki/URL_shortening" contains 64 characters and its corresponding short URL http://bit.ly/c1htE; it contains 16 characters with Bitly's default domain name "bit.ly" and the hash "c1htE" as the back-half. A hash only consists of letters and numbers "a-z, A-Z,0- 9". Attackers use this shortened URL feature to hide links to infected websites or phishing.

Rule: $\begin{cases} \boldsymbol{if}\ TinyURL \ \rightarrow\ Phishing \\ \quad \boldsymbol{else} \rightarrow\ Legitimate \end{cases}$

URL's having "@" Symbol

Using "@" symbol within the URL causes the Web browser to read the right side of the browser address and ignore everything preceding the "@" symbol. For instance, in this URL www.prudentbank.com@www.google.com, the browser will ignore "www.prudentbank.com" and only read www.google.com which it may be used to hide a phishing URL.

Rule: $\begin{cases} \boldsymbol{if}\ URL\ having\ @\ symbol \rightarrow\ Phishing \\ \boldsymbol{else} \rightarrow\ Legitimate \end{cases}$

Hovering of a Mouse over Hyperlink Feature

One of the tactics of phishers is that they use legitimate domain names for their links to send messages to their potential victims while the destination URLs are hidden from them using HTML code. For instance, a phisher may send this link <a href = "http://phishing.com" > www.prudentbank.com </a> to unsuspecting Internet users which looks like a Prudent Bank Website whereas the destination URL "http://phishing.com" is hidden from the user. If the user clicks the link "www.prudentbank.com" it will take him to "http://phishing.com" thinking that they are surfing a legitimate website. To check if a link is malicious or not, a mouse is hovering over the link to view the destination URL.

Rule: $\begin{cases} \boldsymbol{if}\ destination\ URL\ is\ the\ same\ with\ the\ domain \\ \quad name\ and\ the\ link\ leads\ to\ the \\ \quad homepage \rightarrow\ Legitimate \\ \boldsymbol{else\ if}\ the\ destination\ URL\ cannot\ be \\ \quad determined \rightarrow\ Suspicious \\ \boldsymbol{else}\ the\ destination\ URL\ does\ not\ the\ same \\ \quad with\ the\ domain\ name \rightarrow Phishing \end{cases}$

Redirecting using "//"

The presence of "//" in the URL path shows that an innocent user will be redirected to another infected website. For example, http://www.legitimate.com//http://www.phishing.com.

This study examines the position of "//" in a legitimate URL. If the URL begins with "http" then "//" should appear in the 6th position and the 7th position if it begins with "https".

Rule: $\begin{cases} \boldsymbol{if}\ the\ position\ of\ \ "//"\ in\ the\ URL\ >\ 7 \\ \quad \rightarrow\ Phishing \\ \boldsymbol{else} \rightarrow\ Legitimate \end{cases}$

Domain name separated by a dash symbol

It is very rare for a legitimate domain name to be separated by a dash symbol (-). Phishers use this method to trick Internet users by adding a dash symbol (-) within the domain name so that users will think that they are surfing a legitimate webpage. For instance, http://www.pay-pal.com/.

Rule: $\begin{cases} \boldsymbol{if}\ Dash\ symbol\ (-)\ is\ part\ of\ a\ domain \\ \quad name\ \rightarrow\ Phishing \\ \boldsymbol{else}\ \rightarrow\ Legitimate \end{cases}$

Subdomain of a subdomain

A URL might include an Internet country code top-level domain (ccTLD) to identify a particular country. For instance, http://www.prudentbank.com.za/login/. "za" is a ccTLD, and the ".com" portion of the extension shows that the domain name is a commercial entity. Taking the two extensions together ". com.za" is called a second-level domain (2LD) and "prudent bank" is the real domain name. To minimize rules for extracting this feature, first, we remove subdomain "www" from the URL and ccTLD if the extension is part of the URL. Thereafter, the number of dots in the URL is counted. If the number of dots is one, then the URL is legitimate. It is suspicious if the number of the dots is two since the URL has one subdomain. It is declared phishing if the number of dots is more than two since it will contain many subdomains.

Rule: $\begin{cases} \boldsymbol{if}\ \text{the}\ number\ of\ dots\ in\ domain\ portion = 1 \\ \quad \rightarrow\ Legitimate \\ \boldsymbol{else\ if}\ dots\ in\ domain\ portion = 2 \\ \quad \rightarrow\ Suspicious \\ \boldsymbol{else} \rightarrow\ Phishing \end{cases}$

A domain name containing multiple alphabets

It is possible to register domain names in other alphabets such as Chinese, Arabic, French, German, or anything that can be represented with the Unicode standard since 1998. Phishers have taken advantage of this unique feature by finding characters in other alphabets which look similar to the Latin ones to lure users into a phishing website. For instance, in this URL "https://apple.com", the domain name can be registered with "xn--pple-43d.com". The URL is equivalent to "https:// xn--pple-43d.com". Thus, most users will fall for this trick because their browsers will show the green padlock icon, showing that the user is on a secure connection but in fact, a bunch of Cyrillic characters is embedded within the multiple alphabets.

Rule: $\begin{cases} \boldsymbol{if}\ domain\ name\ containing\ multiple \\ \quad alphabets\ \rightarrow\ Phishing \\ \boldsymbol{else}\ \rightarrow\ Legitimate \end{cases}$

Phishing website longevity

We believe that legitimate websites will be hosted and regularly paid for one or more years in advance. It has been shown that a phishing website exists for a short period to avoid being detected [14]. In our data sets, the longest fake domains that have been used are only for six months.

Rule: $\begin{cases} \boldsymbol{if}\ domains\ expire \leq\ six\ months\ \rightarrow\ Phishing \\ \boldsymbol{else} \rightarrow\ Legitimate \end{cases}$

IV. Detection of Phishing URLs

We use a hybrid machine learning classification techniques in detecting phishing URLs. A feature vector matrix is built from our data sets presented in Table I. Each vector-matrix consists of 13 important lexical features described above. We use two variables to classify the data sets: -1 for a legitimate URL and 1 for a phishing URL as shown in equation (1). This gives a feature matrix-vector of 36,874 denoting the total number of the data sets.

There are many machine learning classification algorithms, we classified our data sets using the following classification algorithms. Metrics for classification are discussed thereafter.

A. Support Vector Machines (SVMs) Classifiers

In any classification process, both a parameter and a model technique should be chosen to achieve a high level of performance of the machine learning. Recent methods enable different kinds of models of varying complexity to be selected.

This study uses a linear classifier of the form: $f(X_i) = W.X_i + b$ where . represents the dot product, W denotes the weight vector, X_i is input data, and b denotes a learned bias vector.

Let $\{X_i\}$ denote the features of our data sets for all $i = 1,\ 2,3, \ldots., n,\quad X_i \in \mathbb{R}^d$, and $y_i \in \{-1, 1\}$ denote class labels (indicator variable). Our goal is to classify the data sets correctly. The following mathematical equations need to be satisfied to achieve this goal as shown in equation (1). SVM data sets classification is contained in Algorithm 1.

$$f(X_i) = \begin{cases} \geq 0 & y_i = +1 \\ < 0 & y_i = -1 \end{cases}$$

$$.X_i + b \geq 1 \tag{1}$$

$$W.X_i + b < 1$$

$$y_i(W.X_i + b) \geq 1, \text{ for all } i$$

B. Naïve Bayes Classifiers

Naive Bayes classifiers are a group of classification algorithms based on Bayes' Theorem. The underlying assumption of these classifiers is that all the features used for the classification are autonomous of each other. In other words, it assumes that the existence of a specific feature in a data set is unrelated to the existence of any other feature. The Bayes can consider all the features of data sets and correctly classify them. It provides a way of determining posterior probability $P_r(y|X_i)$ from $P_r(X_i)$, $P_r(y)$, and $P_r(X_i|y)$ as shown in equation (2).

Fig. 3 shows the process of experimenting before arriving at our results.

$$P_r(y|X_i) = \frac{P_r(X_i|y) * P_r(y)}{P_r(X_i)} \tag{2}$$

Above,

$P_r(y|X_i)$ is defined as the posterior probability of class (legitimate or phishing URL) given the predictor (feature).

$P_r(X_i)$ is the probability of a predictor.

$P_r(y)$ is the probability of the class.

$P_r(X_i|y)$ is the probability of the predictor given class.

The variable $y = y_k$ denote the class defined above and variable X_i denote the features of our data sets such that

$$X_i = (X_1, X_2, X_3, \ldots \ldots., X_n)$$

Substituting for X_i in equation (3) and expanding using the chain rule

$$P_r(y|X_1, X_2, .., X_n) = \frac{P_r(X_1|y)P_r(X_1|y)\ldots\ldots P_r(X_n|y)P_r(y)}{P_r(X_1)P_r(X_2)\ldots\ldots P_r(X_n)} \tag{3}$$

The value of the denominator remains static for all values in our data set. Thus, the denominator is eliminated and proportionality is introduced as follows.

$$P_r(y|X_1, X_2, .., X_n) \propto\ P_r(y) \prod_{i=1}^{n} P_r(X_i|y) \tag{4}$$

The above function is further used to classify our data sets, X_i, into two classes: legitimate or phishing URLs. Model in Fig. 3 is developed to classify the data sets.

Algorithm 1: SVM Data Classification

```
Begin
Given a hyperplane $W.X + b$
$f(X_i) = W.X_i + b$ for all $i = 1,2,3, \ldots \ldots n$
The classifier can be expressed as
$f(X_i) = \widetilde{W}.\widetilde{X_\imath} + w_o = W.X_i$
where $W = (\widetilde{W}, w_o)$, $X_i = (\widetilde{\widetilde{X_1}}, 1)$
Let $W = 0$
Considering the data sets and class labels, $\{X_i, y_i\}$
   $f(X_i) = sign\ (\sum w[i]\, x[i] + b)$

if $X_i$ is wrongly classified then $W \leftarrow W + \beta * sign(W.X_i + b)$
Else
Continue until all the data sets are correctly classified
end if
```

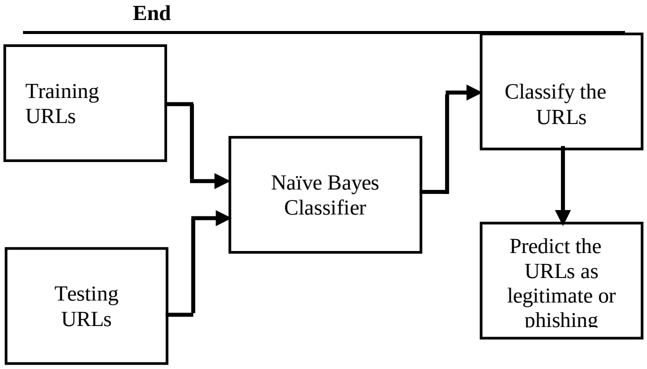


Fig. 3. The Proposed Model to Detect Phishing Attacks.

C. Metrics used for Evaluation

The following metrics are used for evaluation of the proposed scheme to eliminate or minimize misclassification in our data sets. We assume that a legitimate website is negative and a phishing website as positive. include i) True positive rate (TPR) also called sensitivity, ii) False positive rate (FPR) also called specificity, iii) true negative rate (TNR), and false-negative rate (FNR). These prediction outcomes are summarised in Table III.

TABLE III. PREDICTION OUTCOMES FOR PHISHING URLS DETECTION

Expected class			
Classes		True	False
	True	True positive (TP)	False-positive (FP)
	False	False-negative (FN)	True negative (TN)

*True positive rate (*Sensitivity*)*: It is defined as the proportion of legitimate websites that are correctly classified as legitimate. It is mathematically expressed as follows.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (5)$$

*False-negative rate (*Specificity*):* FN is defined as the proportion of phishing websites that are correctly classified as phishing.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad (6)$$

False-positive rate (FPR): It is defined as the proportion of phishing websites that are wrongly classified as legitimate websites. It is mathematically expressed as follows.

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \qquad (7)$$

True negative rate (TNR): It is defined as the proportion of legitimate websites that are wrongly classified as phishing websites. It is mathematically expressed as follows.

$$\text{TNR} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (8)$$

Accuracy: Accuracy (ACC) is determined as the number of all correct predictions divided by the total number of the dataset. It is mathematically expressed as follows.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FN+ FP}} = \frac{\text{TP+TN}}{\text{P+N}} \qquad (9)$$

Error rate: Error rate (ERR) is determined as the number of all wrong predictions divided by the total number of the dataset. It is mathematically expressed as follows.

$$\text{Error rate} = \frac{\text{FP+FN}}{\text{TP+TN+FN+ FP}} = \frac{\text{FP+FN}}{\text{P+N}} \qquad (10)$$

V. EXPERIMENTS

To evaluate the proposed scheme, we used two machine learning techniques: Support Vector Machines (SVM) and Naïve Bayes to classify our train data sets into two classes. Many experiments were performed on the data sets to test whether the input URLs are malicious or benign. The URLs were entered into the python program and extracted the URLs features. The results of the classification are presented in Table IV. The table shows the 5th percentile, 95th percentile, median, and standard deviation (SD) values for the *Accuracy* of each classifier for four different number of runs using all the important features discussed above.

A. Analysis and Discussion of Results

To test the accuracy of the algorithms, we obtained the following experimental results and present them in a tabular form as shown in Table IV.

Also, we conducted more experiments on the classification of the URLs. Fig. 4 shows the graphical representation of phishing and benign values for the next experiment. A total number of 18108 URLs are phishing and 1892URLs are benign.

Moreover, Fig. 5 shows the graphical representation of phishing and benign values. A total number of 22897 URLs are for phishing and 2103 are benign.

Similarly, Fig. 6 shows the graphical representation of phishing and benign values. A total number of 23851 URLs are phishing and 6149 are benign.

TABLE IV. EXPERIMENTAL RESULTS OF THE PHISHING CLASSIFIERS

Experiment	**URLs**	**Phishing**	**Benign**
Exp1	1000	991	9
Exp2	2000	1987	13
Exp3	3000	2947	53
Exp4	4000	3850	150
Exp5	5000	4766	234
Exp6	6000	5683	317
Exp7	7000	6708	292
Exp8	8000	7671	329
Exp9	9000	8518	482
Exp10	10000	9376	624
Exp11	11000	10431	569
Exp12	12000	11498	502
Exp13	13000	12602	398
Exp14	14000	13255	745
Exp15	15000	13943	1057

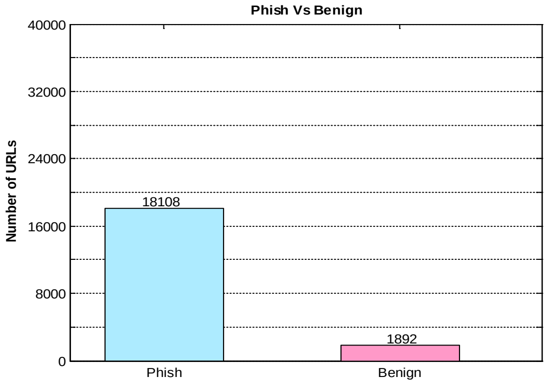


Fig. 4. Graphical Classification for 20,000 URLs.

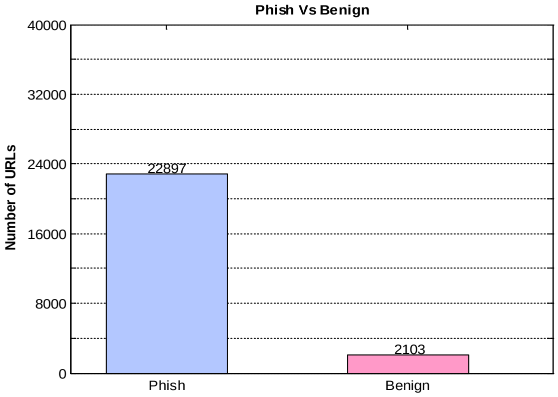


Fig. 5. Graphical Classification for 25,000 URLs.

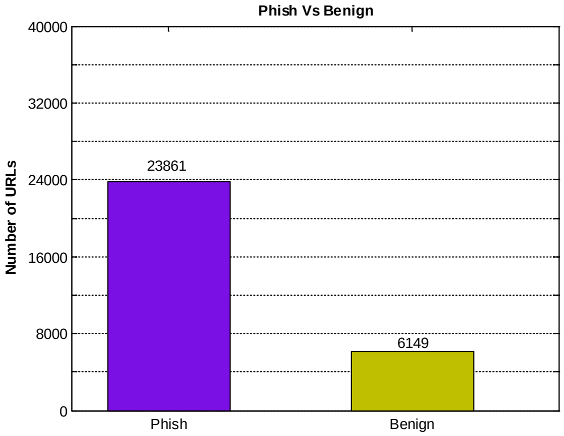


Fig. 6. Graphical Classification for 30,000 URLs.

Finally, Fig. 7 shows the graphical representation of phishing and benign values. A total number of 27629 URLs are phishing and 7371 are benign.

In order to provide further information about confidence intervals of URLs classification, each classifier runs for 100, 150, 200, and 250. Table V shows the 5th percentile, 95th percentile, median, and standard deviation (SD) values for the accuracy of each classifier.

More experiments were performed to ascertain which malicious schemes and attack methods are successful at tricking innocent Internet users to reveal personal information. We use 30 phishing features and randomly distributed them across 40 phishing URLs from our data sets. Thus, one phishing feature could be in many phishing URLs; similarly, one phishing URL could have one or more features. The results of the experiments are presented in Table VI.

We observe that the “Anomalous Request URL” featured in all the selected 40 phishing URLs having a 100% appearance. In addition, spelling errors are 85% having appeared 34. It shows that most of the messages sent by the attackers to innocent users have spelling errors. However, the "Disabling right-click button" feature has the highest percentage (7.5%) with 3 appearances. We ensured that every phishing feature had featured at least once in all the selected phishing URLs.

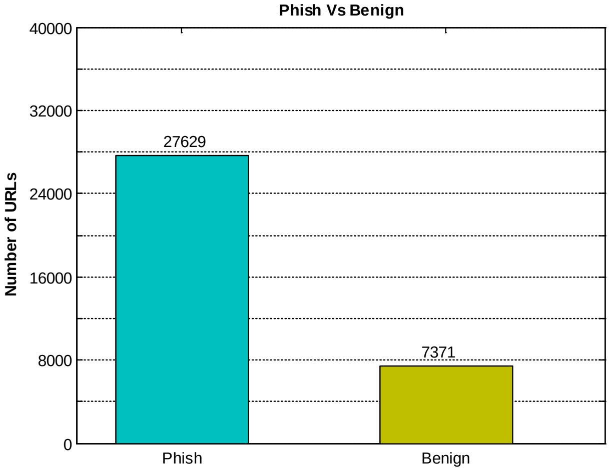


Fig. 7. Graphical Classification for 35,000 URLs.

TABLE V. CLASSIFICATION RESULTS FOR THE CLASSIFIERS

Number of Runs	Classifier	5th Percentile	95th Percentile	Median	SD
100	SVM	95.25	97.31	96.78	0.31
	Naïve Bayes	96.37	98.42	97.81	0.27
150	SVM	92.95	94.10	93.45	0.42
	Naïve Bayes	95.07	95.29	94.62	0.38
200	SVM	89.09	90.73	90.39	0.57
	Naïve Bayes	91.51	93.48	94.62	0.49
250	SVM	86.37	88.06	87.41	0.67
	Naïve Bayes	89.28	91.50	90.59	0.61

TABLE VI. PHISHING FEATURE INDICATORS

Lexical features	**No. of appearance**	**Percentage of appearance (%)**
IP-based URL	23	57.5
Long URL to hide the fake part	28	70.0
Shortened URL	7	17.5
URL’s having “@” Symbol	9	22.5
Using forms with the 'Submit' button	5	12.5
Hovering of a Mouse over Hyperlink	21	52.5
Spelling errors	34	85.0
Redirect pages	29	72.5
Anomalous Request URL	40	100.0
Domain name separated by a dash symbol	13	32.5
Subdomain of a subdomain	28	70.0
Copying Website	15	37.5
Anomalous cookie	7	17.5
Website Traffic	5	12.5
Domain name having multiple alphabets	11	27.5
1.1.1.1 Phishing website longevity	25	62.5
Generic salutation	31	77.5
Pharming attack	6	15.0
Using Non-Standard Port	18	45.0
URL of Anchor	14	35.0
Disabling right-click button	3	7.5
Adding Prefix or Suffix	12	30.0
Status Bar Customization	16	40.0
Age of Domain	23	57.5
Google Index	19	47.5
Server Form Handler (SFH)	5	12.5
Number of Links Pointing to Page	17	42.5
Using Hexadecimal Character Codes	12	30.0
Replacing Similar Characters for URL	21	52.5
Using the pop-up window	7	17.5

VI. CONCLUSION AND FUTURE WORK

Phishing is a type of social engineering attack often used to steal user personal information. In this project, we explore several tactics in which phishers use to trick innocent Internet users into divulging their personal information. We added new features to our design and in addition to some important features, we identified in the literature. An efficient approach is developed for detecting malicious URLs. Hybrid machine learning algorithms are used to classify our data sets. Several experiments were performed to determine the efficiency of our scheme. These experiments showed better performance and achieved a classification accuracy of 97.8% with a low false-positive rate of 1.06%.

In the future, we would consider more machine learning algorithms to compare their accuracy and false-positive rates.

ACKNOWLEDGMENTS

This work is supported by the Centre for Postgraduate Studies, Cape Peninsula University of Technology, Cape Town, South Africa.

DECLARATION OF INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

AUTHORS’ CONTRIBUTIONS

All authors contributed and approved the final manuscript.

DATA AVAILABILITY

The raw data of the IoT devices used to support the findings of this study are available from the corresponding author upon request.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining," Expert systems with applications, Vol.37, No.12, pp.7913-7921, 2010.

[2] M. Ajlouni, W. e. Hadi, and J. Alwedyan, "Detecting phishing websites using associative classification," image, Vol.5, No.23, pp.36-40, 2013.

[3] APWG. (2019, November 13,). Anti-Phishing Working Group Phishing Activity Trends Report. Available: https://docs.apwg.org/reports/apwg_trends_report_q2_2019.pdf

[4] R. B. Basnet, A. H. Sung, and Q. Liu, "Feature selection for improved phishing detection," in Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2012, pp.252-261.

[5] "Learning to detect phishing URLs," International Journal of Research in Engineering and Technology, Vol.3, No.6, pp.11-24, 2014.

[6] A. Bouveret, Cyber risk for the financial sector: a framework for quantitative assessment: International Monetary Fund, 2018.

[7] M. Büchi, N. Just, and M. Latzer, "Caring is not enough: the importance of Internet skills for online privacy protection," Information, Communication & Society, Vol.20, No.8, pp.1261-1278, 2017.

[8] J. A. Chaudhry, S. A. Chaudhry, and R. G. Rittenhouse, "Phishing attacks and defenses," International Journal of Security and Its Applications, Vol.10, No.1, pp.247-256, 2016.

[9] K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: their types, vectors, and technical approaches," Expert Systems with Applications, Vol.106, pp.1-20, 2018.

[10] L. M. Ellram and W. L. Tate, "The use of secondary data in purchasing and supply management (P/SM) research," Journal of purchasing and supply management, Vol.22, No.4, pp.250-254, 2016.

[11] M. Graham and W. H. Dutton, Society and the internet: How networks of information and communication are changing our lives: Oxford University Press, 2019.

[12] J. Hong, "The current state of phishing attacks," 2012.

[13] A. Jain and V. Richariya, "Implementing a web browser with phishing detection techniques," arXiv preprint arXiv:1110.0360, 2011.

[14] L. James, Phishing exposed. Canada.: Syngress, 2005.

[15] Kaspersky. (2019, November 25). How to protect yourself against spam email and phishing. Available: https://www.kaspersky.co.za/resource-center/threats/spam-phishing

[16] J. LaCour, "Phishing Trends and Intelligent Report," 2019.

[17] S. Le Page, G.-V. Jourdan, G. v. Bochmann, J. Flood, and I.-V. Onut, "Using url shorteners to compare phishing and malware attacks," in Proceedings of the 2018 APWG Symposium on Electronic Crime Research (eCrime), 2018, pp.1-13.

[18] L.-H. Lee, K.-C. Lee, H.-H. Chen, and Y.-H. Tseng, "Poster: Proactive blacklist update for anti-phishing," in Proceedings of the Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, 2014, pp.1448-1450.

[19] L. Li, E. Berki, M. Helenius, and S. Ovaska, "Towards a contingency approach with whitelist-and blacklist-based anti-phishing applications: what do usability tests indicate?," Behaviour & Information Technology, Vol.33, No.11, pp.1136-1147, 2014.

[20] S. Li and R. Schmitz, A novel anti-phishing framework based on honeypots: IEEE, 2009.

[21] M. A. Mahmood and L. Rajamani, "APD: ARM Deceptive Phishing Detector System Phishing Detection in Instant Messengers Using Data Mining Approach " Global Trends in Computing and Communication Systems, Vol.269, No.1, pp.490-502, 2011.

[22] Miniwatts Marketing Group. (2019, November 11,). Internet World Starts. Available: https://www.internetworldstats.com/stats.htm

[23] M. Moghimi and A. Y. Varjani, "New rule-based phishing detection method," Expert systems with applications, Vol.53, pp.231-242, 2016.

[24] PhishTank. (2019, November 25,). Statistics about phishing activity and PhishTank usage. Available: http://www.phishtank.com/stats.php

[25] G. Ramesh, I. Krishnamurthi, and K. S. S. Kumar, "An efficacious method for detecting phishing webpages through target domain identification," Decision Support Systems, Vol.61, pp.12-22, 2014.

[26] H. N. Security. (2019). Phishing attacks at highest level in three years. Available: https://www.helpnetsecurity.com/2019/11/07/phishing-attacks-levels-rise/

[27] E. Soegoto and M. Rafi, "Internet role in improving business transaction," in Proceedings of the IOP Conference Series: Materials Science and Engineering, 2018, pp.012059.

[28] V. Suganya, "A review on phishing attacks and various anti phishing techniques," International Journal of Computer Applications, Vol.139, No.1, pp.20-23, 2016.

[29] R. Verma and A. Das, "What's in a url: Fast feature extraction and malicious url detection," in Proceedings of the Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics, 2017, pp.55-63.

[30] A. Wang, W. Chang, S. Chen, and A. Mohaisen, "Delving into internet ddos attacks by botnets: Characterization and analysis," IEEE/ACM Transactions on Networking (TON), Vol.26, No.6, pp.2843-2855, 2018.

[31] S. Wedyan and F. Wedyan, "An Associative Classification Data Mining Approach for Detecting Phishing Websites," Journal of Emerging Trends in Computing and Information Sciences, Vol.4, No.12, 2013.

[32] D. Zhang, Z. Yan, H. Jiang, and T. Kim, "A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites," Information & Management, Vol.51, No.7, pp.845-853, 2014.

Biographies

Philip Abidoye received his M.Sc. from the University of Ibadan, Ibadan, Nigeria in 2006 and a Ph.D. degree from the University of the Western Cape, Cape Town, South Africa in 2015, both in Computer Science.

He is currently a Postdoctoral Fellow in the Department of Information Technology, Cape Peninsula University of Technology, Cape Town, South Africa. He has presented conference papers at international conferences, as well as published many papers in reputable international journals.

Dr. Abidoye is a member of the Institute of Electrical and Electronics Engineers (IEEE), South African Institute of Computer Scientists and Information Technologists (SAICSIT), and Computer Professionals Registration Council of Nigeria (CPN).

His research interests include secure wireless sensor networks, Cloud Computing security, security and privacy in the Internet of Things (IoT).

Boniface Kabaso received a Ph.D. degree in Information Technology from the Cape Peninsula University of Technology, Cape Town, South Africa. His research interests include software development, Internet of Things (IoT), soft computing, and Cloud Computing. Dr. Kabaso has published many research articles and conference papers in top-quality journals and conference proceedings.

Enhanced Pre-processing and Parameterization Process of Generic Code Clone Detection Model for Clones in Java Applications

Nur Nadzirah Mokhtar[1], Al-Fahim Mubarak-Ali[2], Mohd Azwan Mohamad Hamza[3]
Faculty of Computing, Universiti Malaysia Pahang
26300 Gambang, Pahang, Malaysia

***Abstract*—Code clones are repeated source code in a program. There are four types of code clone which are: Type 1, Type 2, Type 3 and Type 4. Various code clone detection models have been used to detect code clone. Generic Code Clone model is a model that consists of a combination of five processes in detecting code clone from Type-1 until Type-4 in Java Applications. The five processes are Pre-processing, Transformation, Parameterization, Categorization and Match Detection process. This work aims to improve code clone detection by enhancing the Generic Code Clone Detection (GCCD) model. Therefore, the Preprocessing and Parameterization process is enhanced to achieve this aim. The enhancement is to determine the best constant and weightage that can be used to improve the code clone detection result. The code clone detection result from the proposed enhancement shows that private with its weightage is the best constant and weightage for the Generic Code Clone Detection Model.**

***Keywords*—Code clone; code clone detection model; java applications; computational intelligence**

I. Introduction

Duplicated codes or better known as code clone are similar source codes that exist in a program [1-3]. Code clone brings maintenance issues in software. The more source codes are cloned in a program, the more memory and time needed in processing the software. At times, it also happens due to the software developer code writing practices [4]. Apart from that, if a source code contains bugs copied to the other parts of the software, the same bugs will be copied together throughout the program. This compromises the security of the software [3]. Code clone occurrence also depends on the deficiency of a programming language. As an instance, the Java programming language. Java is a worldwide open-sourced programming language used to develop open-source applications. In an experiment conducted to see the occurrence of code clones in Java applications, a total of 6% out of 512 000 lines of codes or 30 720 lines of codes from tested Java applications contains clones. One of the reasons for this occurrence is due to the absence of generic modules in Java [5].

At the initial stages of code clone detection, various approaches have been introduced. The approaches include text-based approach [6] [7], metric-based approach [8-10], tree-based approach [11-14], token-based approach [15-17] and graph-based approach [18-20]. The drawback of existing approaches is the lack of detecting all types of code clones [21]. In order to overcome this issue, code clone detection models were introduced to detect code clones that causes bad effect to the software. Code clone detection models is a model with combinatorial and structured processes that helps to detect and display detection result of code clone. Code clone detection models are recent development in the field of software clone and very little in terms of availability as tool, yet the existing code clone detection models have been a frontal movement in terms of having a combined process that detects code clone nevertheless of the diverse code clone jargons and programming languages.

As mentioned before, a model is an effort of unifying different processes to detect all code clone types. The effort can be seen through the Unified Clone Model [22] although this model is still in the design phase. Generic Pipeline Model [23-24] is a code clone detection model that detects on exact which is Type-1 and near-exact clones which is Type-2 in Java applications. An enhanced was proposed for this model by proposing a concatenation process, but it more focused on improving the time rather than improving the clone detection [25]. The disadvantage of this model is it only detect clones for Type-1 and Type-2. The state of the art model can be referred to as the Generic Code Clone Detection Model [26]. This model detects clones from Type-1 until Type-4 in Java applications. Type-3 refers to the source code that has modified semantically and Type-4 refers to the source code that has been modified further compared to Type-3.

This work focuses on improving code clone detection by enhancing the Generic Code Clone Detection Model through determining the best constant and weightage for Generic Code Clone Detection Model. Section 2 describes the Generic Code Clone Detection Model. Section 3 shows the implementation of the proposed enhancement while Section 4 discusses the findings of this work. This paper is summarized and concluded in Section 5.

II. Generic Code Clone Detection Model

Generic Code Clone Detection is a model that was designed and developed with an objective of detecting code clone from Java programming language [26]. It was designed into five processes which are elaborated in detail in upcoming sections. Fig. 1 illustrates the diagrammatic view of the model together with a brief narrative of the processes involved.

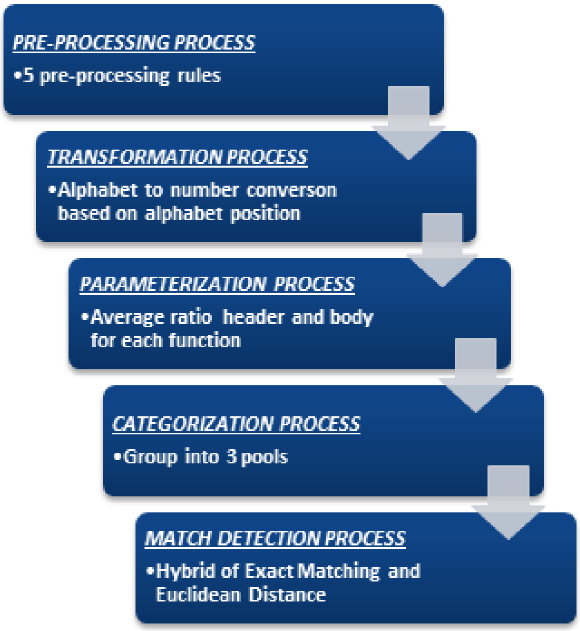


Fig. 1. Illustration of Generic Code Clone Detection Model [26].

A. *Pre-Processing Process [26]*

This model is initiated through this process. Source code alludes to the codes written in a source document of an application. It fills in as the contribution for the procedure. The source codes need to experience five joined rules used to accomplish the point of this process. Table I shows the rundown of these five rules. The yield of this process is standardized source codes or otherwise called source units. The source unit is still as source code. Each source unit speaks to a component of the source code.

B. *Transformation Process [26]*

This process is after the pre-processing process. The main objective of this process is changing the output of the previous process which is the source units into a more calculable format. The calculable format which is in the form of numbers are called as transformed source units and serves as the input in the determination of the parameters that will be used in the next process. The numerical form for this process is acquired from a letter to number substitution concept. The substitution is done based on the location of the alphabet. As an example, b is the third alphabet in the vocabulary sequence; therefore, b is changed to 02. This change is done on other alphabets.

The yield for this process is source units that has been transformed in number form. The source units that has been transformed are split to two branch which are the header (h) and body (b). Header refers to the transformed source unit that starts at the very first of the line of code and ends before the start of the body part of a transformed source unit. The body (b) is the body of a transformed source unit. As an instance in explaining the concept of a header (h) and body (b) in a Java function, assume a Java function named Function C with the written composition of:

```
public static void myMet ()
{
System.out.println("I love java");
}
```

After going through the initial pre-processing process, the source unit of Function C appear as:

public static void myMet systemoutprintln i love java

Therefore, the header and body of a function of Function C:

header (h): public static void myMet

body (b) : systemoutprintln i love java

C. *Parameterization Process*

This process starts after the transformation process. The transformed source units from the previous process serves as the input for this process. The attribute or parameter used for clone detection in this model is the average ratio for both header and body. Before demonstrating the step by step calculation for the average ratio header and body of a function, four important metrics is extracted from transformed source units. Table II shows attained metrics from the transformed source units.

To gain an average ratio of a function, the ratio of the header (h) and body (b) of the respective function must be gained initially. From the previous transformation process, the access modifier of all the function or method that has been changed to the value of public. Therefore, all the functions that has been changed to transformed source unit consist the equal access modifier value after going through the previous process. By using the value of public as the standard value, respective source units are divided with this standard value. It is done so that the header and body ratio value of each code of the transformed source unit is acquired. As an instance in calculating the average ratio for each transformed source unit, let's presume a transformed source unit contains a header, *TSUXa*, with body, *TSUXb*.

TABLE I. FIVE PRE-PROCESSING RULES

Pre-processing Rule	Description
PR-1	Import and package lines are excluded.
PR-2	Comment lines are excluded.
PR-3	Empty statements are excluded.
PR-4	Access modifier of a function is replaced with public.
PR-5	All the written source code lines is changed to lowercase format.

TABLE II. METRIC EXTRACTED FROM TRANSFORMED SOURCE UNITS

Metrics	Description
header code count	Total source code count in the header
body code count	Total source code count in the body
header ratio	header (h) ratio
body ratio	body (b) ratio
average header ratio	header (h) average ratio
average body ratio	body (b) average ratio

Therefore, the ratio of the transformed source unit is:

$$RA = \frac{(A1, A2, A3 \ldots An)}{P1} \quad (1)$$

$$RB = \frac{(B1, B2, B3 \ldots An)}{P1} \quad (2)$$

in which;

P1 is public access modifier value

RA is ratio value of header for each source units that has been transformed

RB is ratio value of body for each source units that has been transformed

A1, *A2 A3..An* is value of header in source units that has been transformed

B1, *B2*, *B3.. Bn* is value of body in source units that has been transformed

Once each function acquired the ratio of header and body, the next step is the calculation of the average ratio header and average ratio body of each transformed source unit. The formula of average ratio header and average ratio body in each transformed source units are:

$$AVRA = \frac{(RA)}{CA} \quad (3)$$

$$AVRB = \frac{(RB)}{CB} \quad (4)$$

in which:

AVRA is the value of average ratio for header in a transformed source unit

AVRB is the value of average ratio for body in a transformed source unit

CA is the total count of source code for header in a transformed source unit

CB is the total count of source code for body in a transformed source unit

The output of this process is the mentioned metrics; in which will be used in the next categorization process.

D. Categorization Process [26]

This process starts after parameterization process. The objective of this process is to pool the source units that has been transformed into a pool of code clones based on the exact ratio value of average ratio header and body for respective functions. This process uses metrics acquired from the parameterization process as input. The categorization is completed by grouping it into three pools using the average ratio value of the header and body of source units that has been transformed.

The first pool is for transformed source units for different functions that has the same value of header. As an instance, if transformed source unit for function E has the same header average ratio value with transformed source unit B, therefore these two transformed source units are categorized into the same group. This process will be continued until all the transformed source units that have the same average value of the header are categorized in the same pool. The second pool is for transformed source units for different functions that has the same value of body.

After the first pooling process, if transformed source unit for function E has the same body average ratio value with transformed source unit B, therefore these two transformed source units are grouped into the same category. This process will be continued until all the transformed source units that have the same average value of the body are categorized in the same pool. The remaining transformed source units is categorized into another category or better known as the third pool.

E. Match Detection Process [26]

This process comes after categorization process and it is the last process in this model. The main objective of this process is detecting code clone. The pool from the previous process is utilized as input for this process. Combination of exact matching and Euclidean Distance is used to detect code clone for this model. Exact matching is used on the first two pools to detect Type-1 and Type-2 clone. Once the detection is done for Type-1 and Type-2 from the first and second pools, the remaining transformed source units from the first and second pools is combined together with the third pool. As for the remaining average ratio header and body value in the third pool, Euclidean distance is used for Type-3 and Type-4 clone detection. As for the Euclidean distance application in this process, presume there are two transformed source units which are M and N. Therefore, the Euclidean distance, ED, between transformed source unit M and transformed source unit N is calculated as:

$$\text{EDMN} = (headerM - headerN)^2 + (bodyM - bodyN)^2 \quad (5)$$

where;

EDMN is Euclidean distance of transformed source unit M and N

headerM is the average ratio header value of M

bodyM is average ratio body value of M

headerN is average ratio header value of N

bodyN is average ratio body value of N

As for the remaining body and header value in the final pool, the mathematical equation which is the Euclidean distance is utilized. Once the equation is utilized upon the remaining average ratio values of the functions, all the functions is gathered to Type-3 and Type-4 depending on the distance value that is gained. Range 0f 0.85 to 1 is categorized as Type-3 while the rest is categorized as Type-4.

III. Proposed Enhancement

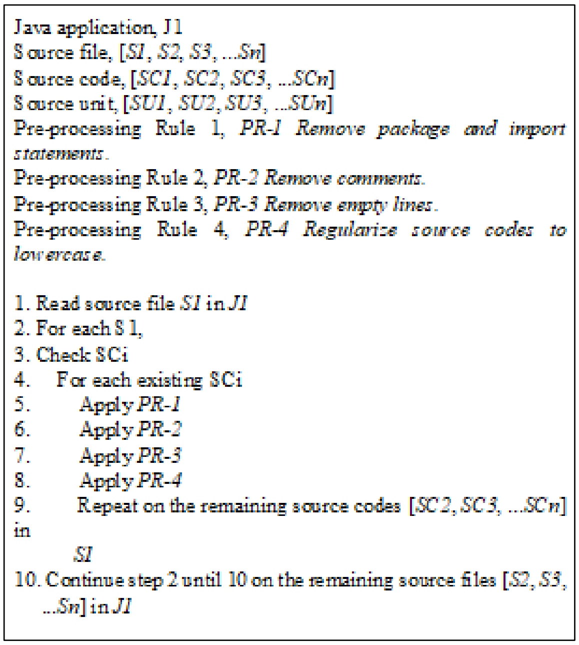
The enhancement of the Generic Code Clone Detection Model [26] is focused on two of its process which is Pre-processing and Parameterization Process.

A. *Enhancement on Pre-Processing Process*

Pre-processing is a process that normalizes source code to produce better source units to be processed for clone detection. The enhancement done in this process is the removal of function regularization rule; which is PR-4: Regularize function access keyword to public. This is to maintain the original function keyword of a function. Therefore, the enhanced pre-processing remains with four pre-processing rules. Fig. 2 shows the enhanced Pre-processing process is elaborated in the form of pseudo code.

B. *Enhancement on Parameterization Process*

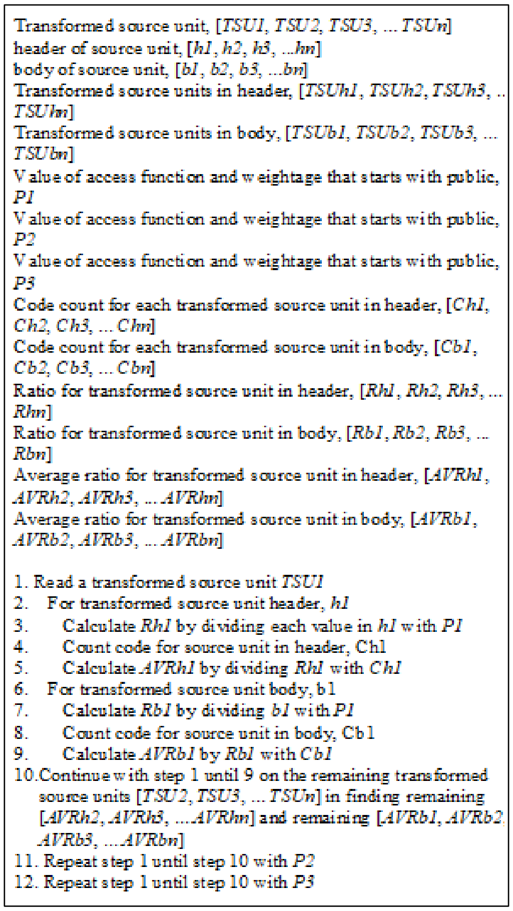
This process aims to create parameters or metrics that will be used for the categorization and match detection process. Therefore, the enhancement done in this process is the change of value access function based on three access functions and their respective weightage. The three access function is public with the weightage of 162102120903, private with the weightage of 16180922012005 and protected with the weightage of 161815200503200504. These values are based on the concept of the alphabet to number that has been explained in the Transformation Process. Fig. 3 shows the enhanced parameterization process is elaborated in the form of pseudo code.



```
Java application, J1
Source file, [S1, S2, S3, ...Sn]
Source code, [SC1, SC2, SC3, ...SCn]
Source unit, [SU1, SU2, SU3, ...SUn]
Pre-processing Rule 1, PR-1 Remove package and import statements.
Pre-processing Rule 2, PR-2 Remove comments.
Pre-processing Rule 3, PR-3 Remove empty lines.
Pre-processing Rule 4, PR-4 Regularize source codes to lowercase.

1. Read source file S1 in J1
2. For each S 1,
3. Check SCi
4.    For each existing SCi
5.        Apply PR-1
6.        Apply PR-2
7.        Apply PR-3
8.        Apply PR-4
9.        Repeat on the remaining source codes [SC2, SC3, ...SCn] in
          S1
10. Continue step 2 until 10 on the remaining source files [S2, S3,
    ...Sn] in J1
```

Fig. 2. Enhanced Pre-Processing Process Pseudo Code.



```
Transformed source unit, [TSU1, TSU2, TSU3, ... TSUn]
header of source unit, [h1, h2, h3, ...hn]
body of source unit, [b1, b2, b3, ...bn]
Transformed source units in header, [TSUh1, TSUh2, TSUh3, ..
TSUhn]
Transformed source units in body, [TSUb1, TSUb2, TSUb3, ...
TSUbn]
Value of access function and weightage that starts with public,
P1
Value of access function and weightage that starts with public,
P2
Value of access function and weightage that starts with public,
P3
Code count for each transformed source unit in header, [Ch1,
Ch2, Ch3, ... Chn]
Code count for each transformed source unit in body, [Cb1,
Cb2, Cb3, ... Cbn]
Ratio for transformed source unit in header, [Rh1, Rh2, Rh3, ...
Rhn]
Ratio for transformed source unit in body, [Rb1, Rb2, Rb3, ...
Rbn]
Average ratio for transformed source unit in header, [AVRh1,
AVRh2, AVRh3, ... AVRhn]
Average ratio for transformed source unit in body, [AVRb1,
AVRb2, AVRb3, ... AVRbn]

1. Read a transformed source unit TSU1
2.    For transformed source unit header, h1
3.        Calculate Rh1 by dividing each value in h1 with P1
4.        Count code for source unit in header, Ch1
5.        Calculate AVRh1 by dividing Rh1 with Ch1
6.    For transformed source unit body, b1
7.        Calculate Rb1 by dividing b1 with P1
8.        Count code for source unit in body, Cb1
9.        Calculate AVRb1 by Rb1 with Cb1
10. Continue with step 1 until 9 on the remaining transformed
    source units [TSU2, TSU3, ... TSUn] in finding remaining
    [AVRh2, AVRh3, ....AVRhn] and remaining [AVRb1, AVRb2,
    AVRb3, ....AVRbn]
11. Repeat step 1 until step 10 with P2
12. Repeat step 1 until step 10 with P3
```

Fig. 3. Enhanced Parameterization Process Pseudo Code.

IV. Result and Discussion

This section is divide into three subsections. The first subsection describes the result of the overall clone pair for Java applications from Bellon's benchmark data [27]. The second subsection describes the result of the overall clone pair based on the clone type for Java applications from Bellon's benchmark data [27]. The third subsection discusses the obtained results.

A. *Overall Clone Pair in Java Application*

Fig. 4 shows the overall clone pair for Java applications from Bellon's benchmark data. Based on Fig. 4, the highest overall clone pair detected for Eclipse-ant is from protected weightage with 7681 clone pairs. The second highest overall clone pair detected for Eclipse-ant is from private weightage with 4454 clone pairs. It is 42% lower compared to the overall clone pairs detected from the protected weightage, that is, the highest overall clone pair detected for Eclipse-ant. The third overall clone pair detected for Eclipse-ant is from the existing GCCD with 2688 clone pairs. It is 65% lower compared to the overall clone pairs detected from the protected weightage, that is, the highest overall clone pair detected for Eclipse-ant. The lowest overall clone pair detected for Eclipse-ant is from public weightage with 2654 clone pairs. It is 65.4% lower compared to the overall clone pairs detected from the protected weightage, that is, the highest overall clone pair detected for Eclipse-ant.

As for the Eclipse-jdtcore application, the highest overall clone pair detected for Eclipse-jdtcore is from protected weightage with 39974 clone pairs. The second highest overall clone pair detected for Eclipse-jdtcore is from private weightage with 15406 clone pairs. It is 61.5% lower compared to the overall clone pairs detected from the protected weightage, that is, the highest overall clone pair detected for Eclipse-jdtcore. The third overall clone pair detected for Eclipse-jdtcore is from the existing GCCD with 11268 clone pairs. It is 71.8% lower compared to the overall clone pairs detected from the protected weightage, that is, the highest overall clone pair detected for Eclipse-jdtcore. The lowest overall clone pair detected for Eclipse-jdtcore is from public weightage with 10767 clone pairs. It is 73.1% lower compared to the overall clone pairs detected from the protected weightage, that is, the highest overall clone pair detected for Eclipse-jdtcore.

As for the j2sdk1.4.0-javax-swing application, the highest overall clone pair detected for j2sdk1.4.0-javax-swing is from protected weightage with 56312 clone pairs. The second highest overall clone pair detected for j2sdk1.4.0-javax-swing is from private weightage with 8993 clone pairs. It is 84% lower compared to the overall clone pairs detected from the protected weightage, that is, the highest overall clone pair detected for j2sdk1.4.0-javax-swing. The third overall clone pair detected for j2sdk1.4.0-javax-swing is from public weightage with 7393 clone pairs. It is 86.9% lower compared to the overall clone pairs detected from the protected weightage, that is, the highest overall clone pair detected for j2sdk1.4.0-javax-swing. The lowest overall clone pair detected for j2sdk1.4.0-javax-swing is from the existing GCCD with 7281 clone pairs. It is 87.1% lower compared to the overall clone pairs detected from the protected weightage, that is, the highest overall clone pair detected for j2sdk1.4.0-javax-swing.

As for the Netbeans-javadoc application, the highest overall clone pair detected for Netbeans-javadoc is from private weightage with 937 clone pairs. The second highest overall clone pair detected for Netbeans-javadoc is from protected weightage with 654 clone pairs. It is 30.2% lower compared to the overall clone pairs detected from the private weightage; which is the highest overall clone pair detected for Netbeans-javadoc. The lowest overall clone pair detected for Netbeans-javadoc is from the existing GCCD and the public weightage with 595 clone pairs. It is 36.5% lower compared to the overall clone pairs detected from the private weightage; which is the highest overall clone pair detected for Netbeans-javadoc. The next subsection discusses the overall clone pair based clone type for each Java application from the Bellon benchmark data.

B. Overall Clone Pair based on Clone Type

Table III shows the overall clone pair based on the clone type for Java applications from Bellon benchmark data. As for Eclipse-ant application, the highest number of Type-1 clone pairs in Eclipse-ant was produced through the protected weightage which is 424 clone pairs. The second highest number of Type-1 clone pairs in Eclipse-ant was produced through the private weightage which is 246 clone pairs. The existing GCCD produced 185 clone pairs for Type-1; which is the same as the enhancement of the GCCD done using public weightage. This is the lowest amount of clone pair for Type-1 in Eclipse-ant. As for Type-2 clone in Eclipse- ant, the highest Type-2 clone pair in Eclipse-ant was produced through protected weightage with 916 clone pairs. The second highest Type-2 clone pair in Eclipse-ant was produced through private weightage with 750 clone pairs. The third highest Type-2 clone pair in Eclipse-ant was produced through protected weightage with 648 clone pairs. The lowest Type-2 clone pair in Eclipse-ant was produced through the existing GCCD with 552 clone pairs. As for Type-3 clone in Eclipse- ant, the highest Type-3 clone pair in Eclipse-ant was produced through protected weightage with 2296 clone pairs. The second highest Type-3 clone pair in Eclipse-ant was produced through private weightage with 2481 clone pairs. The third highest Type-3 clone pair in Eclipse-ant was produced through the existing GCCD with 581 clone pairs. The lowest Type-3 clone pair in Eclipse-ant was produced through the public weightage with 578 clone pairs. As for Type-4 clone in Eclipse-ant, the highest Type-4 clone pair in Eclipse-ant was produced through protected weightage with 4225 clone pairs. The second highest Type-4 clone pair in Eclipse-ant was produced through the existing GCCD with 1370 clone pairs. The third highest Type-4 clone pair in Eclipse-ant was produced through the public weightage with 1243 clone pairs. The lowest Type-4 clone pair in Eclipse-ant was produced through the private weightage with 977 clone pair.

As for the Eclipse-jdtcore application, the highest Type-1 clone pair in Eclipse-jdtcore was produced through protected weightage with 1008 clone pairs. The second highest Type-1 clone pair in Eclipse-jdtcore was produced through the private weightage with 766 clone pairs. The third highest Type-1 clone pair in Eclipse-jdtcore was produced through the public weightage with 627 clone pairs. The lowest Type-1 clone pair in Eclipse-ant was produced through the existing GCCD with 626 clone pairs. As for Type-2 clone in Eclipse-jdtcore, the highest Type-2 clone pair in Eclipse-ant was produced through protected weightage with 2952 clone pairs. The second highest Type-2 clone pair in Eclipse-jdtcore was produced through the existing GCCD with 2886 clone pairs. The third highest Type-2 clone pair in Eclipse-jdtcore was produced through the public weightage with 2882 clone pairs. The lowest Type-2 clone pair in Eclipse-jdtcore was produced through the private weightage with 2660 clone pairs. As for Type-3 clone in Eclipse-jdtcore, the highest Type-3 clone pair in Eclipse-jdtcore was produced through protected weightage with 22854 clone pairs. The second highest Type-3 clone pair in Eclipse-jdtcore was produced through the private weightage with 9634 clone pairs. The third highest Type-3 clone pair in Eclipse-jdtcore was produced through the existing GCCD with 4265 clone pairs. The lowest Type-3 clone pair in Eclipse-jdtcore was produced through the public weightage with 3866 clone pairs. As for Type-4 clone in Eclipse-jdtcore, the highest Type-4 clone pair in Eclipse-jdtcore was produced through protected weightage with 13169 clone pairs. The second highest Type-4 clone pair in Eclipse-jdtcore was produced through the existing GCCD with 3491 clone pairs. The third highest Type-4 clone pair in Eclipse-jdtcore was produced through the public weightage with 3392 clone pairs. The lowest Type-4 clone pair in Eclipse-jdtcore was produced through the private weightage with 2346 clone pairs.

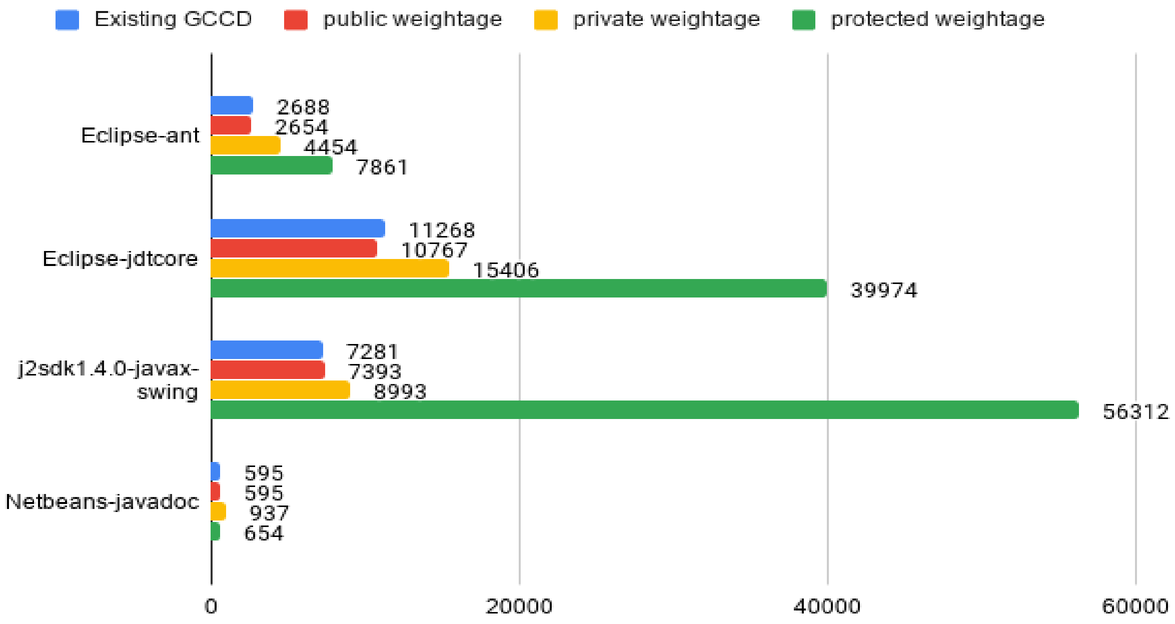


Fig. 4. Overall Clone Pair for Java Applications from Bellon Benchmark Data.

TABLE III. OVERALL CLONE PAIR BASED ON CLONE TYPE FOR JAVA APPLICATIONS FROM BELLON BENCHMARK DATA

Java Application	**Clone Type**	**Existing GCCD**	**public weightage**	**private weightage**	**protected weightage**
Eclipse-ant	Type-1	185	185	246	424
	Type-2	552	648	750	916
	Type-3	581	578	2481	2296
	Type-4	1370	1243	977	4225
Eclipse-jdtcore	Type-1	626	627	766	1008
	Type-2	2886	2882	2660	2952
	Type-3	4265	3866	9634	22845
	Type-4	3491	3392	2346	13169
j2sdk1.4.0-javax-swing	Type-1	877	891	1021	1330
	Type-2	3697	3713	3709	4259
	Type-3	1710	1774	1977	27316
	Type-4	997	1015	2286	23407
Netbeans-javadoc	Type-1	99	99	120	182
	Type-2	341	341	393	425
	Type-3	102	102	304	11
	Type-4	53	53	120	36

As for the j2sdk1.4.0-javax-swing application, the highest Type-1 clone pair in j2sdk1.4.0-javax-swing was produced through protected weightage with 1330 clone pairs. The second highest Type-1 clone pair in j2sdk1.4.0-javax-swing was produced through the private weightage with 1021 clone pairs. The third highest Type-1 clone pair in j2sdk1.4.0-javax-swing was produced through the public weightage with 891 clone pairs. The lowest Type-1 clone pair in j2sdk1.4.0-javax-swing was produced through the existing GCCD weightage with 877 clone pairs. As for Type-2 clone in j2sdk1.4.0-javax-swing, the highest Type-2 clone pair in j2sdk1.4.0-javax-swing was produced through protected weightage with 4259 clone pairs. The second highest Type-2 clone pair in j2sdk1.4.0-javax-swing was produced through the public weightage with 3713 clone pairs. The third highest Type-2 clone pair in j2sdk1.4.0-javax-swing was produced through the private weightage with 3709 clone pairs. The lowest Type-2 clone pair in j2sdk1.4.0-javax-swing was produced through the existing GCCD with 3697 clone pairs. As for Type-3 clone in j2sdk1.4.0-javax-swing, the highest Type-3 clone pair in j2sdk1.4.0-javax-swing was produced through protected weightage with 27316 clone pairs. The second highest Type-3 clone pair in j2sdk1.4.0-javax-swing was produced through the private weightage with 1977 clone pairs. The third highest Type-3 clone pair in j2sdk1.4.0-javax-swing was produced through the public weightage with 1774 clone pairs. The lowest Type-3 clone pair in j2sdk1.4.0-javax-swing was produced through the existing GCCD with 1710 clone pairs. As for Type-4 clone in j2sdk1.4.0-javax-swing, the highest Type-4 clone pair in j2sdk1.4.0-javax-swing was produced through protected weightage with 23407 clone pairs. The second highest Type-4 clone pair in j2sdk1.4.0-javax-swing was produced through the private weightage with 2286 clone pairs. The third highest Type-4 clone pair in j2sdk1.4.0-javax-swing was produced through the public weightage with 1015 clone pairs. The lowest Type-4 clone pair in j2sdk1.4.0-javax-swing was produced through the existing GCCD with 997 clone pairs.

As for the Netbeans-javadoc application, the highest Type-1 clone pair in Netbeans-javadoc was produced through protected weightage with 182 clone pairs. The second highest Type-1 clone pair in Netbeans-javadoc was produced through the private weightage with 120 clone pairs. The lowest Type-1 clone pair in Netbeans-javadoc was produced through the private weightage and the existing GCCD with 99 clone pairs. As for Type-2 clone in Netbeans-javadoc, the highest Type-2 clone pair in Netbeans-javadoc was produced through protected weightage with 425 clone pairs. The second highest Type-2 clone pair in Netbeans-javadoc was produced through the private weightage with 393 clone pairs. The lowest Type-2 clone pair in Netbeans-javadoc was produced through the private weightage and the existing GCCD with 341 clone pairs. As for Type-3 clone in Netbeans-javadoc, the highest Type-3 clone pair in Netbeans-javadoc was produced through private weightage with 304 clone pairs. The second highest Type-3 clone pair in Netbeans-javadoc was produced through the public weighthage and the existing GCCD with 102 clone pairs. The lowest Type-3 clone pair in Netbeans-javadoc was produced through the protected weightage with 11 clone pairs. As for Type-4 clone in Netbeans-javadoc, the highest Type-4 clone pair in Netbeans-javadoc was produced through private weightage with 120 clone pairs. The second highest Type-4 clone pair in Netbeans-javadoc was produced through the public weightage and the existing GCCD with 53 clone pairs. The lowest Type-4 clone pair in Netbeans-javadoc was produced through the protected weightage with 36 clone pairs.

C. Discussion

The main aim of this work is to improve the code clone detection for Java applications by enhancing the Pre-processing and Parameterization process in the Generic Code Clone Detection Model. The pre-processing rule has been reduced from five rules to four rules by removing the regularization of function access modifiers. After that, the Parametrization process was enhanced with three different access functions and weightage. The three access functions are public with the weightage of 162102120903, private with the weightage of 16180922012005 and protected with the weightage of 161815200503200504. These values are based on the concept of the alphabet to number that has been explained in the Transformation Process. The result from these changes has been described in subsection 4.1 and subsection 4.3. Based on the result shown, the protected with the weightage of 161815200503200504 has shown more clone pair detection in three Java applications compared to the other success function. The three Java applications are Eclipse-ant, Eclipse-jdtcore and j2sdk1.4.0-javax-swing. The remaining Java application which is Netbeans-javadoc has more clone pair revealed through private with the weightage of 16180922012005 but has the second most clone pair detected through protected with the weightage of 161815200503200504.

This happens due to the enhancement made to the GCCD model. First is the removal keyword regularization rule from the pre-processing process. As mentioned previously, the pre-processing process of the GCCD at the start does the process of removing source code from uninteresting information. The uninteresting information is removed through the five rules previously that had been adopted in this process. The rules include removing packages and import statements, removing comments, removing empty lines, regularizing function access keyword to public and regularizing source codes to lowercase. These rules were set after taking into consideration in not making many changes to the original source codes. Too many changes in the source codes may cause critical information to be changed or removed; therefore, keeping a minimum set of rules ensures the most of the information of the source code such as the source code line and length is intact. The idea of removing the keyword regularization rule is to minimize the change of a function by sustaining original source code of a function. Furthermore, the different weightage of a constant influence the result. Based on the result, the higher the weightage value, the more clones are detected.

V. Conclusion

The idea of this work is to improve code clone detection in Java applications by enhancing the Generic Code Clone Detection Model. The enhancement involves by enhancing the Pre-processing and Parameterization Process. Based on the result shown, it can be concluded that the best constant and weightage for Generic Code Clone Detection Model is protected with a weightage value of 161815200503200504.

Acknowledgment

The authors would like express gratitude to Ministry of Higher Education Malaysia and Universiti Malaysia Pahang in supporting this work through the Fundamental Research Grant Scheme (FRGS Grant ID: RACER/1/2019/ICT01/UMP//1).

References

[1] J. Yang, Y. Xiong and J. Ma, "A function level Java code clone detection method," 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 2019, pp. 2128-2134.

[2] D. K. Kim, “Enhancing code clone detection using control flow graphs, ” International Journal of Electrical and Computer Engineering (IJECE), Vol.9, No.5, October 2019, pp. 3804~3812.

[3] M. S. Rahman and C. K. Roy, "On the Relationships Between Stability and Bug-Proneness of Code Clones: An Empirical Study," 2017 IEEE 17th International Working Conference on Source Code Analysis and Manipulation (SCAM), Shanghai, 2017, pp. 131-140.

[4] M. Pyl, B. van Bladel and S. Demeyer, "An Empirical Study on Accidental Cross-Project Code Clones," 2020 IEEE 14th International Workshop on Software Clones (IWSC), London, ON, Canada, 2020, pp. 33-37.

[5] M. Dagenais, J. F. F. Patenaude, E. Merlo, and B. Lagu¨e, “Clones occurrence in Java and Modula-3 software systems,” Advances in Software Engineering, 2002, pp. 95–110.

[6] E. Kodhai, S. Kanmani, A. Kamatchi, R. Radhika and B. V. Saranya, "Detection of Type-1 and Type-2 Code Clones Using Textual Analysis and Metrics," 2010 International Conference on Recent Trends in Information, Telecommunication and Computing, Kochi, Kerala, 2010, pp. 241-243.

[7] A. Marcus and J. I. Maletic, "Identification of high-level concept clones in source code," Proceedings 16th Annual International Conference on Automated Software Engineering (ASE 2001), San Diego, CA, USA, 2001, pp. 107-114.

[8] Vishwachi and S. Gupta, "Detection of near-miss clones using metrics and Abstract Syntax Trees," 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, 2017, pp. 230-234.

[9] Z. Li and J. Sun, "An iterative, metric space based software clone detection approach," The 2nd International Conference on Software Engineering and Data Mining, Chengdu, 2010, pp. 111-116.

[10] M. Sudhamani and L. Rangarajan, "Code clone detection based on order and content of control statements," 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), Noida, 2016, pp. 59-64.

[11] Y. Yang, Z. Ren, X. Chen and H. Jiang, "Structural Function Based Code Clone Detection Using a New Hybrid Technique," 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Tokyo, 2018, pp. 286-291.

[12] L. Büch and A. Andrzejak, "Learning-Based Recursive Aggregation of Abstract Syntax Trees for Code Clone Detection," 2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER), Hangzhou, China, 2019, pp. 95-104.

[13] J. Zhang, X. Wang, H. Zhang, H. Sun, K. Wang and X. Liu, "A Novel Neural Source Code Representation Based on Abstract Syntax Tree," 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE), Montreal, QC, Canada, 2019, pp. 783-794.

[14] H. Yu, W. Lam, L. Chen, G. Li, T. Xie and Q. Wang, "Neural Detection of Semantic Code Clones Via Tree-Based Convolution," 2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC), Montreal, QC, Canada, 2019, pp. 70-80.

[15] T. Kamiya, S. Kusumoto, and K. Inoue, “CCFinder: a multilinguistic tokenbased code clone detection system for large scale source code,” IEEE Transactions on Software Engineering. 28(7), pp. 654–670.

[16] W. Toomey, "Ctcompare: Code clone detection using hashed token sequences," 2012 6th International Workshop on Software Clones (IWSC), Zurich, 2012, pp. 92-93.

[17] P. Wang, J. Svajlenko, Y. Wu, Y. Xu and C. K. Roy, "CCAligner: A Token Based Large-Gap Clone Detector," 2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE), Gothenburg, 2018, pp. 1066-1077.

[18] C. M. Kamalpriya and P. Singh, "Enhancing program dependency graph based clone detection using approximate subgraph matching," 2017 IEEE 11th International Workshop on Software Clones (IWSC), Klagenfurt, 2017, pp. 1-7.

[19] Y. Higo, U. Yasushi, M. Nishino and S. Kusumoto, "Incremental Code Clone Detection: A PDG-based Approach," 2011 18th Working Conference on Reverse Engineering, Limerick, 2011, pp. 3-12.

[20] P. Gautam and H. Saini, "Type-2 software cone detection using directed acyclic graph," 2017 Fourth International Conference on Image Information Processing (ICIIP), Shimla, 2017, pp. 1-4.

[21] J. Harder, "The limits of clone model standardization," 2013 7th International Workshop on Software Clones (IWSC), San Francisco, CA, 2013, pp. 10-11.

[22] C. J. Kapser, J. Harder and I. Baxter, "A common conceptual model for clone detection results," 2012 6th International Workshop on Software Clones (IWSC), Zurich, 2012, pp. 72-73.

[23] B. Biegel and S. Diehl, “Highly Configurable and Extensible Code Clone Detection,”. 2010 17th Working Conference on Reverse Engineering, 2010, pp. 237–241.

[24] B. Biegel and S. Diehl, “JCCD: a flexible and extensible API for implementing custom code clone detectors,” Proceedings of the 2010b Proceedings of the IEEE/ACM International Conference on Automated Software Engineering (ASE '10), 2010, pp. 167-168.

[25] A. Mubarak-Ali, S. Sulaiman and S. M. Syed-Mohamad, "An enhanced generic pipeline model for code clone detection," 2011 Malaysian Conference in Software Engineering, Johor Bahru, 2011, pp. 434-438.

[26] A. Mubarak-Ali and S. Sulaiman, “Generic Code Clone Detection Model for Java Applications,” IOP Conference Series: Materials Science and Engineering, Volume 769, The 6th International Conference on Software Engineering & Computer Systems, 25-27 September 2019, Pahang, Malaysia.

[27] S. Bellon, R. Koschke, G. Antoniol, J. Krinke and E. Merlo, "Comparison and Evaluation of Clone Detection Tools," in IEEE Transactions on Software Engineering, vol. 33, no. 9, pp. 577-591, Sept. 2007.

Improving Intrusion Detection System using Artificial Neural Network

Marwan Ali Albahar[1], Muhammad Binsawad[2], Jameel Almalki[3], Sherif El-etriby[4], Sami Karali[5]
Umm Al Qura University[1,3,5], King Abdulaziz University[2], Menoufia University[4]

Abstract—**Currently, network communication is more susceptible to different forms of attacks due to its expanded usage, accessibility, and complexity in most areas, consequently imposing greater security risks. One method to halt attacks is to identify different forms of irregularities in the data transmitted and processed during communication. Detection of anomalies is a vital process to secure a system. To this end, machine learning plays a key role in identifying abnormalities and intrusion in communication over a network. The term regularization is one of the major aspects of training machine learning models, in which, it plays a primary role in several successful Artificial neural network models, by inducing regularization in the model training. Then, this technique is integrated with an Artificial Neural Network (ANN) for classifying and detecting irregularities in network communication efficiency. The purpose of regularization is to discourage learning a more flexible or complex model. Thus, the machine learning model generalizes enough to perform accurately on unseen data. For training and testing purposes, NSL-KDD, CIDDS-001 (External and Internal Server Data), and UNSW-NB15 datasets were utilized. Through extensive experiments, the proposed regularizer reaches higher True Positive Rate (TPR) and precision compared L1 and L2 norm regularization algorithms. Thus, it is concluded that the proposed regularizer demonstrates a strong intrusion detection ability.**

Keywords*—*New regularizer; anomaly detection; NSL-KDD dataset; CIDDS-001 dataset; UNSW-NB15

I. Introduction

Now-a-days, network communication's threats and attacks are growing as it is widely utilized in every field. To prevent such attacks, it is a crucial and necessary task to classify network communication as normal and suspicious. Such task is generally known as anomaly detection, dealing with unlikely events in network communication. The standard approach to detect an anomaly is computing the accurate mathematical model of normal data. Every new receiving instance is compared with the model of normality and, accordingly, an anomaly score is computed. The score will describe the deviations of the new instance compared to the average data instance and, if the deviation is relatively high, then the instance will be considered as suspicious and classified as anomalous and hence processed adequately [1] [2] [3].

In machine learning, generally, we are looking for the best-fitting model among other models in a large solution space. Similarly, in the context of ANN, solution space is defined as the space of all approximated or precise functions that a network can represent.

Network depth and activation functions are used to determine the size of this solution space. One hidden layer with an activation function makes the space of functions very huge, so this space grows exponentially when the depth of the network is increased; hence, finding a most-fit solution becomes a difficult task.

Multiple optimizer functions tend to minimize the loss function, of which Stochastic Gradient Descent (SGD) being very common. Using SGD as an optimizer, one can seek a solution by moving in the opposite direction of the gradient of loss function. Due to complexity and richness of the solution space, this method of learning might overfit the learning model and affects the generalization error or performance significantly on unseen data while giving good results on training data [4]. To solve this issue, the concept of regularization is introduced in machine learning to avoid the complexity of the learning model. There are different regularization algorithms used to avoid overfitting of the machine learning model [4]. For example, in iterative learning, the most common regularization algorithm is early stopping, and, in the neural network, the commonly used regularization algorithm is a dropout. Generally, in statistics and machine learning, the regularization term is used in combination with the loss or error function. This method is beneficial as it incorporates the model complexity into the function to be minimized. Such methods are used in many algorithms such as Support Vector Machines (SVMs) [5] as optimization problems.

However, the existing regularization algorithms come with drawbacks due to the nature of the regularizers. In a challenging setting, where the number of features is greater than the number of samples and correlated, the existing regularization algorithms either do not promote sparsity or poorly perform because of the absence of relevant information.

The purpose of this paper is to implement a new regularization algorithm to search for the optimal solution in a large solution space by taking into consideration the relationship between weight matrix entries. Hence, the limit of space is increased and can be controlled by squeezing and expanding this space based on the penalty term λ. Consequently, it provides the ability to find the least complex learning model. To differentiate between normal and various malicious connections, we plan to examine the algorithm from a multiclass classification perspective.

In this paper, we introduced new regularization design considerations and a general outline of an intrusion detection technique based on using the standard deviation to decay the weight matrices in order to get the regularization term. Compared with well-known regularization techniques, we embedded the proposed regularizer with ANN model for classification tasks and employed NSL-KDD, CIDDS-001 (External and Internal Server Data), and UNSW-NB15 datasets with separate testing and training sets to evaluate the efficiency in detecting anomalies.

The main contributions of this paper are summarized as follows:

1) We present the design and implementation of an ANN intrusion detection system based on a new regularizer.
2) We study the performance of the model with different regularization parameters impacting accuracy.

The outline of the paper is as follows: We provide the related works in Section II. Then, we give background and formalization in Section III. Next, we present the used datasets in Section IV. In the same section, we also study anomaly detection and the new regularization technique. In Section V, results and discussion are presented. The limitation of proposed regularization technique provided in Section VI. Finally, we conclude our study in Section VII.

II. Related Works

The first IDS or anomaly detection system was introduced by Dr. Dorothy in SRI international, and it is still an actively and heavily researched topic due to its broad applications in network communication ([6],[7]). Supervised learning techniques are popular methods for solving such problems. These techniques give more satisfactory results when statistical and regression techniques are incorporated [21].

A novel intrusion system and a multilevel hybrid classifier were proposed in [8]. The proposed system is combined the unsupervised Bayesian clustering with the supervised tree classifiers to detect the intrusions. Based on the Modular Multiple Classifier System (MCS), the authors in [9] proposed an unlabeled network anomaly IDS, where every module was created to model network services or a specific group of similar protocols. Moreover, they conducted experimental studies on the KDD Cup 1999 dataset, which revealed that the proposed anomaly IDS was able to accomplish high attack detection along with the low rate of false alarm. In [10], authors developed an intrusion detection system based on the AdaBoost algorithm. Within this algorithm, the decision rules were provided for the continuous and categorical features and the decision stumps were used as weak classifiers. The combination of the weak classifiers for the continuous and categorical features with the strong classifier allowed handling the relation between these features without the need for any forced conversations. According to the authors' experimental analysis, they reported that the algorithm had low error rates and computational complexity. Data mining techniques were utilized in [11]; the authors devised a novel framework for intrusion detection accordingly. For building classifiers, the authors proposed a classification algorithm which uses fuzzy association rules. However, the outcomes regarding the unseen attacks were not promising. In [12], the authors used a supervised learning classifier system for intrusion detection. To learn signatures for network intrusion detection, they presented a biologically inspired computational approach which can learn adaptively and dynamically. For the futuristic establishment of the intrusion detection system, authors in [13] presented a reference for the comparison of the efficiency of different machine learning techniques, including SVM and the tree classification. Moreover, the authors proposed a method to compute the mean value through sampling different ratios within the normal data for every measurement, resulting in obtaining a better rate of accuracy when observing the data in the real world. A novel machine-learning algorithm was proposed in [14], namely, Boosted Subspace Probabilistic Neural Network (BSPNN), which combined a semiparametric and an adaptive boosting approach to attain better trade-off between the generality and the accuracy. Hence, the method depicted prominent improvements with respect to detection accuracy, comparatively low computational complexity, and negligible false alarms. A new approach for intrusion detection was proposed in [15]. This approach is based on ANN and fuzzy clustering (FC-ANN). To evaluate the proposal, the authors conducted an experiment using the KDD Cup 1999 dataset. Experimental results demonstrated that FC-ANN enhanced the detection stability and the detection precision. For the prediction of the anomaly detection, a random-effects logistic regression model was proposed in [16].

Imbalanced class distribution is an inevitable problem in real network traffic due to the large size of traffic and low frequency of certain types of anomalies. Authors in [17] used sampling approaches to combat imbalanced class distributions for network intrusion detection. It performed flow-based classification on a network flow dataset: CIDDS-001. The system was able to detect attacks with up to 99.99% accuracy.

In [18], the statistical and complexity analysis of CIDDS-001 dataset is considered. The authors utilized the k-nearest neighbor classifier on CIDDS-001 to build an IDS. Their system achieved an overall accuracy of 99.6% with 2nn and a minimum accuracy of 99.3% with 5nn. Using the same dataset, the authors in [19] conducted an analytical study to assess the performance of KNN and k-means clustering algorithms when classifying traffic. Both algorithms achieved over 99% accuracy. In [20], authors proposed an effective anomaly-based intrusion detection system using a gradient boosted machine (GBM). Three different datasets, NSL-KDD, UNSW-NB15, and GPRS dataset, were utilized with either tenfold cross-validation or hold-out method. In [21], the authors proposed an improved IDS based on hybrid feature selection and two-level classifier ensembles. Two intrusion datasets (NSL-KDD and UNSW-NB15) have been employed to evaluate the performance. Based on the statistics and significance tests, on the NSL-KDD dataset, the proposed classifier shows 85.8% accuracy, 86.8% sensitivity, and 88.0% detection rate. By taking advantage of the multiple classification abilities of neural networks and the fuzzy logic, authors in [22] developed a novel model for the intrusion detection system. A new learning algorithm was proposed in [23] for adaptive intrusion detection using naïve Bayesian and boosting classifiers. Additionally, they conducted an experiment using the KDD Cup 1999 dataset. The experiment proved that the proposed algorithm offered higher detection rates with a remarkable reduction in the number of false positives for multiple types of network intrusion. A GA combined with the KNN for feature weighting and selection was proposed in [24]. The proposed model was applied on the KDD Cup 1999 dataset for identifying DDoS/DoS attacks. The result showed that the accuracy for unknown attacks was found to be 78%, whereas the accuracy for known attacks was calculated to be 97.24%. Based on the Pittsburgh, iterative rule learning (IRL), and Michigan approaches, the authors in [25] proposed three different types of genetic fuzzy systems for intrusion detection. A novel feature representation approach was proposed in [26]. This approach is called the cluster center and nearest-neighbor approach (CANN), in which the distance between data and its nearest neighbor and data sample and its cluster center were measured and summed. The authors conducted the experiments using the KDD Cup 1999 dataset, showing that the CANN classifier performed similarly or slightly better than SVM and k-NN.

Two dimensionality reduction techniques, namely, PCA and fuzzy PCZ, were used and compared in [27], where the authors classified the test samples of connections into attack or normal category by applying KNN algorithm. In addition, they conducted experiments using KDD Cup 1999 dataset. The results showed that fuzzy PCA performed better than the PCA in detecting the DoS and U2R attacks. In [28], the authors proposed a deep learning approach using recurrent neural networks (RNN-IDS). The experimental results demonstrated that the RNN-IDS was ideal for modeling a classification model with relatively high accuracy, and its performance was also superior compared to conventional machine learning classification techniques in multiclass and binary classification. The authors in [29] built an anomaly detection system using backpropagation algorithm optimized by Conjugate Gradient (CG) algorithm. Then, they analyzed the use of CG optimization (Polak-Ribiere, Fletcher Reeves, Powell Beale). Based on their experiment results, the average accuracy was 93.2% for two classes “intrusion” and “normal”. Applications of LSTM to RNN for modeling the IDS modeling were proposed in [30]. The ideology of the experiment was dependent on the hyperparameter values, the rate of learning, and changes in the performance; the size of the hidden layer had a significant impact on the performance. According to their experiments, the average rate of detection was computed to be 98.8%. The authors in [31] proposed a learning model, namely, PSO-FLN for fast learning network (FLN), based on particle swarm optimization (PSO). A deep learning model was proposed in [32]. The model is based on the DBN and stacked nonsymmetric deep autoencoder (NDAE). They used KDD Cup 1999 and NSL-KDD datasets to evaluate their model, which accurately detected the Probe attacks and the DoS. Nevertheless, R2L attacks were barely identified, while no detection of the U2R attacks was recorded. The precision value was found to be 99.99%, with an overall accuracy of 97.85%.

III. Background and Formalization

The most critical issue in machine learning is developing a generalized training model that will perform accurately on training data and at least will provide almost the same results on unseen data.

There are many algorithms used whose primary goal is to decrease classification error on unseen data at the cost of increased training error. In other words, we may say that reducing the model’s generalization error without any effect on training error is known as regularization.

Many techniques have been used to improve the generalization performance of the learning model [33]. Some add constraints to the machine learning model, such as putting constraint on model parameters values, and others add further statistical terms in the objective function that are known as a soft constraint on model parameters [34].

Developing a more effective regularization algorithm is a crucial task in the field of machine learning; hence, it is the main focus of research in this field. In a statistical model of learning algorithms, such constraints and penalties are used to encode prior knowledge. On occasion, these penalties and constraints are designed to promote generalization by expressing generic preferences for a simple classification model. However, it is necessary to incorporate such penalties and constraints to make an undetermined problem determined.

As explained above, there are multiple strategies to incorporate regularization in machine learning algorithms [35], [36], [37]. Among these methods, L1- and L2-norm are the most common regularization methods. L2 regularization is also referred to as *Tikhonov regularization*, and, in statistics, as ridge regression. It is combined with the cost function as a complexity term.

L2 regularization is the squared Euclidean of all feature weights of the hidden layer, and, in the case of multiple hidden layers, it is the sum of all such squared norms including the output layer of the neural network [38].

Another regularization parameter,λ, is multiplied with regularization in order to put a penalty on and control the strength of the magnitude of weights. Due to this regularization, the model results in much smaller weights for each layer. Similarly, L1 regularization produces many zeros in the weight matrix and makes it sparse, hence, controlling the complexity of the model. Both L1 and L2 regularizations have a well-defined probabilistic interpretation which is similar to adding a Gaussian prior over the distribution of weight matrix W in case of L2 and Laplacian in case of L1 [39]. However, several tried and tested regularization methods exist for both neural networks and other machine learning algorithms (Random forests, SVM, etc.) [5], [37], [40],[41]. For succinctness, we will focus purely on methods used on ANNs. For simplicity, we can split the methods into categories, with one being sparsity-based regularization and the other not.

The sparsity-based methods to be considered are L1 and L2 norms.

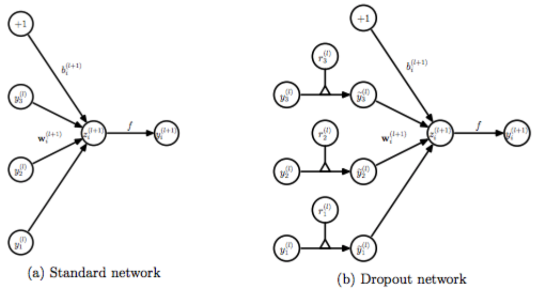
Both methods take a sum over the absolute value and square, respectively.

There is a great amount of previous work comparing L1 and L2 along with other regularization methods in a variety of problem domains [42].

On the other hand, we can also apply methods such as early stopping. This would reduce the number of parameters the network learns; thus, this is considered a form of regularization. The goal of early stopping along with other forms of regularization is to reduce generalization error or increase generalization accuracy while allowing training error to increase.

The most seminal regularization method and one of the more significant breakthroughs in machine learning is dropout [43].

Dropout is an intuitively brilliant discovery that *drops out* or deactivates and removes a portion of neurons randomly according to an arbitrary value. Pushing a neural network to acquire more stable and strong characteristics together with various random subsets of other neurons. Depending on the problem's context, it can be used in combination with sparsity regularizers to good effect (see Fig. 1).



(a) Standard network

(b) Dropout network

Fig. 1. Comparison between a standard neural network and a network implementing dropout

IV. Anomaly Detection and New Regularization Technique

A. Anomaly Detection

Anomaly detection can be framed in many ways. Outlier detection, for instance, can often fall under this umbrella. Here, let us define an anomaly as something that significantly differs from the rest of the data or otherwise grossly misfits the distribution of data. We trained and tested in a supervised context (classification) using a feedforward network to test differences across other regularization techniques and our new regularization in our problem domain.

B. DATA

Due to the nature of the problem, the following 3 datasets were chosen to carry out analysis based on the proposed regularization, L1 and L2 norm regularizations. The competition task was to build a network intrusion detector to analyze the performance of the proposed regularizer with existing regularizers and a predictive model capable of distinguishing between normal connections and attack connections.

1) NSL-KDD Dataset.
2) UNSW-NB15 Dataset
3) CIDDS-001 Dataset

1) NSL-KDD dataset: NSL-KDD dataset is a replacement of KDD-CUP dataset and it solves some problems in the KDD CUP 1999 dataset. In NSL-KDD dataset there are 4 attack categories that represent anomalous data and 1 normal category which shows that the corresponding instances are normal. The dataset is quite imbalance and due to this nature, training a classifier is a challenging task. Various types of attacks categories are shown in Table I.

TABLE I. Attack classes based on different attack types

Attack Class	Training Set	Testing Set
DOS	back, land, neptune, pod, smurf, teardrop	back, land, neptune, pod, smurf, teardrop, mailbomb, processtable, udpstorm, apache2, worm
Probe	ipsweep, nmap, portsweep, satan	ipsweep, nmap, portsweep, satan, mscan, saint
U2R	buffer-overflow, loadmodule, perl, rootkit	buffer-overflow, loadmodule, perl, rootkit, sqlattack, xterm, ps
R2L	fpt-write, guess-passwd, imap, multihop, phf, spy, warezclient, warezmaster	fpt-write, guess-passwd, imap, multihop, phf, spy, warezmaster, xlock, xsnoop, snmpguess, snmpgetattack, httptunnel, sendmail, named

2) UNSW-NB15 dataset: UNSW-NB15 dataset is created by IXIA PerfectStorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) and its purpose is generating a hybrid of real modern normal activities and synthetic contemporary attack behaviors. It contains nine different types of attacks Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms. The whole dataset contains 2,540,044 records and it is available to download as one file or split into several different CSV files. There is also one list of event files which contains information about the number of events categorized by attack category and attack subcategory for all 2.5M records. From that dataset, the training and the test dataset are produced, wherein the training dataset contains175,341 records and 82,332 records in the test dataset [44].

3) CIDDS-001 Dataset: Another dataset we used for our experiment is the CIDDS-001 dataset [45]. This is a labeled flow-based dataset used for intrusion detection system. The following attributes from the dataset are used for training the model: *Src IP, Src Port, Dest IP, Dest Port, Proto, Duration, Bytes, Packets, Flags*.

There are two types of server through which this data is collected (o*pen stack and external server)*. Data from both servers contain the aforementioned attributes, the only difference is in the attack categories.

Data from o*pen stack server* contains the following three categories:

n*ormal, victim and attacker*. While data from the *external server* contains the following 5 categories: n*ormal, victim, attacker, unknown, suspicious*.

C. New Regularization Technique

In the machine learning field, the commonly applied regularization techniques are L1-norm and L2-norm. During optimization, these regularizers consider the complexity of weights to induce the networks towards a more general mapping. L1-norm imposes the sum of the absolute values as a penalty, while L2-norm imposes the sum of the squared values as a penalty. The purpose of this article is to introduce a new regularization that employs the standard deviation of the weight matrix and then multiplies it by λ to make the regularization term. Consequently, the regularizer computes the weights standard deviation of the weights to the loss function.

After studying the L1 and L2 regularizers, we found one

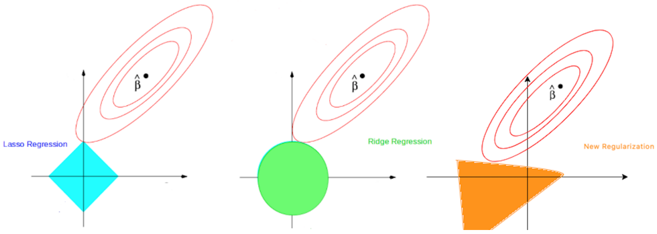


Fig. 2. Contours of L1, L2, and new regularizers

significant drawback, that is regulating the weights' individual values without taking into consideration the relationship between weight matrix entries. To resolve this downside, the new regularization technique utilizes the standard deviation to get the regularization term. This is to construct an adaptive form of weight decay. Thus, the regularizer does not allow the learning model to adapt widespread values from weight space.

The contour of the new regularizer was displayed highlighting the efficacy and potency of the new regularizer. 2 represents the feasible region of L1, L2, and new regularization techniques. The contours of each regularizer represent different loss values. The behavior of the L2-norm is circular and incorporates L1, while the new regularization acts like a parabola and takes values beyond the L2-norm limit. This helps in a sense, it increases the limit of values (space) to be adopted, and based on the penalty term λ this space can be expanded. The formalization as follows (See equations 1), with ω denoting the standard deviation of weight matrix w_i.

$$\lambda \sum_{i=1}^{k} \omega \tag{1}$$

During the training process, λ denotes the regularization parameter that sets a penalty to restrict weights from selecting high values. In other words, the loss function in our case will become (see equations 2):

$$min_w \{f(X, y : w) + \lambda\sigma(w)\} \tag{2}$$

Therefore, if the weight values of all layers are large, the weight values of the selected λ will be large. Thus, the weight values cannot be equal, as they will have more freedom to search in a large space. Consequently, our regularization technique is more effective compared to the L1 and L2 regularization techniques.

The model was trained using the ***Nesterov ADAM*** optimizer, with $tanh$ activation functions. The model was trained over 100 epochs with a batch size of 32. The labeled data were classified with a feedforward network.

D. Artificial Neural Network (ANN) Based IDS with New Regularization Method

In this section, we present the diagrammatic representation of data preprocessing and training ANN model which employed our new regularizer as shown in Fig. 3. There are generally four steps involved in this process (Fig. 2) as follows: There are generally four steps involved in this process (Fig. 3) as explained below.

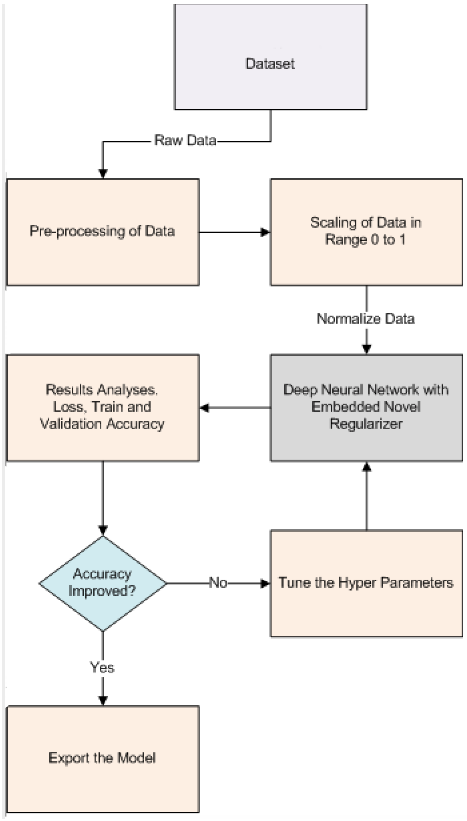

Dataset
Raw Data
Pre-processing of Data
Scaling of Data in Range 0 to 1
Normalize Data
Deep Neural Network with Embedded Novel Regularizer
Results Analyses, Loss, Train and Validation Accuracy
Accuracy Improved?
No
Tune the Hyper Parameters
Yes
Export the Model

Fig. 3. New regularizer based IDS workflow

1) Data Preprocessing: Artificial Neural Network uses only numerical data for training and testing. So, the initial step is to transform nominal and textual data into numerical data. To do this, the following steps were performed:

- All the nominal and textual attributes were converted by using one-hot encoding (nominal to binary conversion in Weka). Conversion of attributes to one-hot encoding leads to increasing of attributes in attributes. Therefore, the number of units in ANN is adjusted according to attributes.
- Each category of attack types was converted by one-hot encoding.

2) Data Scaling: After data preprocessing, each dataset contains attributes of numerical values and one-hot encoded values. The numerical values were normalized according to the formulation given in equation 3.

$$\bar{X_i} = \frac{X_i - min(X_i)}{max(X_i) - min(X_i)} \tag{3}$$

For $i = 1, ..., n$ where n represents the number of records, and x represents a specific column in the dataset. Next, duplicate records were removed from the dataset to restrict classifiers from giving biased results.

3) Training the ANN model: After data preprocessing and data scaling phases, our next task is to implement ANN model. Python was picked to be the implementation language and the Keras framework was employed for ANN. The ANN model

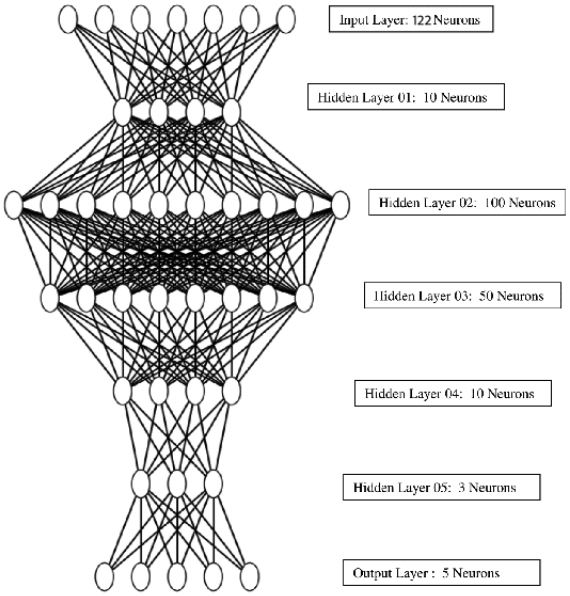

Input Layer: 122 Neurons
Hidden Layer 01: 10 Neurons
Hidden Layer 02: 100 Neurons
Hidden Layer 03: 50 Neurons
Hidden Layer 04: 10 Neurons
Hidden Layer 05: 3 Neurons
Output Layer : 5 Neurons

Fig. 4. ANN architecture with embedded new regularizer

is incorporated with a new regularizer to test our method. Employing the mathematical description in equation 1, the proposed regularizer is applied as a function. In fact, the kernel_regularizer was assigned with this new function rather than built-in regularizers in the ANN model. The ANN predictive model includes two hidden five- and three-unit layers, respectively. The last layer is composed of two units according to class values. $tanh$ is the activation function employed in each layer, except for the last layer, where the $softmax$ activation function is utilized. In the first two layers, the weight matrix was initialized with Gaussian random distribution. Due to the large number of neurons in each layer, we show only the reduced version of the model as depicted in Fig. 4. The first layer consists of 122 neurons, the first hidden layer 10 neurons, and the second layer 100 neurons.

Likewise, the third hidden layer has 50 neurons. In the fourth hidden layer, the input size is reduced to 10. Hence, we added 10 neurons to it. Further, these neurons are connected to 3 neurons in the fifth hidden layer which is further connected to 5 output neurons each for one of the five specific categories. In each layer, a Kernel matrix is initialized with uniform distribution and $tanh$ activation function except the last layer which has $softmax$ activation function. Further, for binary classification, the last layer has 2 output neurons. Finally, the model is compiled with an $adam$ optimizer with a default learning rate and other parameters.

4) HyperParameters Adjustment: After each 100-epoch run of the ANN model, precision and loss values were evaluated and the hyperparameters were adjusted accordingly. Activation functions and kernel initializer distributions were determined after several iterations and examining the depth of the ANN model. According to our optimal desired outputs, regularization parameter λ was also adjusted. The number of layers and hyperparameters remained the same for each regularizer. λ parameter was constantly updated and fixed to the value resulting in the highest and best accuracy for the corresponding regularizer.

V. Results and Discussion

To produce results based on the proposed method, we implemented our model for multiclass classification *(normal and four different attack categories (5-class))*. In addition, we applied the 10-fold cross-validation on each dataset. All simulations were carried out on a server having 32 GB RAM, GeForce GTX 1080 GPU of 8 GB GDDR5X memory, and 2560 NVIDIA CUDA cores. We compared results for multiclass problems in each case and demonstrated our results. For each dataset, the corresponding attack categories were considered as classes and the ANN with new, L1, and L2 regularizations is trained by using 10-fold cross-validation. For every attack class in each dataset, the performance measures described in equations 5–9 were computed and presented. In the following sections, results for each dataset based on our new regularization are compared with other regularization algorithms. In each type of classification, the proposed regularization demonstrates a good performance and is superior to L1 and L2 regularizations. Furthermore, other hyperparameters and results on each classification category are discussed in detail.

A. Evaluation Protocols

For multiclass classification, the loss function used is categorical cross entropy as given in equation 4.

$$CrossEntropy = -\sum_i^C t_i \log f(s_i) \tag{4}$$

where C is the number of classes, t_i is the ith class and $f(s_i)$is the ith output after activation function f.
To evaluate our model, training and validation accuracy are reported for the data partitions as explained in Results and Discussion. Accuracy is calculated based on the following mathematical representation. Apart from accuracy, other performance measures, that is, TPR also known as Recall, False Positive Rate (FPR), Precision (Pre), and F1 measures, are calculated based on equations 5, 6, 7, 8, and 9, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

$$Pre = \frac{TP}{TP + FP} \tag{8}$$

$$F1 = 2 * \frac{TPR \times Pre}{TPR + Pre} \tag{9}$$

where TP, TN, FP and FN denote true positives, true negatives, false positives, and false negatives, respectively.

B. Performance Measurement of Different Regularization Techniques

1) NSL-KDD dataset: ANN model with embedded new regularization is trained on 25192 samples (**The 20% of NSL-KDD training set (KDDTrain+)** by using 10-fold cross-validation. The trained model is then tested on a separate test dataset containing 22544 samples (**KDDTest+**). After that the results were recorded in Tables II, III, and IV. NSL-KDD dataset is an imbalanced dataset; therefore, the individual performance measures for each class are significantly affected. For example, R2L attack type has a total of 224 samples and the performance is lower for this category type. In such a situation, the classifier is biased towards more frequent samples, for example, a normal category having 9711 samples.

New regularization: NSL-KDD dataset has four different attack categories. For each attack category, different performance measures were computed (see Table II). Experimental results for TPR are also demonstrated in Fig. 5.

TABLE II. PERFORMANCE MEASURES FOR NSL-KDD DATASET BY USING NEW REGULARIZATION

Labels	TPR	FPR	Pr.	F1	Acc
Normal	0.980	0.0073	0.990	0.985	98.5%
DoS	0.978	0.0054	0.989	0.983	
R2L	0.924	0.0087	0.520	0.665	
U2R	0.969	0.0033	0.977	0.973	
Probe	0.956	0.0067	0.941	0.948	

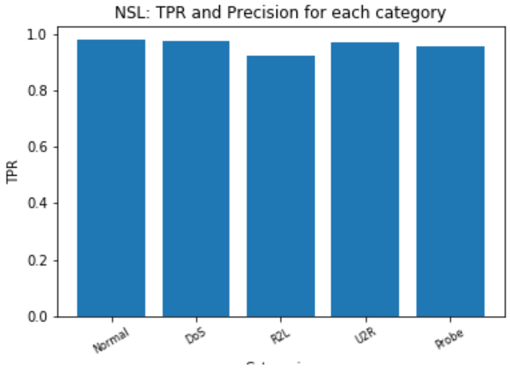


Fig. 5. TPR observerd using the new regularizer for NSL-KDD dataset

L1-Norm regularization: For the sake of comparison, we also used L1-norm regularization for the 5 classes of NSL-KDD dataset. Finally, we computed the performance measures results (see Table III).

TABLE III. PERFORMANCE MEASURES FOR NSL-KDD DATSET USING L1-NORM REGULARIZATION

Labels	TPR	FPR	Pr.	F1	Acc
Normal	0.975	0.011	0.986	0.981	95.4%
DoS	0.96	0.013	0.974	0.967	
R2L	0.938	0.01	0.502	0.654	
U2R	0.948	0.01	0.937	0.942	
Probe	0.943	0.006	0.95	0.947	

L2-norm regularization: Further, we trained the classifier by using L2-norm regularization. The result of NSL-KDD datasets is shown in Table IV.

TABLE IV. PERFORMANCE MEASURES FOR NSL-KDD DATASET USING L2-NORM REGULARIZATION

Labels	TPR	FPR	Pr.	F1	Acc
Normal	0.978	0.011	0.986	0.982	97.2%
DoS	0.967	0.007	0.985	0.976	
R2L	0.929	0.011	0.472	0.626	
U2R	0.956	0.004	0.975	0.966	
Probe	0.948	0.009	0.927	0.938	

Based on our analysis of the above results, we observed that, in terms of average TPR and FPR, our proposed technique outperformed the L1 and L2 regularizations. The average TPR for the proposed regularizer is 96.2%, while the L1 and L2 regularizations' average TPR was 95.27% and 95.56%, respectively. Similarly, the average FPR is lower using the proposed regularizer, being 0.63%. For L1 and L2 regularizations, the average FPR is 0.97% and 0.8%, respectively. Regarding the training time of each regularizer, our proposed regularization took 177.3 seconds, whereas L1 and L2 regularizers took almost 176.5 seconds. Obviously, the training time for all models is almost the same. While in testing, the parameters are kept static (as we do not change them during testing); thus, the role of tuning the regularization has vanished during testing. Consequently, the testing time was less than training. For the NSL-KDD dataset, the testing time for all models was 46.02 seconds. Hence, we undoubtedly can state that our proposed regularizer performed better than other regularizers on the NSL-KDD dataset.

2) UNSW-NB 15 dataset: In addition, we provided the comparison of different performance measures for the UNSW-NB15 dataset using new, L1-norm, and L2-norm regularizers. We tested these different regularizations using 10 classes of the UNSW- NB15 dataset on 175,341 samples given in an explicit training set (NSW_NB15_Train). Similarly, the models are tested on 82,332 samples (UNSW_NB15_Test), and then we computed the results (see Tables V, VI, and VII).

New regularizaton : We embedded the proposed regularization with ANN model and then tested it using 10 different categories of the UNSW-NB15 dataset as shown in Table V. TPR results can also be viewed from Fig. 6.

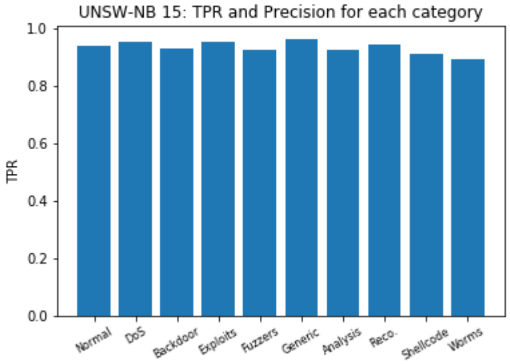


Fig. 6. TPR observed using proposed regularizer on UNSW-NB 15 dataset

L1-norm regularization: Table VI represents our simulation results using L1-norm regularization *(using the same ANN model having an equal number of layers and units).*

TABLE V. NEW REGULARIZATION RESULTS ON UNSW-NB 15 DATASET

Labels	TPR	FPR	Pr.	F1	Acc
Normal	0.941	0.009	0.98	0.96	94.58%
DoS	0.952	0.01	0.878	0.914	
Backdoor	0.933	0.006	0.641	0.76	
Exploits	0.956	0.007	0.972	0.964	
Fuzzers	0.924	0.006	0.95	0.937	
Generic	0.964	0.034	0.966	0.965	
Analysis	0.927	0.006	0.635	0.753	
Reconnaissance	0.944	0.005	0.92	0.932	
Shellcode	0.911	0.002	0.72	0.804	
Worms	0.892	0.001	0.504	0.644	

TABLE VI. L1-NORM RESULTS ON UNSW-NB 15 DATASET

Labels	TPR	FPR	Pr.	F1	Acc
Normal	0.937	0.015	0.969	0.953	92.4%
DoS	0.942	0.013	0.854	0.896	
Backdoor	0.919	0.007	0.586	0.716	
Exploits	0.932	0.009	0.964	0.948	
Fuzzers	0.908	0.009	0.928	0.918	
Generic	0.949	0.039	0.961	0.955	
Analysis	0.925	0.008	0.584	0.716	
Reconnaissance	0.925	0.005	0.926	0.925	
Shellcode	0.902	0.003	0.7	0.789	
Worms	0.862	0.001	0.407	0.553	

L2-norm regularization: We embedded L2-norm regularization with ANN model having an equal number of layers and units *(as that used for the proposed regularization)*. The results are shown in Table VII. Based on the analysis of our results, the average TPR computed for this dataset using the proposed regularization is 93.43%. However, for L1- and L2-norm regularizations, the average TPR is 92% and 92.7%. Here, again our model outperformed the existing regularizations in terms of TPR. Similarly, our proposed regularization surpassed the existing regularizations in terms of FPR. The average FPR achieved using the proposed regularizer is 0.86%, whereas the average TPR for L1 is 1.06% and for L2 is 0.96%. Regarding the training time, the proposed regularization took 425.9 seconds, while L1 and L2 took 424.7 seconds *(which is an acceptable difference)*. Noteworthy, the testing time was much less than training *(the testing time was 126.2 seconds)*.

TABLE VII. L2-NORM REGULARIZATION RESULTS ON UNSW-NB 15 DATASET

Labels	TPR	FPR	Pr.	F1	Acc
Normal	0.94	0.011	0.977	0.958	94.3%
DoS	0.948	0.011	0.873	0.909	
Backdoor	0.925	0.006	0.62	0.743	
Exploits	0.943	0.008	0.969	0.956	
Fuzzers	0.918	0.007	0.938	0.928	
Generic	0.957	0.036	0.964	0.961	
Analysis	0.923	0.008	0.599	0.726	
Reconnaissance	0.93	0.006	0.915	0.922	
Shellcode	0.91	0.003	0.689	0.784	
Worms	0.877	0.001	0.427	0.574	

3) CIDDS-001 dataset: Here, we carried out several experiments on the CIDDS-001 dataset using different regularizations. CIDDS-001 dataset has two parts:

1) External Server Dataset
The model is trained over all the dataset provided using 10-fold cross-validation, except for a set of 339030 samples which were kept separate for testing purposes. Apart from normal samples, there are 4 different attack categories in this dataset which are $victim$, $attacker$, $unknown$, and $suspicious$.

2) Open Stack dataset
The same approach is applied to this dataset. The ANN model is trained over all dataset using the 10-fold cross-validation, except for a set of 2789002 samples. There are only 2 attack categories which are $victim$ and $attacker$.

We carried out our experiments on both datasets and the following performance measures were observed for each regularization.

New regularization: The experimental results on the two types of dataset using the new regularization are given in Tables VIII and IX. TPR results for each category are also shown in Fig. 7 and 8, respectively.

TABLE VIII. SIMULATION RESULTS ON CIDDS-001 OPENSTACK DATASET USING THE PROPOSED REGULARIZATION

Labels	TPR	FPR	Pr.	F1	Acc
Normal	0.988	0.019	0.996	0.992	97.87%
Victim	0.979	0.007	0.928	0.953	
attacker	0.969	0.005	0.945	0.957	

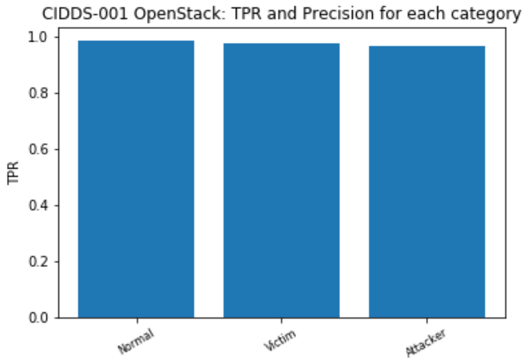


Fig. 7. TPR observed for OpenStack server dataset using the new regularization

L1-norm regularization: Results obtained using L1-norm regularization for each of the two types of dataset are shown in Tables X and XI.

L2-norm regularization: Tables XII and XIII demonstrated the results using L2-norm regularization for each of the two types of the dataset.

TABLE IX. SIMULATION RESULTS ON CIDDS-001 EXTERNAL SERVER DATASET USING THE PROPOSED REGULARIZATION

Labels	TPR	FPR	Pr.	F1	Acc
Normal	0.979	0.006	0.969	0.974	96.8%
Victim	0.94	0.002	0.901	0.92	
attacker	0.957	0.002	0.941	0.949	
unknown	0.978	0.006	0.968	0.973	
suspicious	0.986	0.011	0.993	0.99	

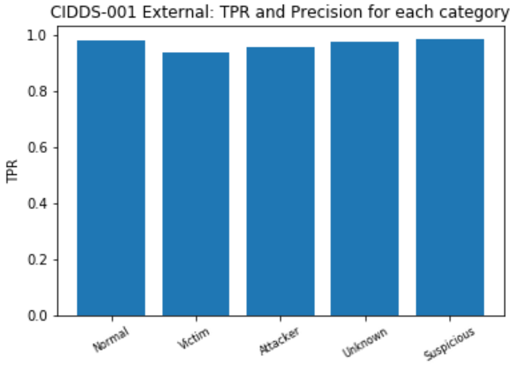


Fig. 8. TPR observed for External server dataset using the new regularization

TABLE X. SIMULATION RESULTS FOR CIDDS-001 OPEN STACK DATASET USING L1-NORM REGULARIZATION

Labels	TPR	FPR	Pr.	F1	Acc
Normal	0.969	0.025	0.995	0.982	96.6%
Victim	0.965	0.016	0.848	0.903	
attacker	0.95	0.016	0.852	0.898	

TABLE XI. SIMULATION RESULTS ON CIDDS-001 EXTERNAL SERVER DATASET USING L1-NORM REGULARIZATION

Labels	TPR	FPR	Pr.	F1	Acc
Normal	0.965	0.008	0.962	0.964	95.3%
Victim	0.928	0.009	0.674	0.781	
attacker	0.937	0.007	0.8	0.863	
unknown	0.964	0.01	0.949	0.956	
suspicious	0.968	0.014	0.992	0.98	

TABLE XII. SIMULATION RESULTS FOR CIDDS-001 OPEN STACK DATASET USING L2-NORM REGULARIZATION

Labels	TPR	FPR	Pr.	F1	Acc
Normal	0.981	0.019	0.996	0.988	96.0%
Victim	0.979	0.009	0.91	0.943	
attacker	0.968	0.01	0.903	0.935	

TABLE XIII. SIMULATION RESULTS FOR CIDDS-001 EXTERNAL SERVER DATASET USING L2-NORM REGULARIZATION

Labels	TPR	FPR	Pr.	F1	Acc
Normal	0.972	0.011	0.947	0.959	96.4%
Victim	0.935	0.003	0.868	0.9	
attacker	0.947	0.005	0.856	0.899	
unknown	0.969	0.009	0.953	0.961	
suspicious	0.974	0.013	0.992	0.983	

- External Server Dataset
  Average TPR of 96.8%, 95.24%, and 95.93% was observed for the external server dataset. Similarly, the average FPR computed is 0.55%, 0.95%, and 0.82% for the proposed, L1, and L2 regularizations, respectively. The training time was 1028, 1019, and 1022 seconds for the proposed, L1, and L2 regularizations. In this case, the testing time was the same for all regularizations which was 76.6 seconds. Based on the analysis of the above results, it can be concluded that the proposed regularization outperformed the other regularizations.

- OpenStack Server Dataset
  For this dataset, the average TPR computed was 97.86%, while the FPR was 1.03%. As for the L1 and L2 regularizers, the average TPR is 96.13% and 97.6%. Similarly, the average FPR for L1 and L2 is 1.9% and 1.26%, respectively. As far as the training time is concerned, the training time took 1728.4, 1720.0, and 1723.2 seconds for the proposed, L1, and L2 regularizations, respectively. In terms of testing, the time taken was 97.7, 100.2, and 92.8 seconds, respectively. Hence, from the results above, we can conclude that the performance of the proposed regularization is slightly higher than other regularizers.

VI. LIMITATION OF NEW REGULARIZATION

Several researchers employed multiple regularization algorithms, the most common ones being lasso regularizations and ridge regression. However, some disadvantages are inherent in the regularization framework. For example, In a challenging setting, where the number of instances is very low and the dimensionality is very high, it is impractical to utilize these regularizations. Likewise, our regularization algorithm had multiple limitations, as follows:

- It cannot be employed for selecting or reducing features.

- It is challenging to choose a suitable value of λ, due to the fact that it is a continuous value. In addition, the process of picking a suitable value from multiple attempts will be computationally costly and time consuming.

VII. CONCLUSION

The field of ANN regularizers is one that is still ripe for new research and innovation. From attempts in adaptive weight decay to new techniques altogether, many innovations in improving generalization through reducing model complexity are possible. In this paper, we proposed a new regularization technique for anomaly detection based on the standard deviation of the weight matrix. Based on the analysis of our experimental results, it is evident that our proposed regularization algorithm makes the ANN capable of identifying good patterns in data and classifying them efficiently. Moreover, the proposed regularizer has outperformed the existing regularization algorithms when incorporated with ANN. As a result, the overall average accuracy achieved on NSL-KDD, UNSW-NB15, and CIDDS-001 datasets using 10-folds cross-validation is 98.53%, 94.58%, and 97.87%, respectively.

FUNDING

This work was supported by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah . The authors, therefore, gratefully acknowledge DSR technical and financial support.

Conflict of interest

The authors declare that they have no conflict of interest.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Data Availability

Datasets used to support the findings of this study are included within the article.

References

[1] M. Markos and S. Singh, "Novelty Detection: A Review-Part 1: Statistical Approaches," *J. Signal Processing*, vol. 83, 2003. 2481 2497,2003.

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. (CSUR), pp. 41(3)–15, 2009.

[3] M. Markou and S. Singh, "Novelty detection: a review – part 2: neural network based approaches, Signal Process," *Signal Process*, vol. 12, pp. 2499–2521, 2003.

[4] P. Murugan and S. Durairaj, "Regularization and optimization strategies in deep convolutional neural network," 2017.

[5] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," vol. 609, pp. 223–239, 2010.

[6] K. Das, "Detecting Patterns of Anomalies," *Carnegie Mellon University*, 2009.

[7] H. T. M, "The Science of Anomaly Detection, Numenta," 2015.

[8] C. Xiang, P. C. Yong, and L. S. Meng, "Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees," pp. 918–924, 2008.

[9] G. Giacinto, R. Perdisci, M. D. Rio, and F. Roli, "Intrusion detection in computer networks by a modular ensemble of oneclass classifiers," 2008.

[10] W. Hu, W. Hu, and S. Maybank, "AdaBoost-based algorithm for network intrusion detection," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, no. 2, pp. 577–583, 2008. Cited By :137.

[11] A. Tajbakhsh, M. Rahmati, and A. Mirzaei, "Intrusion detection using fuzzy association rules," 2009.

[12] K. Shafi and H. A. Abbass, "An adaptive genetic-based signature learning system for intrusion detection," pp. 12036–12043, 2009.

[13] S.-Y. Wu and E. Yen, "Data mining-based intrusion detectors," pp. 5605–5612, 2009. doi.org/10.1016/j.eswa.2008.06.138 ID: 271506.

[14] T. P. Tran, L. Cao, D. Tran, and C. D. Nguyen, "Novel intrusion detection using probabilistic neural network and adaptive boosting," *arXiv preprint*, vol. 911, no. 0485 6, pp. 83–91, 2009.

[15] G. Wang, J. Hao, J. Ma, and L. Huang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering," pp. 6225–6232, 2010.

[16] M. S. Mok, S. Y. Sohn, and Y. H. Ju, "Random effects logistic regression model for anomaly detection," pp. 7162–7166, 2010.

[17] R. Abdulhammed, M. Faezipour, A. Abuzneid, and A. AbuMallouh, "Deep and machine learning approaches for anomaly-based intrusion detection of imbalanced network traffic," *IEEE Sensors Letters*, vol. 3, p. 1–4, Jan 2019.

[18] A. Verma and V. Ranga, "On evaluation of network intrusion detection systems: Statistical analysis of cidds-001 dataset using machine learning techniques," *Pertanika Journal of Science & Technology*, vol. 26, pp. 1307–1332, march 2018.

[19] A. Verma and V. Ranga, "Statistical analysis of cidds-001 dataset for network intrusion detection systems using distance-based machine learning," *Procedia Computer Science*, vol. 125, p. 709–716, 2018.

[20] B. A. Tama and K.-H. Rhee, "An in-depth experimental study of anomaly detection using gradient boosted machine," *Neural Computing and Applications*, vol. 31, pp. 955–965, Apr 2019.

[21] B. A. Tama, M. Comuzzi, and K.-H. Rhee, "Tse-ids: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system," *IEEE Access*, vol. 7, p. 94497–94507, 2019.

[22] M. M. T.Jawhar and M. Mehrotra, "Design network intrusion detection system using hybrid fuzzy-neural network," *International Journal of Computer Science and Security*, vol. 4, no. 3, pp. 285–294, 2010.

[23] C. M. Rahman, D. M. Farid, and M. Z. Rahman, "Adaptive intrusion detection based on boosting and naive bayesian classifier," *International Journal of Computer Applications*, vol. 24, no. 3, pp. 11–19, 2011.

[24] M.-Y. Su, "Real-time anomaly detection systems for Denial-of-Service attacks by weighted k-nearest neighbor classifiers," pp. 3492–3498, 2011.

[25] M. S. Abadeh, H. Mohamadi, and J. Habibi, "Design and analysis of genetic fuzzy systems for intrusion detection in computer networks," pp. 7067–7075, 2011.

[26] W.-C. Lin, S.-W. Ke, and C.-F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," 2015.

[27] A. Hadri, K. Chougdali, and R. Touahni, "Intrusion detection system using PCA and Fuzzy PCA techniques," in *Advanced Communication Systems and Information Security (ACOSIS), International Conference on. IEEE*, pp. 1–7, 2016.

[28] C. Yin, Y. Zhu, J. Fei, and X. He, "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks," *IEEE Access*, vol. 5, no. 2017, pp. 21954–21961, 2017.

[29] U. N. Wisesty and Adiwijaya, "Comparative study of conjugate gradient to optimize learning process of neural network for Intrusion Detection System (IDS).," in *In Science in Information Technology (ICSITech), 2017 3rd International Conference on. IEEE*, pp. 459–464, 2017.

[30] J. Kim, J. Kim, H. L. T. Thu, and H. Kim, "Long short term memory recurrent neural network classifier for intrusion detection,," in *2016 International Conference on Platform Technology and Service (PlatCon) IEEE*, pp. 1–5, 2 2016.

[31] M. H. Ali, B. A. D. A. Mohammed, M. A. B. Ismail, and M. F. Zolkipli, "A new intrusion detection system based on Fast Learning Network and Particle swarm optimization," *IEEE Access*, vol. 6, no. 2018, pp. 20255–20261, 2018.

[32] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A deep learning approach to network intrusion detection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 1, pp. 41–50, 2018.

[33] M. Tanaka, V. Sladek, and J. Sladek, "Regularization techniques applied to boundary element methods," *Applied Mechanics Reviews*, vol. 47, pp. 457–499, 1994.

[34] B. S. Kautz and Y. Jiang, "A general stochastic approach to solving problems with hard and soft constraints," *The Satisfiability Problem: Theory and Applications*, pp. 573–586, 1997.

[35] P. Simard, B. Victorri, Y. LeCun, and J. Denker, "Tangent prop-a formalism for specifying selected invariances in an adaptive network,," *Advances in neural information processing systems*, pp. 895–903, 1992.

[36] S. J. Hanson and L. Y. Pratt, "Comparing biases for minimal network construction withback-propagation," *Advances in neural information processing systems*, pp. 177–185, 1989.

[37] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout:a simple way to prevent neural networks from overfitting," *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[38] R. Moore and J. DeNero, "L1 and L2 Regularization for Multiclass Hinge Loss Models," in *Symposium on Machine Learning in Speech and Natural Language Processing (MLSLP), WA*, pp. 1–5, 2011.

[39] A. Y. Ng, "Feature selection, L1 vs. L2 regularization, and rotational invariance," in *ICML*, 2004.

[40] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[41] H. Deng and G. Runger, "Feature selection via regularized trees," in *Proc. 12th IEEE International Joint Conference on Neural*, pp. 1–8, 2012.

[42] H. Peng, L. Mou, G. Li, Y. Chen, Y. Lu, and Z. Jin, "A Comparative Study on Regularization Strategies for Embedding-based Neural Networks," 2015.

[43] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580," 2012.

[44] N. Moustafa and J. Slay, ""The significant features of the UNSW-NB15 and the KDD99 sets for Network Intrusion Detection Systems", the 4th International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS," 2015.

[45] M. Ring, S. Wunderlich, D. Grüdl, D. Landes, and A. Hotho, "Flow-based benchmark data sets for intrusion detection," in *Proceedings of the 16th European Conference on Cyber Warfare and Security (ECCWS)*, pp. 361–369, ACPI, 2017.

Successive Texture and Shape based Active Contours for Train Bogie Part Segmentation in Rolling Stock Videos

Kaja Krishnamohan[1], Ch.Raghava Prasad[2], P.V.V.Kishore[3]
Research Scholar, Department of ECE, Koneru Lakshmaiah Education Foundation, Guntur (DT), Andhra Pradesh, INDIA.[1]
Department of ECE, Koneru Lakshmaiah Education Foundation, Guntur (DT), Andhra Pradesh, INDIA.[2,3]

***Abstract*—Train Rolling Stock Examination (TRSE) is a procedure for checking damages in the undercarriage of a moving train at 30kmph. The undercarriage of a train is called bogie according to railway manuals. Traditionally, TRSE is performed manually by set of highly skilled personnel of the railway near to the train stations. This paper presents a new method to segment the TRSE bogie parts which can assist trained railway personnel for better performance and consequently reduce train accidents. This work uses visualization techniques as a pair of virtual eyes to help checking of each bogie part remotely using high speed video data. Our previous AC models are being supervised by a weak shape image which has shown to improve segmentation accuracies on a closely packed inhomogeneous train bogie object space. However, the inner texture of the objects in the bogies is found to be necessary for better object segmentation. Here, this paper proposes an algorithm for bogie parts segmentation as successive texture and shape-based AC model (STSAC). In this direction, texture of the bogie part is applied serially before the shape to guide the contour towards the desired object of interest. This contrasts with the previous approaches where texture is applied to extract object shape, loosing texture information completely in the output image. To test the proposed method for their ability in extracting objects from videos captured under ambient conditions, the train rolling stock video database is built with 5 videos. In contrast to previous models the proposed method has produced shape rich texture objects through contour evolution performed sequentially.**

***Keywords*—Automation of Train Rolling Stock Examination; level sets; shape priors; texture priors; export system models**

I. Introduction

Visual automated testing of machines by computer algorithms has been gaining momentum in the past few decades. This increase can be attributed to factors such as high-resolution visual sensors, high speed cameras and more significantly the higher processing power of computers. Progressively, these advancements can be noticed in manufacturing industries, where the assembly lines are monitored visually by high speed cameras to identify defects in products manufacturing processes and packaging. Consequently, the manufacturing industry was revolutionized by visual monitoring technologies thereby improving productivity and quality of production. The long-term dependencies were higher revenues and lowered labour costs. Steadily visual automation has become industry's biggest challenge in promising new solutions to multitude of problems. One such problem that hadn't been explored was Train rolling stock examination.

Train Rolling Stock Examination (TRSE) is a budgeted system on the Indian Railways operational space. The TRSE is currently being executed at every major train station across the Indian subcontinent and the world over to man the safety of passenger trains. Trains on Indian subcontinent carry around 10 million passengers per day. This has been the primary mode of commercially affordable long distance transportation on the planet. Safety of the train during transit is the most significant factor for the rail companies around the world and the foremost job for Indian Railways. Considering the number of train accidents from the past decades, the train transportation has been one of the safest mode of travel and is mostly attributed to rolling stock examination personal.

Train Rolling Stock Examination (TRSE) is a procedure for checking damages in the undercarriage of a moving train at 30kmph. The undercarriage of a train is called bogie according to railway manuals. The bogie consists of dynamic machinery on which the passenger car moves. It is made of wheels, break units, suspension, holding rods, springs, axle box, etc. There are around one hundred components in the bogie that cater for the train movement. The bogie parts have to be constantly monitored during transit as there go through extremities of pressures. The pressure on the bogie parts come due to inter part stroking between them during high speed motion of the train. This causes wear and tare in the bogie parts, which if not checked in time have caused extensive damage to the train causing derailment and human loss. To periodically check the bogie parts during transit, the long lasting and most trusted process is train rolling stock examination.

Traditionally, TRSE is performed manually by set of highly skilled personnel of the railway near to the train stations. Fig. 1 shows a rolling pit with trained railway employees noting the results of their examination (not in frame). These personal are trained for years to use their visual and auditory sensors to identify weaknesses in the bogie parts that can potentially cause an accident. Consequently, the noted risk factors are relayed to the nearest station maintenance crew for necessary repairs. Though the process of TRSE is full proofed, the system in the past hasn't been successful in preventing accidents and loss of life. The fact that the system is heavily dependent on human performance in naturalistic environments, which is dynamic in lighting, temperature, winds and water. Finally, it also dependents on human emotional health at the time of the hour.

The goal of any railway company is to provide a safe transportation system. Despite their committed efforts through

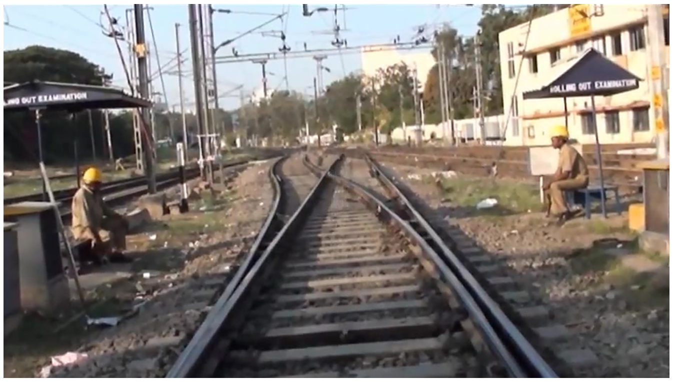


Fig. 1. Manual Train Rolling Stock Examination at an Indian Railway Station.

times there are accidents, and many are during train movement. Manual TRSE needs an extra degree of support to perform without glitches. Technology started providing solutions to this age-old problem only in last two decades. Despite some conformations using sensors, there were no real solutions on the visual frontiers. Train Rolling Stock Examination (TRSE) is a conditional health monitoring system for damage detection in moving passenger bogies used to prevent trains from derailment. Currently, TRSE has been performed manually by experts through audio visual inspection on a moving train at both ends of a train station.

The commercially available sensors and signal processing technology used for TRSE has successfully identified only 20% of the total detect causing derailments. The large set of anomalies found during TRSE are predominantly visual in nature. Thus, a visual automation solution using computer vision algorithms can make the process more robust in preventing train accidents. To this end, the primary task is to extract bogie objects from a video sequence of moving train.

Hence, this chapter proposes a novel segmentation model on the videos of train bogies using a serial texture and shape-imposed level set evolution. The present generation of level set models use texture and shape priors for segmenting objects in an image. However, these models regularize the shape of the contour using the texture in the shape region resulting in a boundary shaped object with no texture. This limits the algorithms capacity to handle segmentation in real time video applications. To preserve texture during shape segmentation, we propose a novel serial texture and shape prior level set evolution model. We first present a theoretical framework for the model with various parameters for regularized contour evolution.

To demonstrate the applicability of the proposed method, experimentation and analysis is performed on benchmark image data and the real-time application, TRSE. The consistency of the algorithm is validated against the state – of – the – art level sets on TRS high speed video datasets. The results show that the proposed method is feasible in practice for segmenting texture preserving shapes in real time videos. The rest of the paper is arranged as follows. The second section gives the background motivation highlighting various gaps in current methods. Methodology and experimentation are provided in sections three and four, respectively. Final section of this paper draws conclusions on the proposed method.

II. Related Background

The promising and motivational research that inspired the formation of this thesis was industrial imaging solutions [1]. Industrial computer vision applications included a variety of image acquisition systems that use the captured videos to detect patterns during the product assembly. The most widely used are CMOS image sensors and hyperspectral sensors [2]. These sensors along with the embedded software has shown to be a valuable asset in bottling and beverage cans quality testing and discarding the faulty bottles on a high-speed assembly line [3].

The automobile industry and its robots use computer vision systems for wheel alignment to mirror inspections [4]. Largely, the operations performed by the software programs are designed to process the image of the object in question to make decisions on its quality and maintenance. The image processing methods used range from simple edge detection to as complex as filtering in frequency domain [5]. Consequently, offline testing of industrial products is on the rise from the last few decades due the availability of commercially viable sensors [6].

Subsequently, vision-based models have shown to provide accuracies on the higher side in most of the real-world industrial applications. This has motivated us to take up the study and investigate the problem to make the manual system of TRSE into an assistant for railway personal. The next session gives an insight into the current operating models of TRSE that are being prototyped and undergoing testing in railways across the world.

The railways around the world and the IR have adopted technologies for locomotive design, coaches, signalling systems, accident prevention with GPS, track maintenance [7], [8], [9], [10], [11]. The quintessential component would be maintenance of train coaches and bogies that are most likely to get damaged during their high-speed movement.

IR uses a highly trained human workforce to do the maintenance checks under the banner called Rolling Stock Examination (RSE) [12], [13]. The checking of the train happens during the train movement at less than 30KMPH near to the railway stations. The check log that is prepared during each RSE involves the visual analysis of the train undercarriage during running is accessible at [14].

The first models that were developed involved sensors along the track that measure parameters such as temperature, pressure, break wear and tear, acceleration etc along with a normal camera module [15]. This prototype is currently being tested in the code name KRATES - Konkan Railway Automated Train Examination System [16].

In this system, the objective of the camera was to have a visual examination at a remote location and no algorithms were proposed to automate the visual information. Apart from this, the camera is an RGB video camera with a frame rate of 30fps, which gives blurry images of bogie for automated processing.

However, the RSE involves checks for hanging parts, lose couplings in bogie parts, break bindings, broken components, hot axle boxes and flat tyres. The following disadvantages were listed for manual TRSE which are limited to, Human factors – Biased Judgment, Heavy Workload – Wrong Judgment, Communication Lags – Between RSE and Maintenance dept., Ambient Nature – Weather Dependent and Commercially Draining.

A complete automated TRSE is quite possible with a large sensor network placed along the tracks. In the current scenario it is still a long-term plan for most of the rail networks due to issues like technology development, deployment and commerce. Despite these issues, this problem is quite challenging and an assistance to manual TRSE is proposed in through this work.

Narayanaswami [17] helps to unfold the connection between automation technologies in transportation to prevent accidents. Inspired by the ideas in [17], this work applies machine vision algorithms for discovering train undercarriage parts from videos of recorded. There are only few works on train safety research with computer vision, targeted at monitoring of rails, ballast and a few on rolling stock.

Sabato et al. [18] proposed methods to inspect train tyres as well as ballast using algorithms developed on 3D digital image correlation (DIC) calculations. It's a dual camera system with a marked displacement installed on a train running at 60Kmph that builds rails as a 3D image. The 3D DIC along with pattern projection models were applied to identify deformation of railway tracks.

Hart, et al. [19] extracted bogie parts for inspection using multispectral imaging camera sensors. The proposed dual camera model recorded multi spectral data as RGB (Red, Green and Blue) along with a thermal sensor to capture a running train which are further treated as panoramic view models.

The computer vision algorithm in [19] has been designed to spot elevated hot bogie parts such as wheel joints, Axel box, brake shoes, air conditioning blowers. Nevertheless, this system has been successful in identifying defective regions, but the motion blur in the video data poses a challenge to distinguish cold parts from hot objects.

Kim et al. [20], has proposed a curve fitting view to the problem of automated train break examination using image processing. The developed techniques use a trench hole establishment under the tracks to capture the break panels of a moving train. The method uses a fitting curve on the recorded images to progressively train the system to identify brake alignment attributes. Despite its excellent performance in real time, the setup cost creates a bottleneck for actual implementation.

The US patent from Sanchez, et al. [21], applies artificial vision for monitoring rolling stock using cameras mounted on the train. Currently, high speed trains such as TGV and bullet train use camera mounts to manually monitor the trains movements. However, videos captured using a camera system onboard a train is bound to induce numerous noises into the video data.

Kazanskiy and Popov [22], introduced a framework to integrate a lighting structure with anti-glare to record high contrast undercarriage videos which is further compressed for quick processing to discover trains on tracks for monitoring rolling stock. This method gave a recipe for automating rolling stock in real time, notwithstanding the procedure for bogie object extraction.

Freid [23], provided an experimental setup under the train with lights focused on the bogie which is captured with a video camera. The work develops an algorithm using straightforward edge recognition techniques for isolating axle box and analyzing its heating profile by using thermal cameras. This model gives an understanding of the TRSE problem for automation and the need for research.

In [24] and [25], the authors offered a 3D reconstruction of the bogie parts for monitoring rail wheel surfaces and contact strips. The methods show effectiveness in identifying surface defects using 3D models by perfectly reconstructing moving parts. However, they are computationally inefficient for processing in real time. Further, it shows the difficulty in modelling defective surfaces for every possible problem beforehand in 3D.

The literature illustrates relatively small number of computer vision state-of-the-art algorithms that are being researched for TRSE. Moreover, the models from literature are inefficient to incorporate the TRSE process for micro level examination of bogie parts specifically. The goal of remotely monitoring system for TRSE is to identify defective and non-working parts that can be repaired timely to prevent mishaps. Earlier proposed frameworks offer very little inclination for research towards remote monitoring of TRSE. Hence, this paper proposes a new orientation to the TRSE with solutions to assist trained railway personnel for better performance. This work uses visualization techniques as a pair of virtual eyes to help checking of each bogie part remotely.

TRSE with video data has been attempted previously [26], [27], [28], [29] using active contours with shape prior models. The performance reported by these models were exceptionally good in terms of segmentation accuracy. However, these models were limited by their ability to provide the required accuracies due to homogeneous nature of pixels in the video sequence.

This gap in segmentation accuracy has been improved by applying local texture information around the object of convergence in the objective function defining the active contours. Moreover, the methods were derived from Chan Vese active contour models which will be discussed exclusively in the following sections. Traditionally, texture and shape based active contours use texture information in a region for shape segmentation [30], [31], [32], [33], [34]. In contrast, this work proposes to segment shape rich texture objects through contour evolution performed sequentially on a shape prior model.

This work proposes to segment video objects using the texture and shape based active contour models successively. The novelty lies in extracting shape and texture of the bogie parts accurately. Contrasting to the preceding results in [26] or correlated works [35], [36], [37], the proposed model in this paper will serially supervise texture and shape segmentation of TRSE video objects with inhomogeneity.

III. Methodology

This section starts by presenting a brief introduction about Chan Vese active contour models [38]. Further, it provides detailed methodology about the proposed successive texture and shape-based model for bogie part video segmentation.

A. Introduction to Chan Vese Active Contours

Active contours or level sets as they are mathematically named are a set of curves in image space that propagate towards the object edges defined by the image gradients. The most popular class of ACs are Chan Vese (CV) model without edges [38], which are described in the following section.

CV active contours evaluate to find a contour C in the image space $f\left(x\right) \forall x \in \left(R, R\right)$ through the energy function modelled as

$$E_{cv} = \omega_1 \int\limits_{C} ds + \nu \iint\limits_{C^I} C\left(x\right) dx +$$
$$\omega_2 \left[\frac{1}{2} \iint\limits_{C^I} \left(f(x) - \Phi^{(I)}\right)^2 dx + \frac{1}{2} \iint\limits_{C^E} \left(f(x) - \Phi^{(E)}\right)^2 dx \right] \tag{1}$$

Where E_{cv} gives the energy function of the CV active contour. The first two terms in eq (1) are parameters controlling the contour's length and area with $\omega_1 > 0, \omega_2 > 0$ and $\nu > 0$. The last bracketed components try to adapt the model $C\left(x\right)$ to the image. The energy function in eq (1) is minimized iteratively to find the object boundaries through an initial contour $f : c \rightarrow R^2$ to $\Phi^{(I)}$ and $\Phi^{(E)}$. Where $\Phi^{(I)}$ and $\Phi^{(E)}$ define the inside and outside portions to Φ.

Eq (1) has been modified by applying the level set models in [38] as

$$E_{cv}(C, \Phi^{(I)}, \Phi^{(E)}) = \min_{C, \Phi^{(I)}, \Phi^{(E)}} \omega_2 \left[\iint\limits_{C^I} \left(f(x) - \Phi^{(I)}\right)^2 M\left(C(x)\right) \right.$$
$$\left. + \iint\limits_{C^E} \left(f(x) - \Phi^{(E)}\right)^2 (1 - M\left(C(x)\right)) dx \right] \tag{2}$$
$$+ \omega_1 \int\limits_{C} \left|\nabla M(C(x))\right| dx$$

Where, $M_\varepsilon(C) = \frac{1}{2}\left[1 + \frac{2}{\pi}\tan^{-1}\left(\frac{C}{\varepsilon}\right)\right]$, $C \in \mathbb{R}$ gives the Heaviside function. The above equation is iteratively updated through gradient descent minimization as

$$C^t = -\delta(C)\omega_2((f(x) - \Phi^{(I)})^2 - \ (f(x) - \Phi^{(E)})^2) - \omega_1 \nabla . \frac{\nabla C(x)}{|\nabla C(x)|} \tag{3}$$

Here $\delta(C) = \frac{1}{\pi} . \frac{\varepsilon}{\varepsilon^2 + C^2}$ is the delta function and iterative adaptations of $\Phi^{(I)}$ and $\Phi^{(E)}$ are initiated with

$$\Phi^{(I)} = \frac{\iint\limits_{C} f(x) M(C(x)) dx}{\iint\limits_{C} M(C(x)) dx} \tag{4}$$

$$\Phi^{(E)} = \frac{\iint\limits_{C} f(x)(1 - M(C(x))) dx}{\iint\limits_{C} (1 - M(C(x))) dx} \tag{5}$$

Where $\Phi^{(I)}$ gives the intensity averages on the inside and $\Phi^{(E)}$ gives the same on the outside of the contour. The major drawback in CV model has been the assumption that the above intensities are homogeneous. This assumption fails to characterize image pixels globally. Hence, pixel intensity inhomogeneity segmentation problem persists in CV models which results in improper segmentations which has shown improvement using shape prior CV [39]. However, texture of the object in real time applications plays an important health detection factor and hence we propose our model called successive texture and shape-based AC model (STSAC).

B. STSAC Model

This section presents a new AC model that fits the evolving contour to the objects in an image based on additional information in the form of texture apart from previously used shape image. Consequently, a serial texture and shape influenced level set model is being formulated by describing each model separately.

1) Texture Features: Here, texture features are extracted using the most successful texture abstraction algorithm called Local Binary Patterns (LBP) [40]. The algorithm operates in a set pixel neighbourhood to identify the texture locally across the entire image. The central pixel in a region is compared with the neighbourhood pixels to binarize the region around the central pixel.

For a colour video frame $v\left(x, 3\right) \in R^+$, where x gives pixel location and the number 3 represents three colour planes RGB. The algorithm uses either 3 (r=1) or 12 (r=2.5) pixel neighbourhoods. The lbp texture code is modelled across a central pixel $\left(x_c, 3\right)$ as

$$lbp\left(x_c, 3\right) = \sum_{i=1}^{3} \sum_{j=1}^{P} s\left(g_p - g_c\right) 2^P \tag{6}$$

Where,

$$s\left(x\right) = \begin{cases} 1 & \forall \ x \geq 0 \\ 0 & Otherwse \end{cases} \tag{7}$$

The g_c denotes the gray value at $\left(x_c\right)$ and g_p represents the gray value in the neighbourhood of g_c. The variable P is the number of pixels around g_c. Using the above texture function, the active contour texture energy functional is formulated as

$$\begin{aligned} E_{texture}\left(\varphi\right) &= d^2\left(\varphi_0, \varphi_t\right) \\ &= \iint\limits_{C} \left(-H\left(\varphi_0\right) + H\left(F_{tf}\right)\right)^2 \delta\left(x\right) dx \end{aligned} \tag{8}$$

The term $F_{tf} = lbp\left(v\left(x\right)\right)$.

2) Shape Image: To achieve clean object boundaries on bogie video data, in addition to texture, the chapter proposes to use shape knowledge. Here, the shape is modelled as a zero-active contour for the first video frame of the bogie video sequence. Signed distance function computes the relation between the shape prior and the initial contour. Eventually, the evolving contour computes this signed distance function to move towards the zero contour through gradient minimization.

The energy functional with prior shape knowledge is given as

$$E^{S}(C,\phi_{S}^{(I)},\phi_{S}^{(E)})=\iint\limits_{C}\left(H\left(\phi\left(x\right)\right)-H\left(\phi_{S}\left(x\right)\right)\right)^{2}\delta\left(\phi\right)dx \tag{9}$$

Where E^S gives the shape energy function of the active contour. The energy function in eq'n (9) is minimized iteratively to find the object boundaries through an initial contour $f : c \to R^2$ to ${\phi_S}^{(I)}$ and ${\phi_S}^{(E)}$. Where ${\phi_S}^{(I)}$ and ${\phi_S}^{(E)}$ define the inside and outside portions to shape boundary.

3) Level Set Formulation: The intensity inhomogeneity in train bogie videos for accurate segmentation of bogie parts with both shape and texture measures simultaneously can be formulated as a serial texture shape based level set energy functional defined as

$$E^{S\_TS}=E_{texture}+\tau.E^{Shape} \tag{10}$$

Where τ is the delay between texture prior information and shape prior information applied as pre-knowledge to the ACs. The serial texture and shaped energy function is defined as

$$\begin{aligned}E^{S\_TS}(C,\Phi^{(I)},\Phi^{(E)})=&\min_{C,\Phi^{(I)},\Phi^{(E)}}\omega_{2}\left[\iint\limits_{C^{I}}\left(f(x)-\phi^{(I)}\right)^{2}H\left(C(x)\right)\right.\\&\left.+\iint\limits_{C^{E}}\left(f(x)-\phi^{(E)}\right)^{2}(1-H\left(C(x)\right))dx\right]\\&+\omega_{1}\int\limits_{C}\left|\nabla H(C(x))\right|dx\\&+\lambda\iint\limits_{C}\left(-H\left(\varphi_{0}\right)+H\left(F_{tf}\right)\right)^{2}\delta\left(x\right)dx\\&+\tau.\xi\iint\limits_{C}\left(H\left(\phi\left(x\right)\right)-H\left(\phi_{S}\left(x\right)\right)\right)^{2}\delta\left(\phi\right)dx\end{aligned} \tag{11}$$

where λ and ξ are the controls for the texture and shape which can be applied as prior information. All the variables and parameters in eq (11) have the same representation as the previous model in Section III.A. The contour evolution is achieved by applying the gradient descent model to achieve a minimization as

$$\begin{aligned}C^{t}=&\begin{bmatrix}\xi\left(g_{tex}\left(\left|\nabla f_{x}\right|\right)\left|\nabla\phi\left(x\right)\right|\right)\delta\left(\phi\right)\\+\gamma\left(\left(\left(f_{x}-\phi^{I}\right)^{2}+\ \left(f_{x}-\phi^{E}\right)^{2}\right)\delta\left(\phi\right)+\mu\nabla.\dfrac{\nabla C(x)}{\left|\nabla C(x)\right|}\delta\left(\phi\right)\end{bmatrix}\\&+\tau.\lambda(H\left(\phi\left(x\right)\right)-H\left(\phi_{S}\left(x\right)\right))^{2}\delta\left(\phi\right)\end{aligned} \tag{12}$$

Where g_{tex} is the texture function defined using local binary pattern (LBP). The proposed serial texture shape active contour model has shown improved performance in the segmentation of objects in real time video sequences that are inhomogeneous with the surroundings.

4) STSAC Contour evolution: For contour evolution in eq (12), is implemented on a machine using the model from [39], which is given by

$$C_{new}=C_{old}+\Delta t\frac{dC}{dt}+\Delta tC_{Texture} \tag{13}$$

After the texture evolution is completed when the following stopping criteria $\left\|\nabla\phi^{n}\left(x,y\right)\right\|\leq\varepsilon$ is attained. Where ε is the minimum gradient between the last two consecutive contour evolutions in texture prior active contour model. Once, the texture prior stops, the outer edges forded due to texture evolution show a poor boundary or shape representation due to inhomogeneity in the object boundaries. To reconstruct the shape of the textured segmentation, we now apply the shape prior model on the textured contour. This is unlike the previous models, where the texture and shape information are fused into a single prior model for the active contour. This single prior model work well if the texture and shape models are perfectly aligned in 2D space. However, in real time computer vision applications getting a perfectly fused texture and shape model is quite a difficult process. Hence, our proposed serial texture shape based contour evolution can handle both texture and shape influence accurately than the previous models for real time computer vision applications. The shape contour evolution

$$C_{new}=\left(C_{Old}+\Delta t\frac{dC_{old}}{dt}\right)^{Texture}+\Delta tC_{Shape} \tag{14}$$

Where $\frac{dC}{dt}$ predicts the rough variations in the right-hand side of the eq's(13, 14) and $\Delta t=\frac{0.48}{\max(|C_{old}|)}$ gives the step size in time. Fig.2 shows the comparison of STSAC against the previous texture shape fused models.

Fig. 2(g) and (h) are the outputs from the serial textured shape based active contour model (STSAC), which are better than the previously proposed models. In the next section we present the experiments and related analysis of the proposed method STSAC for TRSE video datasets in Table I.

TABLE I. VIDEO DATASETS CAPTURED FOR TESTING THE PROPOSED METHODS.

Experiments	Name	Number of Frames
D-1/E-1	Bogie Video Recorded at 6.40AM	$90\times36=3240$
D-2/E-2	Bogie Video Recorded at 12.40PM	$90\times40=3600$
D-3/E-3	Bogie Video Recorded at 4.20PM	$90\times32=2880$
D-4/E-4	Bogie Video Recorded at 6.50PM	$90\times26=2340$
D-5/E-5	Defective Video on 12.40PM Train	$40\times4+90\times38=3600$

IV. RESULTS AND DISCUSSION

This section discusses the datasets capturing mechanism and their characteristics in detail. Next, an extensive experimentation of the proposed algorithm on the considered TRSE problem for extracting bogie parts is presented. The results obtained are evaluated and analysed with benchmark algorithms already proposed on TRSE.

A. The Datasets

To test the proposed methods for their ability in extracting objects from videos captured under ambient conditions, the train rolling stock video database is built. The videos are recorded near to an Indian Railway station using the setup shown in Fig. 3. The figure shows an arrangement not more than 3 feet from the moving train. All the videos were recorded when the train was entering the station for a halt.

The handbook on train rolling stock examination was followed during the video capture. Accordingly, all trains in the dataset were recorded when the train was running at around 30KMPH. However, Digital single lens reflex (DSLR) record

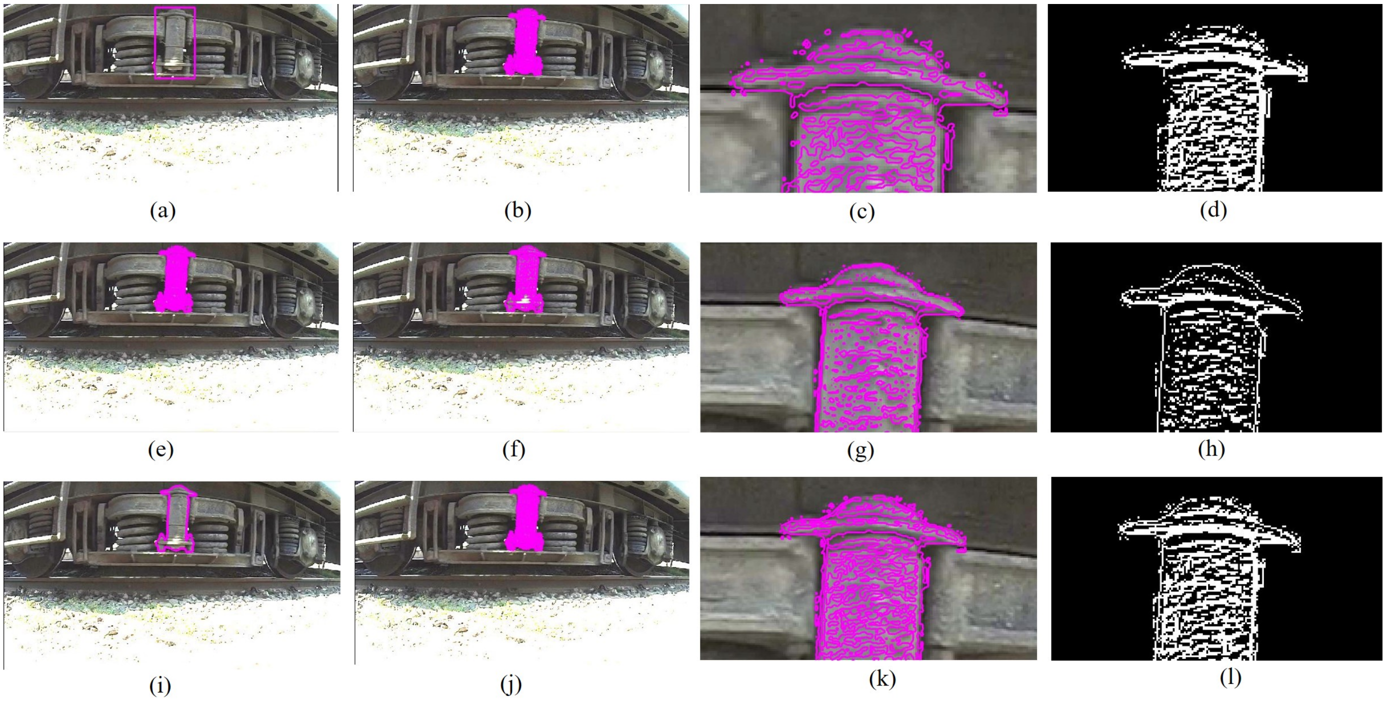


Fig. 2. Comparison of only texture and shape prior based active contour models with the proposed serial texture shape based active contour model. (a) Initial contour, (b) Only texture based AC model after 50 iterations, (c) Zoomed View of the object from (b), (d) Zoomed view of the segmented result from (b), (e) Our proposed initial texture model, (f) serial shape model, (g) Zoomed from (f), Segmented result from the proposed model STSAC in (f), (i) shape only AC model and (j) fused shape texture model, (k) zoomed shape textured model, (l) zoomed segmented output from (k).

at around 30 frames per second(fps), which induce a consider amount of motion blur to the video object data. Hence, to do away with blurring, the videos were recorded with Isaw sports action camera as can be seen in Fig. 3.

Fig. 3. Train Bogie Video Recording using High Speed Sports action Camera.

The Isaw sports action camera captures video at a max frame rate of 240fps. The camera also possesses a wide-angle lens with a 52^0 angle, that is capable of full bogie into a video frame from the centre. Fig. 4 shows an array of bogie video frames that were recorded by the visual sensor near the tracks. A total set of 4 train bogie videos were filmed at separate time stamps on a day. The advantage of in this approach gives an opportunity to test the proposed methods ability to overcome the effects of ambient lighting on segmentation quality.

In this paper, operation on each video sequence is considered as an experiment. Hence, 4 experiments were performed for testing the proposed methods. Moreover, a 5^{th} experiment is added to test the proposed algorithms capabilities in segmenting defective bogie parts. Unfortunately, there were no comprehensive defects in the recorded videos. Hence, the 5^{th} video has been handcrafted by extracting frames and deliberately inducing defects. Thereupon the defective frames were incorporated back to form a defective train sequence which resulted in a defective train bogie video for the 5^{th} experiment.

Fig. 4. Video Frames of a recorded Bogie for Experimentation.

Fig. 5 shows the video frames that have been photoshopped with defects to bogie parts. The purpose of the proposed segmentation algorithms is resolved to satisfaction if it manages to segment the defective part through the prior knowledge of the healthy bogie part. This capability of the proposed frameworks in this work increases the scope for automation. Finally, Table I shows the experimental valuations performed on the five different datasets throughout the thesis. Fig. 6 gives a visualization of the datasets from Table I.

B. Bogie Parts Segmentation

Fig. 7 projects the results of the segmentation on bogie video frames for the 10 different parts as shown in column 1.

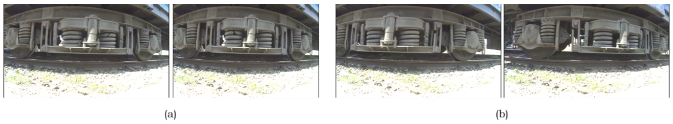
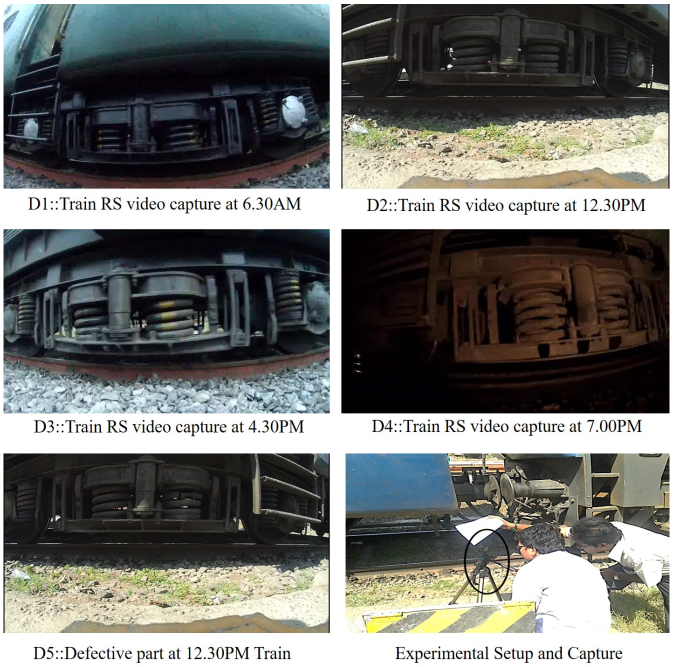


Fig. 5. Bogie part defects induced with photoshop. (a) Spring breaks and (b) Binding rod breaks.


D1::Train RS video capture at 6.30AM

D2::Train RS video capture at 12.30PM

D3::Train RS video capture at 4.30PM

D4::Train RS video capture at 7.00PM

D5::Defective part at 12.30PM Train

Experimental Setup and Capture

Fig. 6. Visualization of datasets.

Visual analysis of the Fig. 7 shows that the proposed STSAC segments bogie parts by retaining both shape and texture. The STSAC method requires the selection of bogie part that the user needs to check from the recorded video sequence along with the initial location of object. Fixing these hyperparameters can automate the part extraction process using STSAC.

The first column in Fig. 7 is the shape prior model extracted from the train bogie frame. Texture extracted using LBP is in second column and the next four columns show the curve evolution for all ten bogie parts after fixed number of 15 iterations. The last column shows the segmented bogie parts from the proposed STSAC algorithm. A visual comparison with similar algorithms is shown in Fig. 2.

The standard parameters ξ, γ and μ are kept constant and low for smooth contour propagation on the video frame. The initial texture contour evolution stopped in most of the cases at around 28^{th} iteration and from then on, the shape evolution lasted for 56^{th} iteration.

Compared to the shape or texture only based models the computation cost is little on the higher side for STSAC model as it had to run serially. The maximum number of iterations recorded were 68, in case of springs. For the entire train of 17 bogies in our dataset 1, our program took 0.35 hours for evaluation on an 8GB RAM with 2.4GHz intel processor on MATLAB software.

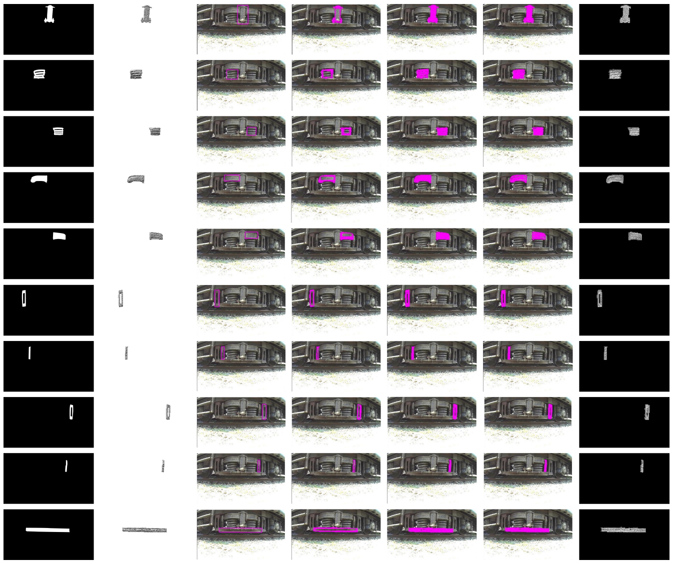


Fig. 7. Bogie part segmentation outputs on the 1^{st} dataset using the proposed STSAC model. (Zoom In for better visibility).

The results demonstrate that the STSAC is capable of segmenting the region of interest objects given their texture and shape information accurately. However, from Fig. 1, we see that the similar AC models with only texture or shape or both had lost either texture or shape in the final output segment on real time video data. Our proposed model has retained good amount of texture information when shape is being reconstructed during the curve evolution process. The segmented outputs could be evaluated further for identifying their defects or their running life by comparing them with the available reference models.

This part of the work is achieved by comparing the segmented bogie part with the reference parts captured from the railway workshop. The reference parts are binarized and are called Ground Truth frames. These images appear as the first row in Fig. 7.

C. Defect Detection

Fig. 8 shows the defect in the biding screw and the contour evolution using the proposed algorithm to identify the defect.

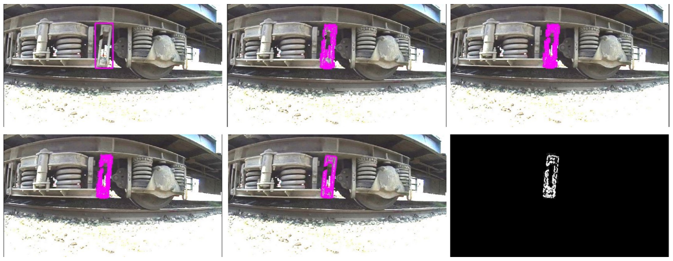


Fig. 8. Defect identification using the proposed STSAC model.

The previous models came close to the above result but showed multiple regions around the defect region making it difficult to identify the actual breakage point as shown in fig.8. Similar results were obtained for cut defects of dimension 0.2mm thick on all bogie parts. For defects less than 0.2mm, our proposed model failed to extract the defect in the segmented output.

D. Bogie Video Data Analysis

This section parametrically evaluates the performance of the previously proposed algorithms against the proposed STSAC model across the datasets in Table I. The following parameters are computed for benchmarking the performance of the proposed algorithm against state-of-the-art previous models.

Detection sensitivity (DS) is defined as the parameter for measuring the performance of the segmentation methods. DS calculates the number of times an object is segmented correctly in the entire video sequence against the times it could not. The number of instances in the entire video sequence an object has been identified or segmented accurately is called True Positive (TP). Similarly, the opposite of TP is True Negative (TN), which gives the number of instances an object of interest could not be segmented properly. The expression for DS is

$$DS = \frac{TP}{TP + FN} \tag{15}$$

A DS measure of 0 indicates a failed segmentation and a one indicates a successful segmentation. The parameter is important for the proposed TRSE application to measure the accuracy of the methods in segmenting the required bogie parts in the entire video sequence.

The Mean Absolute Distance (MAD) calculates the deviation in the segmentation result of an algorithm from the actual required output. It is calculated by subtracting the obtained result with the ground truth GT of the object given as

$$MAD = \left| \frac{\left|S^O - GT\right| - \sum_{i=1}^{n} \left|S^O - GT\right|}{n} \right| \tag{16}$$

where 'n' is the number of pixels.

The Boundary Mean Absolute Distance (BMAD) parameter illustrates the deviation between boundaries of resulting segmented objects S_O^b to ground truth GT_b. The following expression for BMAD is

$$BMAD = 1 - \left| \frac{\left|S_b^O - GT_b\right| - \sum_{i=1}^{n} \left|S_b^O - GT_b\right|}{n_{Boundary}} \right| \tag{17}$$

Where, '$n_{Boundary}$' is the number of pixels in the GT object boundary.

The Normalized Mutual Information (NMI) is a measure to determine how close the resulting segmented object is to the ground truth object. It gives a degree of common information in images. NMI is given as

$$NMI = \frac{MI\left(I, GT\right)}{\sqrt{H\left(I\right) H\left(GT\right)}} \tag{18}$$

Where $MI\left(I, GT\right) = -\sum_{I,GT} p\left(I_i, GT_i\right) . \log\left(\frac{p(I_i, GT_i)}{p(I_i)p(GT_i)}\right)$ is the mutual information and $H\left(\bullet\right)$ is the entropy. MNI has a scale of 0 to 1, where 1 means highest segmentation accuracy.

Finally, the train rolling stock examination is being performed as a real time operation. Consequently, it becomes necessary for the proposed algorithms to compete in speed of execution. Hence, Number of Iterations – Model Speed (MS) measures the number of iterations in which the initial contour deforms and encompasses the object of interest.

Initially, first two parameters, DS and MAD are calculated and averaged on a set of 2000 bogie video frame segmentation outputs. Table II presents the computed values of various AC models on our rolling video datasets in Table I. The parameters are average across datasets. The values point to a conclusion that there has been a direct link between DS and the ambient lighting in which videos are recorded. Similarly, it can be seen that the STSAC performed quite well over the AC models with only shape (SP_AC), texture (TP_AC) and fused shape texture (FSTP_AC) prior models.

E. Parametric Analysis Against the State-of-the-Arts

Here, the proposed STSAC model is being validated against the state-of-the-art AC models parametrically with DS, MAD, BMAD, NMI and NI defined in the previous section. The computed parameters are plotted in Fig. 9. Fig. 9 plots the average DS of the localized AC models shape, texture and fused prior models and our proposed STSI_AC for all the bogie parts. The values plotted in Fig. 9 are averaged across all 4 datasets from Table I. The proposed approach showed that it can detect texture in inhomogeneous regions provided a weak shape and texture prior model as references.

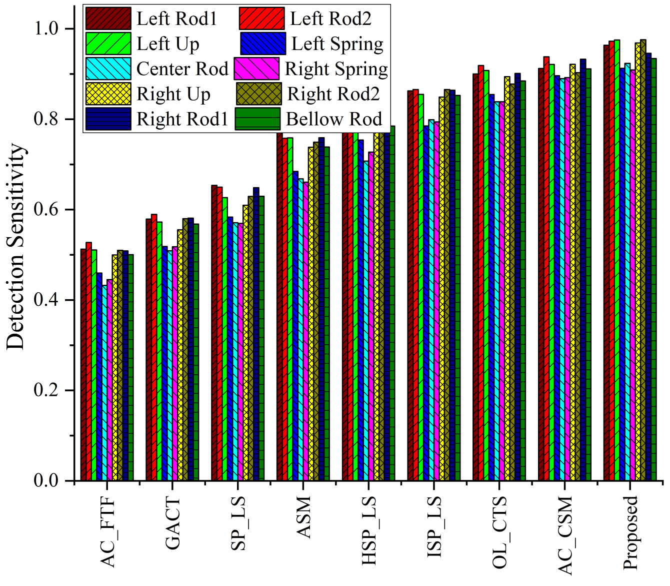


Fig. 9. Detection sensitivity of the proposed approach (STSAC) with the state of the art similar models.

The other plots for the remaining four parameters are presented in Fig. 10(a), (b), (c) and (d) show that the proposed STSI_AC is the better segmentation algorithm in the batch.

All the parameters are computed like the previous chapter. The proposed STSAC has been successfully implemented to extract texture and shape serially using AC models. However, the occlusion resistance has been on the lower side when compared to the shape prior models. Apart from that the model in this work provides additional information on the segmented bogie part for better decision making on the quality of the object.

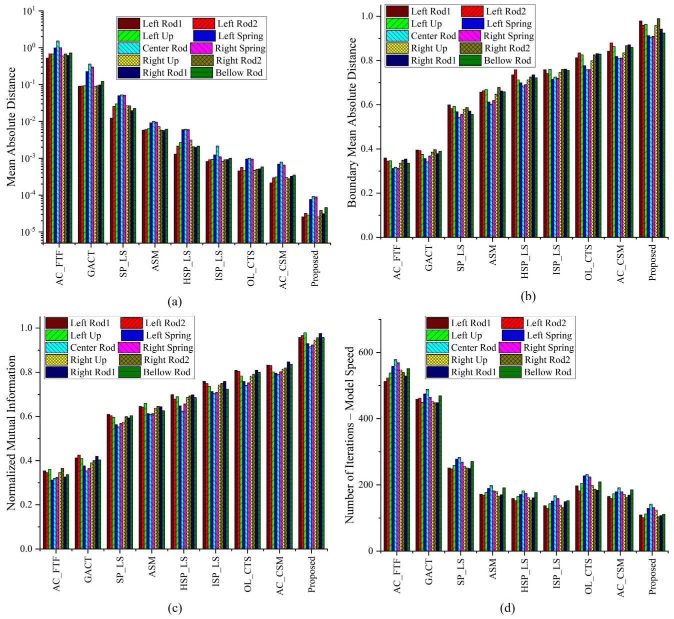


Fig. 10. Performance measures of the proposed algorithm against the state-of-the-art models for automated train rolling stock video segmentation: (a) Mean absolute distance, (b) Boundary mean absolute distance, (c) Normalized mutual information and (d) Number of iterations.

V. Conclusion

To improve the quality of the segmented bogie parts, this thesis proposes the second active contour model called serial texture and shape influenced active contour (STSAC). Traditionally, texture and shape based active contours use texture information in a region for shape segmentation. In contrast, this work proposes to segment shape rich texture objects through contour evolution performed serially on a shape prior model. This resulted in an improvement in segmented bogie parts over the previous model. This work presents a real-time computer vision problem and generates a formidable solution using novel serial texture shape prior active contour models. The objective of the real-time computer vision problem is to segment a train bogie part for inspection using the high-speed video data of the train moving at 30KMPH. The video of the moving train was captured with a high – speed camera at 240fps. Serial texture shape prior active contours algorithm has been developed which uses first the texture prior and then the shape prior serially to extract objects texture by preserving its shape. This was quite different from similar algorithms which uses either texture or shape or both in fused form as priors, resulting in less than accurate segmentation outputs on real time video data. However, the proposed model had bettered the segmentation outputs both visually and parametrically over the existing models. Hence, the proposed method shows prospects of inducing as a platform for segmenting train bogie parts for automated train rolling stock examination in real time.

References

[1] Maik Rosenberger and Rafael Celestre. Smart multispectral imager for industrial applications. In *2016 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 7–12. IEEE, 2016.

[2] Petra Tatzer, Markus Wolf, and Thomas Panner. Industrial application for inline material sorting using hyperspectral imaging in the nir range. *Real-Time Imaging*, 11(2):99–107, 2005.

[3] Frank Joachim Grote and Carsten Buchwald. Beverage bottling plant having an apparatus for inspecting bottles or similar containers with an optoelectric detection system and an optoelectric detection system, December 20 2016. US Patent 9,522,758.

[4] Hong-Dar Lin and Kuan-Shen Hsieh. Automated distortion defect inspection of curved car mirrors using computer vision. In *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, page 361. The Steering Committee of The World Congress in Computer Science, Computer …, 2015.

[5] H Golnabi and A Asadpour. Design and application of industrial machine vision systems. *Robotics and Computer-Integrated Manufacturing*, 23(6):630–637, 2007.

[6] Desmond K Moru and Diego Borro. A machine vision algorithm for quality control inspection of gears. *The International Journal of Advanced Manufacturing Technology*, 106(1-2):105–123, 2020.

[7] Jay H Heizer and Barry Render. *Operations management*, volume 1. Pearson Education India, 2008.

[8] Fei Yan, Chunhai Gao, Tao Tang, and Yao Zhou. A safety management and signaling system integration method for communication-based train control system. *Urban Rail Transit*, 3(2):90–99, 2017.

[9] Michal Kuffa, Daniel Ziegler, Thomas Peter, Fredy Kuster, and Konrad Wegener. A new grinding strategy to improve the acoustic properties of railway tracks. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 232(1):214–221, 2018.

[10] Xiaoxuan Wang, Hailin Jiang, Wenzhe Sun, and Tao Tang. Efficient dual-association resource allocation model of train-ground communication system based on td-lte in urban rail transit. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 2006–2011. IEEE, 2016.

TABLE II. Average DS and MAD for each dataset with the proposed model and the state of the art level set models.

Methods	D-1		D-2		D-3		D-4		D-5	
	DS	MAD	DS	MAD	DS	MAD	DS	MAD	DS	MAD
AC_FTF [37]	0.45	0.89	0.49	0.86	0.51	0.84	0.32	0.95	0.39	0.82
GACT [41]	0.44	0.91	0.51	0.81	0.52	0.79	0.35	0.93	0.33	0.84
SP_LS [28]	0.58	0.84	0.59	0.8	0.59	0.81	0.4	0.9	0.44	0.89
ASM [42]	0.79	0.39	0.78	0.32	0.77	0.3	0.72	0.35	0.7	0.3
HSP_LS [43]	0.68	0.47	0.71	0.52	0.75	0.56	0.42	0.55	0.43	0.86
ISP_LS [27]	0.7	0.43	0.75	0.44	0.78	0.42	0.51	0.48	0.45	0.87
OL_CTS [44]	0.75	0.41	0.75	0.4	0.75	0.4	0.43	0.31	0.47	0.82
AC_CSM [45]	0.84	0.34	0.88	0.38	0.89	0.33	0.47	0.32	0.82	0.31
STSAC (Proposed)	0.91	0.31	0.94	0.25	0.92	0.28	0.83	0.34	0.88	0.35

[11] Yong Qin, Baojun Yuan, and Si Pi. Research on framework and key technologies of urban rail intelligent transportation system. In *Proceedings of the 2015 International Conference on Electrical and Information Technologies for Rail Transportation*, pages 729–736. Springer, 2016.

[12] Arzoo Naghiyev, Sarah Sharples, Brendan Ryan, Anthony Coplestone, and Mike Carey. Expert knowledge elicitation to generate human factors guidance for future european rail traffic management system (ertms) train driving models. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 231(10):1141–1149, 2017.

[13] Satish Chandra et al. *Railway engineering*. Oxford University Press, Inc., 2008.

[14] JS Mundrey. *Railway track engineering*. Tata McGraw-Hill Education, 2009.

[15] Ashutosh Kumar Banerji. Railways and rolling stock engineers-challenges ahead. *Technical Note*, 2005.

[16] Konkan railway automated train examination system, 2008. https://www.youtube.com/watch?v=8RU54XZc9so.

[17] Sundaravalli Narayanaswami. Urban transportation: innovations in infrastructure planning and development. *The International Journal of Logistics Management*, 2017.

[18] Alessandro Sabato and Christopher Niezrecki. Feasibility of digital image correlation for railroad tie inspection and ballast support assessment. *Measurement*, 103:93–105, 2017.

[19] John M Hart, Esther Resendiz, Benjamin Freid, Steven Sawadisavi, CPL Barkan, and N Ahuja. Machine vision using multi-spectral imaging for undercarriage inspection of railroad equipment. In *Proceedings of the 8th world congress on railway research, Seoul, Korea*, volume 18, 2008.

[20] HyunCheol Kim and Whoi-Yul Kim. Automated inspection system for rolling stock brake shoes. *IEEE transactions on instrumentation and measurement*, 60(8):2835–2847, 2011.

[21] Angel Luis Sanchez-Revuelta and Carlos-Javier Gomez Gomez. Installation and process for measuring rolling parameters by means of artificial vision on wheels of railway vehicles, September 15 1998. US Patent 5,808,906.

[22] NL Kazanskiy and SB Popov. Integrated design technology for computer vision systems in railway transportation. *Pattern Recognition and Image Analysis*, 25(2):215–219, 2015.

[23] Benjamin Freid, Christopher PL Barkan, Narendra Ahuja, John M Hart, Sinisa Todorvic, and Nicholas Kocher. Multispectral machine vision for improved undercarriage inspection of railroad rolling stock. In *Proceedings of the Ninth International Heavy Haul Conference Specialist Technical Session–High Tech in Heavy Haul, Kiruna, Sweden*, pages 11–13, 2007.

[24] Leszek Jarzebowicz and Slawomir Judek. 3d machine vision system for inspection of contact strips in railway vehicle current collectors. In *2014 International Conference on Applied Electronics*, pages 139–144. IEEE, 2014.

[25] Yu Zhang, Jia-yuan Hu, Jin-long Li, Hai-qing Wang, et al. The application of wtp in 3-d reconstruction of train wheel surface and tread defect. *Optik*, 131:749–753, 2017.

[26] PVV Kishore and Ch Raghava Prasad. Computer vision based train rolling stock examination. *Optik*, 132:427–444, 2017.

[27] Ch Raghava Prasad and PVV Kishore. Performance of active contour models in train rolling stock part segmentation on high-speed video data. *Cogent engineering*, 4(1):1279367, 2017.

[28] PVV Kishore and Ch Raghava Prasad. Shape prior active contours for computerized vision based train rolling stock parts segmentation. *International Review on Computers and Software (I. RE. CO. S.)*, 10:1233–1243, 2015.

[29] PVV Kishore and Ch Raghava Prasad. Train rolling stock segmentation with morphological differential gradient active contours. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1174–1178. IEEE, 2015.

[30] Farhan Riaz, Sidra Naeem, Raheel Nawaz, and Miguel Coimbra. Active contours based segmentation and lesion periphery analysis for characterization of skin lesions in dermoscopy images. *IEEE journal of biomedical and health informatics*, 23(2):489–500, 2018.

[31] Tiejun Yang, Yaowen Chen, and Zhun Fan. Vegetation segmentation based on variational level set using multi-channel local wavelet texture and color. *Signal, Image and Video Processing*, 12(5):951–958, 2018.

[32] Shiyu Luo, Ling Tong, and Yan Chen. A multi-region segmentation method for sar images based on the multi-texture model with level sets. *IEEE Transactions on Image Processing*, 27(5):2560–2574, 2018.

[33] Nawal Houhou, Jean-Philippe Thiran, and Xavier Bresson. Fast texture segmentation model based on the shape operator and active contour. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[34] Mingqi Gao, Hengxin Chen, Shenhai Zheng, and Bin Fang. Feature fusion and non-negative matrix factorization based active contours for texture segmentation. *Signal Processing*, 159:104–118, 2019.

[35] Priyambada Subudhi and Susanta Mukhopadhyay. A novel texture segmentation method based on co-occurrence energy-driven parametric active contour model. *Signal, Image and Video Processing*, 12(4):669–676, 2018.

[36] Guo Zhao, Shiyin Qin, and Danyang Wang. Interactive segmentation of texture image based on active contour model with local inverse difference moment feature. *Multimedia Tools and Applications*, 77(18):24537–24564, 2018.

[37] Qinggang Wu, Yong Gan, Bin Lin, Qiuwen Zhang, and Huawen Chang. An active contour model based on fused texture features for image segmentation. *Neurocomputing*, 151:1133–1141, 2015.

[38] Tony F Chan and Luminita A Vese. Active contours without edges. *IEEE Transactions on image processing*, 10(2):266–277, 2001.

[39] Tony Chan and Wei Zhu. Level set based shape prior segmentation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 1164–1170. IEEE, 2005.

[40] Syed Inthiyaz, BTP Madhav, and PVV Kishore. Flower segmentation with level sets evolution controlled by colour, texture and shape features. *Cogent Engineering*, 4(1):1323572, 2017.

[41] Berta Sandberg, Tony Chan, and Luminita Vese. A level-set and gabor-based active contour algorithm for segmenting textured images. In *UCLA Department of Mathematics CAM report*. Citeseer, 2002.

[42] Mina Esfandiarkhani and Amir Hossein Foruzan. A generalized active shape model for segmentation of liver in low-contrast ct volumes. *Computers in biology and medicine*, 82:59–70, 2017.

[43] Xiong Yang, Shu Zhan, Dongdong Xie, Hong Zhao, and Toru Kurihara. Hierarchical prostate mri segmentation via level set clustering with shape prior. *Neurocomputing*, 257:154–163, 2017.

[44] Yu Zhong and Anil K Jain. Object localization using color, texture and shape. *Pattern Recognition*, 33(4):671–684, 2000.

[45] Weiming Hu, Xue Zhou, Wei Li, Wenhan Luo, Xiaoqin Zhang, and Stephen Maybank. Active contour-based visual tracking by integrating colors, shapes, and motions. *IEEE Transactions on Image Processing*, 22(5):1778–1792, 2012.

Exerting 2D-Space of Sentiment Lexicons with Machine Learning Techniques: A Hybrid Approach for Sentiment Analysis

Muhammad Yaseen Khan[1]
Mohammad Ali Jinnah University
Karachi, Pakistan

Khurum Nazir Junejo[2]
Ibex CX
Karachi, Pakistan

Abstract—**Sentiment mining from the textual content on the web can give valuable insights for discernment, strategic decision making, targeted advertisement, and much more. Supervised machine learning (ML) approaches do not capture the sentiment inherent in the individual terms. Whereas the unsupervised sentiment lexicon (SL) based approaches lag behind ML approaches because of a bias they have towards one sentiment than the other. In this paper, we propose a hybrid approach that uses unsupervised sentiment lexicons to transform the term space into a two-dimensional sentiment space on which a discriminative classifier is trained in a supervised fashion. This hybrid approach yields higher accuracy, faster training, and lower memory footprint than the ML approaches. It is more suitable for scenarios where training data is scarce. We support our claim by reporting results on six social media datasets using five sentiment lexicons and four ML algorithms.**

***Keywords*—Hybrid approach; machine learning; sentiment analysis; sentiment lexicons; sentiment space; social media analysis**

I. Introduction

Humans are sentient beings that express emotions through sentiments. The behaviour of an individual is often guided by his (or her) emotions but can be studied by his (or her) sentiments. Sentiments are expressed through writings, speech, and actions. Recently, there is a drastic increase in usage of the online medium such as articles, blogs, e-shopping, online social networking (OSN) sites, e-newspapers, and magazines for expression of sentiments. Many people now present their analysis and stories in the forms of comments, tweets, reviews, and feedback on almost every aspect of life [1]. The automatic quantification of sentiments hidden within these texts can lead to many insights that can help in contextual advertisement [2], determine the popularity of an election or an advertising campaign [3], identify trends in political discourses [4], movie and product review mining [5], [6], and many more application areas.

Sentiment analysis (SA)(or opinion mining) focuses on discovering techniques that decipher these emotions and sentiments from raw text comments, reviews, etc. SA widely focuses on prediction or categorization of *polarity* encompassed within a text. The categorization could be into two categories such as *positive* or *negative* [7]; or even into a third *neutral* category as well [8]. Positive and negative text is sometimes also referred to as *subjective* text, whereas neutral text is referred to as *objective*. Similarly, SA also aims to predict *emotions* (happiness, sorrow, joy, anger, etc.) that are expressed within a text [9]–[11].

Quantifying the sentiments in text documents is not an easy task as they come from various domains, cover a wide variety of topics, and are often unorganized and unstructured. Predominantly two types of approaches exist for SA; supervised machine learning (ML) approaches, and the (unsupervised) sentiment lexicon (SL) based approaches. Supervised approaches are based on ML algorithms such as random forests, support vector machines, logistic regression, etc. They require a labelled set of text documents (referred to as training data) to learn the predictive model. On the other hand, lexicon-based approaches use pre-defined lexical dictionaries [12], thus not requiring labelled training examples. Lexicon based approaches can also be thought of as an expert system or a knowledge-based approach. Supervised approaches have the advantage of achieving higher accuracy, but their reliance on labelled data is a bottleneck as it requires a tedious process of reading through each text document and labelling it as positive or negative accordingly.

On the other hand, SL based approaches do not require labelled training data and thus can be applied directly without learning any training model. However, they suffer from a coverage problem, i.e., they fail to assign a sentiment label to each document. To reach the best of both worlds, we propose a hybrid approach that uses SL to transform the document term space into a two-dimensional sentiment feature space where an ML classifier is learned in a supervised fashion. This hybrid approach yields higher accuracy (or similar accuracy with fewer training examples) than the ML approaches, takes lesser time and memory to train than SL approaches. It is suitable for scenarios where training data is scarce. We support our claim by reporting results on six different online social media datasets (BBC, Digg, MySpace, Runner World, Twitter, and YouTube), using five SL (Afinn, Happiness index, SenticNet, SentiStrength, and SentiWordNet), and four ML algorithms (support vector machines, naïve Bayes, decision trees, and LogitBoost). Thus briefly, the main contribution of this are:

- Proposal and implementation of a hybrid technique for sentiment analysis.
- Study the effectiveness of the proposed approach, in comparison to baseline methods of pure machine learning and lexicon-based methods.

- Evaluation of proposed methodology with the different lexicon, machine learning algorithms and sentiment datasets.

The rest of the paper is organized as follows; literature review is given in Section II, followed by the description of our methodology in Section III. Dataset details, evaluation setup, and performance metrics are also described in this section. Section IV presents the results and the discussion about the performance of the various models. Finally, we conclude and state our future direction in Section V.

II. Related Work

The papers [11] and [13] provide detailed surveys of various ML and SL approaches used in the literature for sentiment analysis. For this study, we discuss here the most standard and widely used SL and ML approaches, followed by a detailed survey of hybrid approaches. We, therefore, structure the related work in three subsections according to these three types of approaches.

A. Lexicon-Based Techniques

All the lexicon-based approaches have in common a dictionary of words (or phrases) having some score that hints towards their polarity. They differ in terms of the source of these words, dictionary size, and the methodology used to assign them a score [1]. The process of building such a lexicon is subjective; therefore, all these dictionaries only have a small overlap. Similarly, a word may be deemed by one expert to have a positive sentiment, whereas others may deem it as neutral or even negative. Furthermore, many words inherently do not contain a positive or negative orientation, but it is the context in which they are used that makes their polarity positive or negative [14]. Due to the aforementioned reasons, there is no standard lexicon dictionary. Often there might be a tweet, or a blog, that contains no word that has a polarity, in which case, it is said that the lexicon does not cover that particular document and no score is assigned to it. This problem is referred to as the coverage problem [15]. Lexicon based approaches have a major benefit of not requiring any data for training, and thus can be used as off the shelf solutions. We chose the four lexicons provided by the *iFeel* utility [12], [16], namely, *SenticNet*, *SentiWordNet*, *Happiness Index*, and *SentiStrength*. The fifth lexicon used is *AFINN* [17]. These five approaches are described below.

1) Happiness Index: Happiness Index proposed by [18], calculates average psychological scores and frequency for the Affective Norms for English Words (ANEW) dictionary [19]. ANEW is a set of one thousand and thirty-four words bearing scores for psychological valence (good–bad), dominance (strong–weak), and arousal (active–passive) and their semantic differentials. Based on these dimensions, words are assigned a happiness score on the scale between 1 to 9. In our study, we consider the words with scores between 1 to 4 as negative words, whereas words with scores between 5 to 9 are regarded as positive words.

2) SentiStrength: SentiStrength is a hybrid approach that combines supervised and unsupervised classification methods [20]. It consists of two thousand three hundred and ten words exhibiting sentiments based on the Linguistic Inquiry and Word Count (LIWC) dictionary [21]. Each word has a human-assigned sentiment score from -5 to 5. Words having a score from -5 to -1 are considered as negative, whereas words having a score from 1 to 5 are regarded as positive words. SentiStrength divides the given documents into words and removes punctuations and emoticons, but in our study, we already remove these artefacts during the pre-processing phase. An associated sentiment score defined by SentiStrenght is then mapped to each word. The scores are then summed up for the positive and negative category. The category with the higher score is then marked as the category of that particular document.

3) SenticNet: SenticNet is a concept-level knowledge base that provides a set of sentics, semantics, and polarity for 100,000 concepts. It uses AI and semantic web techniques on web content to recognize, process, and interpret natural language opinions. SenticNet assigns a sentiment score to each concept between the range from -1 to 1 [22]. In this study, we consider words with values less than 0 as negative words, whereas words with scores higher than 0 are considered as positive words.

4) AFINN: AFINN [23] is a sentiment lexicon worked out by Finn Årup Nielsen group that is based on data generated by Twitter. It is based on 1000 tweets used in “Twitter mood maps reveal emotional states of America” [24]. This lexicon has a total of 2,477 words labelled within the integral range of ±[1...5], where the positive and negative signs indicate whether the word is positive or negative, respectively. The larger the score, the more intense, is the sentiment.

5) SentiWordNet: SentiWordNet [8] is based on popular English lexical dictionary ‘WordNet’ [25]. SentiWordNet has more than 117,000 words, including nouns, verbs, and adjectives. Each word is assigned a positive and as well as a negative score within the [0. . . 1] range.

B. Machine-Learning Methods

Supervised ML aims to learn a predictive model from information encompassed in a given (training) dataset and then apply that model to another (testing) dataset for predictions. After document pre-processing, supervised learning is performed using a cross-validation approach, and yielded results are saved accordingly. [26] provides a detailed survey of supervised ML methods used for sentiment analysis. For the purpose of this study we choose three benchmark machine learning algorithms namely, $LogitBoost$ (LB), $naïve\ Bayes$ (NB), and support vector machines (SVM).

1) Naïve Bayes: NB is a probabilistic classifier that has been widely used for text classification in general and sentiment analysis in particular [27]. Based on Bayes theorem, it assumes a naïve independence i.e. all attributes are independent of each other given the category label. Even though this assumption does not hold in most cases, the resulting model is easier to fit and works remarkably well for large dimensional problems. NB predicts the class with the most probable hypothesis. Thus NB assigns the label $\hat{y} = c_k$ for the document X according to the following equation:

$$\hat{y} = \underset{k\in\{P,N\}}{\arg\max}\ p(C_k)\prod_{i=1}^{n} p(x_i|C_k) \tag{1}$$

Where x_i represents the i^{th} attribute (or feature) of document X, and P and N represent the positive and negative class labels, respectively.

2) Decision Trees: Decision trees are well known non-parametric supervised learning methods used for classification and regression. Although initially proposed more than three decades ago [28], newer versions are still popular today [29]. DT predicts the value of the target variable by inferring simple if-then-else type decision rules from the dataset. DT are visualized as a tree structure in which internal nodes perform a check on the attribute values, whereas the leaf nodes correspond to an outcome of the target attribute. For prediction, each record is traversed through the root node until it reaches the leaf node where it is assigned a prediction label that is associated with that particular leaf node. Attributes are selected using an information-theoretic measure called entropy. We use an advanced implementation of the DT algorithm named as classification and regression trees (CART) [30].

3) Support Vector Machines: SVM is a widely used discriminative classifiers for classification of text and sentiments [31], [32]. It is a binary classifier that projects each document as a point in a higher dimensional feature space such that the points belonging to the different categories are separated as far as possible. A hyperplane is then learned in this feature space to discriminate between the points of the two categories. The learned hyperplane is an optimal hyperplane i.e. it maximizes the gap (or distance or margin) between the closest points of the two categories. This help SVM achieve a better generalization over the unseen data. Decision function for SVM is as follows [33]:

$$\hat{y_i} = \begin{cases} P & \text{if } \mathbf{w^T} \cdot \mathbf{X}_i - b \geq 1 \\ N & \text{if } \mathbf{w^T} \cdot \mathbf{X}_i - b \leq -1 \end{cases} \tag{2}$$

Optimization of above objective function through maximizing marginal width and penalizing the vectors that fall within the hyper plane leads to the following decision function:

$$\arg\min_{\mathbf{w},\xi,b} \begin{cases} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \\ \sum_{i=1}^{n}\alpha_i[y_i(\mathbf{w^T}\cdot\mathbf{X_i} - b) - 1 + \xi_i] \end{cases} \tag{3}$$

Where $\mathbf{w}$ is defined as the weight vector; ξ_i as the error term i.e. vectors found within the marginal boundary of hyperplane; and $C > 0$ as the regularization parameter [33].

4) LogitBoost: LogitBoost is a boosting algorithm that has been shown to classify text documents successfully [34]. Boosting is an ensemble approach that combines many weak classifiers to come up with a strong or good classifier that primarily reduces bias and variance. It sequentially fits multiple weak classifiers in a way such that more weight is given to observations in the dataset that were misclassified by the previous classifiers in the sequence. In essence, the training data is re-weighted to produce multiple classifiers in sequence [35]. Like AdaBoost, LogitBoost also performs additive logistic regression using maximum Bernoulli likelihood as a criterion.

C. Hybrid Approaches

Zhang et al. [36] propose an entity level hybrid approach for sentiment analysis for Twitter data. SL is used to perform sentiment analysis on a set of pre-defined entities. The sentiments assigned to these entities are used to identify new tweets having a similar sentiment. These tweets are then used as an automatically labelled training data to train an SVM, whereas we transform the term space using the SL into a two-dimensional feature space instead of generating the labelled data. Furthermore, [36] approach requires pre-defined entities but no manually labelled training data, in contrast, we do not require the former but need the later. Additionally, we are not performing an entity-level sentiment analysis.

Martin et al. [37] combine SA and ML approaches in an ensemble setting. Separate ML classifiers are learned for film reviews in Spanish and its corresponding English translation. A third model is then learned by using the SL (SentiWordNet) on the translated English corpus. The decision of the three classifiers is combined using stacking or voting to output the final label. [38] also propose a hybrid ensemble method to infer sentiment from documents using statistical methods and knowledge-driven linguistic patterns. Our approach, on the other hand, does not ensemble SL and ML classifiers; rather, it uses the SL approach to transform the term space.

In [39], Prabowo et al. propose a hybrid approach that uses multiple SL, rule-based, statistical, and ML approaches in a cascading fashion. Their approach starts by predicting sentiment from one of the algorithms, and if it fails to assign a sentiment, the text is then passed to the next algorithm, and so on until it is eventually assigned a sentiment by one of the algorithms. They experiment with ten different sequences of general inquirer based classifier [40], statistics based classifier, decision tree classifier (ID3) [28], RIPPER [41], and SVM. [42] also propose a hybrid approach that cascades four classifiers: (a) an emoticon classifier, (b) a slang language classifier, (c) an improved domain-specific classifier, and (d) the SentiWordNet classifier. The input document is classified in a two-stage process. In the first stage, it is classified through the first two classifiers; then in the second stage, it is classified through the last two classifiers to get a more accurate classification. [43] also propose a hybrid approach consisting of a sequence of the following five components: (a) sentiment rules, (b) semantic lexicon, (c) ambiguity management process, (d) negation handling process, and (e) linguistic variables. Our approach, to the contrary, is not a cascading classifier approach; we only learn a single discriminative classifier. That is because the SL is used by us to serve to transform the term space only.

Wiebe et al. [44] propose a hybrid approach for classifying sentences into subjective and objective categories. First, lexicons of subjective and objective clues are used to label the data as subjective or objective. The patterns for each category are then extracted from this automatically labelled data. These patterns then serve as a new training data for NB classifier that is used to label the whole of the unlabeled data even that which were left out during the initial labelling by the lexicons. Whereas, our approach focuses on classifying objective texts into positive and negative classes; secondly, we use SL for feature transformation rather than sampling a portion of the unlabeled data for further feature extraction.

The closest approach is given by Ghorbel et al. [45], which shows a hybrid approach for classifying movie reviews in the French language. They translate the French words into English and then find their polarity score using the SentiWordNet sentiment lexicon, after employing some word disambiguation. The polarity score is then used in conjunction with the French text, its POS tags, and some other features to build a feature vector to train an SVM model. Our approach differs with their approach in the sense that we do not use the sentiment scores as features to complement the textual attributes instead we transform the term space into a two-dimensional feature space using these sentiment scores.

In [46] use a three-step hybrid approach. First, they project the data onto an SL and then augment it with a word embedding. This transformed input space is then served as input to ML algorithms. They use the unsupervised Word2Vec embeddings developed by researchers at Google. It exploits the co-occurrence of words in a corpus to detect the meanings and semantic relations between words by training a Deep Neural Network. These approaches have recently gained popularity for sentiment analysis [47]–[49]. Our approach differs from these techniques as we do not learn any embeddings using deep learning and nor we project the input terms using these embeddings rather we transform the document term space through SL into a two-dimensional sentiment space.

Similarly, Mudinas et al. [50] propose a hybrid approach that uses an SL to generate a feature vector for an ML algorithm (SVM). They use SL and POS tagging to generate a feature vector containing sentiment words, adjectives, and lexicon-based sentiment scores, which are then used to train an SVM model. We, on the other hand, do not project our data on to the term space of the SL; instead, we transform the document term space using SL into a two-dimensional sentiment space. Furthermore, they perform concept-level sentiment analysis with the aim of learning optimal weights for these concepts, whereas we are doing document-level sentiment analysis.

The approach closest to what we propose in this paper is given by [51], where hybrid approach uses different sentiment and emoticon lexicons to transform the document term space into seven feature vector space consisting of the frequencies of positive words, very positive words, negative words, very negative words, booster words, negation words, positive emotions, and negative emoticons. Like our approach, this feature vector is then used to train an SVM model. Our approach is different from [51] in the sense that we employ lexicons to find the polarities of each word which are then ensembled into two scores, one for the positive category and one for the negative category. Our discriminative model is learned on this transformed two-dimensional sentiment space.

III. Methodology

Sentiment classification is the prediction of a discrete-valued sentiment. It determines the sentiment of a textual document, whether it be a tweet, Facebook post, product review, SMS, etc. Therefore our target is to predict the overall sentiment of the textual document as either positive or negative. Hence, the problem formulated as such becomes a binary sentiment classification problem in which the class (or category) label P refers to documents exhibiting a positive sentiment, whereas class label N refers to documents where the negative sentiment is predominant. The class labels are just symbolic labels that do not carry any semantics or any additional knowledge. For this study, class P is treated as the positive class.

The binary sentiment classification problem of text documents is formally defined as follows. Let $L = \{\langle \mathbf{x}_i, c_i \rangle\}_{i=1}^{\|L\|}$ be a labeled set of documents such that $c_i \in C = \{P, N\}$ represents the sentiment of the i^{th} document $\mathbf{x}_i$ and $\|L\|$ is the total number of documents in L; learn a classification model that assigns a class label c_i to each document in the unlabeled set $U = \{\langle \mathbf{x}_i \rangle\}_{i=1}^{\|U\|}$. It is assumed (although not guaranteed in practice) that the joint probability distribution of the text documents and the target variable C is identical in the labeled (L) and unlabeled sets (U). Therefore, our task is to approximate the unknown target function $\Phi^{'} : U \rightarrow C$ by the classifier function $\Phi : U \rightarrow C$ such that the number of $x_j \in U$ for which $\Phi(\mathbf{x}_j) \neq \Phi^{'}(\mathbf{x}_j)$ is as less as possible. For lexicon-based approaches, the classifier function $\Phi : U \rightarrow \{P, N\}$ is directly derived from lexicon dictionary and hence there is no learning of classifier function $\Phi : U$ from the labeled set L. We represent the text document x_j as a integer valued vector $\mathbf{x}_i = \langle x_{i1}, x_{i2}, \ldots, x_{i\|A\|} \rangle$ in a $\|A\|$ dimensional vector space such that x_{ij} indicates the value of the feature j for the text document i and $\|A\|$ is the number of unique features in the set $L \cup U$ after standard pre-processing has been applied.

Our proposed approach combines the lexicon and learning-based approaches into a single hybrid approach. We describe its schema and steps in the subsequent subsection. Data and its cleaning process are also discussed in this section.

A. Proposed Hybrid Approach

The proposed approach consists of two major parts, the unsupervised feature transformation, and the supervised discriminant classification. The unsupervised part relies on SL to transform the $\|A\|$ dimensional term space into a two-dimensional sentiment vector space where each dimension corresponds to one of the sentiment class in the set C. Each dimension represents the opinion of the terms based on their respective sentiment polarity score in one of the lexicons. The terms with positive polarity contribute towards the membership of the document for positive sentiment class P, whereas the negative polarity terms contribute towards the membership of the document for the negative sentiment class N. The aggregated polarity score of all these terms is obtained as a linear sum of individual polarity scores normalized by the total number of words in the document. The resultant two scores S^{P} and S^{N} are thus the normalized sentiment scores for the positive and negative classes for document X, respectively. A term contributes towards the score only if it occurs in that document. If the same term occurs more than once in the document, then it contributes to the score S^{P} (or S^{N}) each time.

The sentiment scores, S^{P} and S^{N} define the two-dimensional sentiment space in which documents are aligned along the dimension that corresponds to the sentiment prevalent in them. In this space, documents belonging to one sentiment class are easily discriminable from the documents belonging to the other sentiment class, as illustrated for a dataset in Fig. 2. In this figure, the documents having positive sentiment as the true label (coloured in purple) nicely

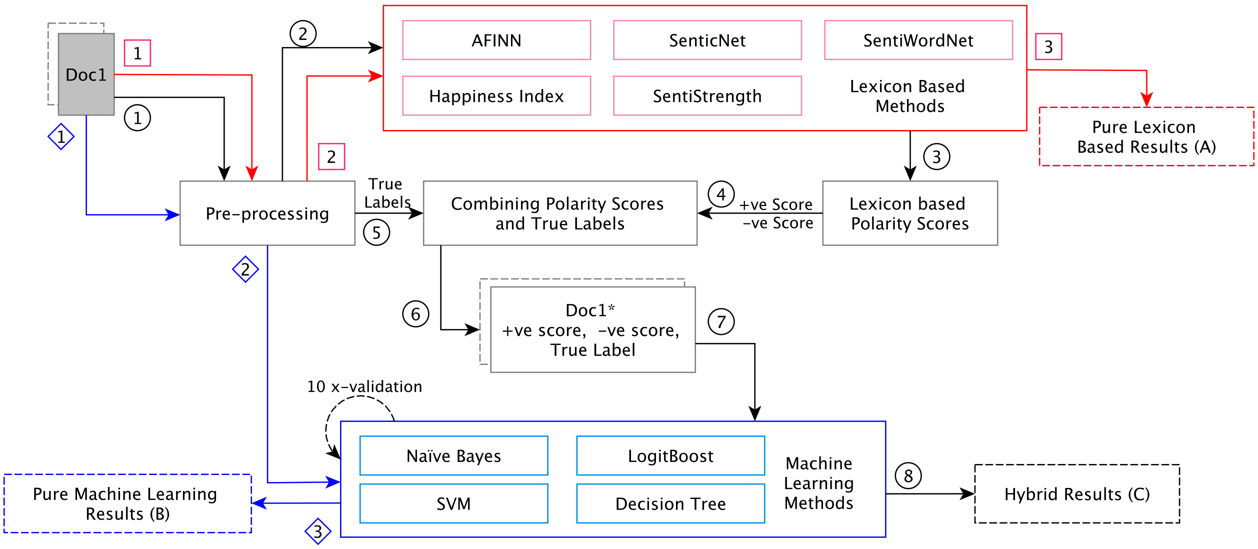


Fig. 1. Block Diagram of the Overall Process. The Arrows in Red Color (Tagged with in Squares) Correspond to the Sentiment Lexicon Approaches, whereas Blue Colored Arrows (Tagged with in Diamonds) Correspond to Pure Machine Learning Approaches, and Arrows in Black Color (Tagged with in Circles) Correspond to the Proposed Hybrid Approach.

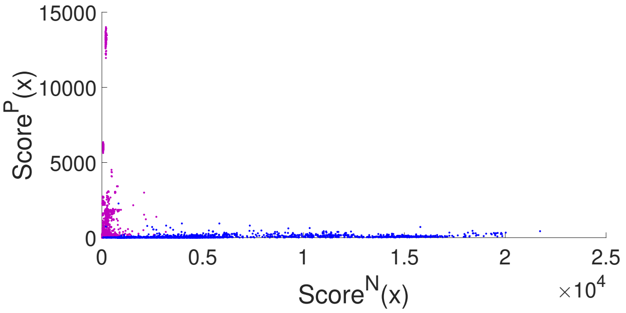


Fig. 2. The 2D Feature Space after Applying SL Approach. Each Point in this Sentiment Space Corresponds to a Document in the Dataset. Purple and blue points are documents with P and N as true class label, respectively.

align along the y-axis whereas the documents having negative sentiment as the true label (coloured in blue) align along the x-axis. Thus these sentiment scores can be thought of as the confidence values of a document's membership in the positive and negative class.

There are multiple hypotheses possible in this two-dimensional space. A potential hypothesis could be to assign the highest sentiment score class to the document. This is also known as the max rule and corresponds to a straight line at a 45 degrees angle from the origin. Though quite handy and intuitive, these rules fail to achieve good generalizations when there exists a class imbalance in the data set or the distribution of the training and test is significantly different. A better hypothesis may be a maximum margin hypothesis such as learned by SVM, or the hypotheses of some other discriminative classifier. Therefore, we train different supervised discriminative classifiers in this sentiment space to find the best hypotheses that separate the two sentiment classes. Fig. 1 depicts the overall schema of our methodology.

```
Data: Let X be the text to process; D be the
      SenticNet dictionary
Result: A tuple of information ⟨coverage, positive
        score, and negative score⟩
c ← be a variable for coverage intially set to False
$S^{\mathrm{P}}$ ← be a variable for positive score initially set to 0
$S^{\mathrm{N}}$ ← be a variable for negative score intially set to 0
Remove all punctuation marks in X
Remove all stopwords in X
if $\|\mathbf{X}\| = 0$ then
    return ⟨c, $S^{\mathrm{P}}$, $S^{\mathrm{N}}$⟩
end
for each token $w$ in X do
    if $w \in \mathbf{D}$ then
        c ← True
        if $\mathbf{D}[w] > 0.0$ then
            $S^{\mathrm{P}} \leftarrow S^{\mathrm{P}} + \mathbf{D}[w]$
        else
            $S^{\mathrm{N}} \leftarrow S^{\mathrm{N}} + \mathbf{D}[w]$
        end
    else
        continue
    end
end
return ⟨c, $S^{\mathrm{P}}$, $S^{\mathrm{N}}$⟩
```

Algorithm 1: A Sample of Altered Algorithm (SenticNet) for Proposed Hybrid Approach.

To generate the aforementioned two-dimensional sentiment space, we alter the decision rule of the SL approaches. Instead of providing a decision about whether the document is positive or negative, we make them output the positive or negative scores only. Algorithm 1 depicts the altered algorithm for SenticNet that outputs the S^{P} and S^{N} scores instead of outputting the label. The true labels of the document are appended to these two scores to obtain the training data for the supervised approach. Thus the dimensionality of our problem is reduced

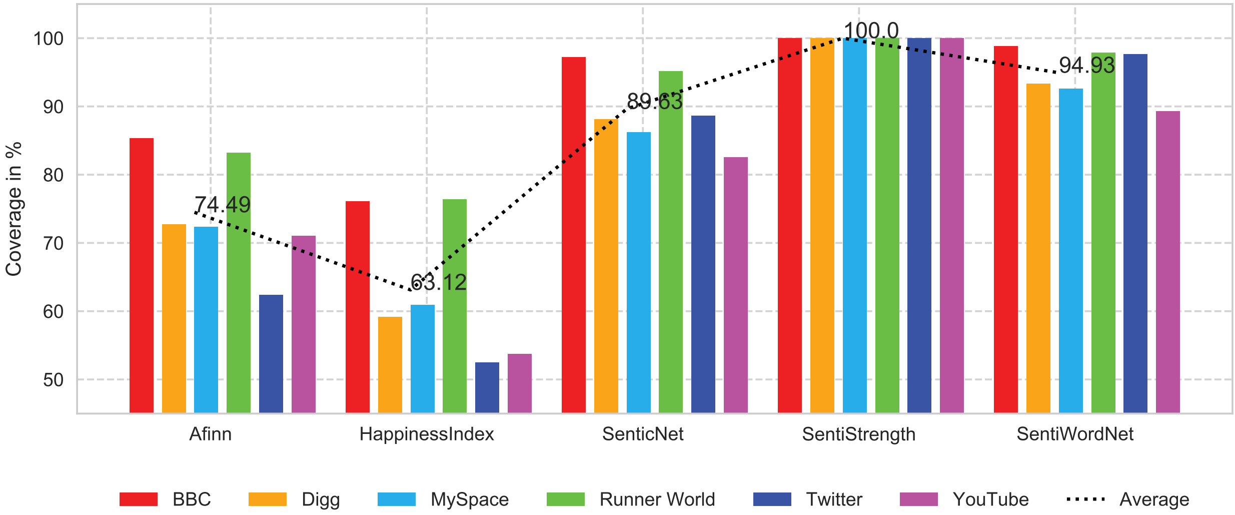


Fig. 3. Coverage of Lexicon based Methods. Each Number Indicates the Percentage of Records Covered in Dataset.

from the number of unique terms ($\|A\|$) in the dataset to just a two-dimensional sentiment space. The resultant feature space is more comfortable to visualize and gives insights on the separability of the sentiments in the transformed feature space. It is also faster for the ML algorithms to build their model because of the small feature space. It is to be noted that our hybrid approach is not specific to any SL or ML algorithm. Any SL can be used to generate the two-dimensional sentiment space over which any ML algorithm can be applied to learn the final decision function.

TABLE I. STATISTICS OF DATASETS: NUMBER OF RECORDS (DOCUMENTS) IN IT AND PERCENTAGE OF CLASS DISTRIBUTION.

Dataset	# Records	% of +ve / -ve Records
BBC	1,000	34.70% / 65.30%
Digg	1,077	46.89% / 53.11%
MySpace	1,041	87.32% / 12.68%
Runner World Forum	1,046	78.87% / 21.13%
Twitter	4,242	77.63% / 22.37%
YouTube	3,407	77.49% / 22.51%

B. Pre-Processing and True Label Generation

We chose the datasets used by [20] to evaluate the performance of the classifiers. The dataset belongs to six online domains, namely, BBC, Digg, MySpace, Runner World, Twitter, and YouTube. Details of these datasets are given in Table I. As given in Table I, the datasets have different sizes and have a very different distribution of positive and negative documents. Each record in these datasets is labeled (scored) by humans using Mechanical Turk [52] based on positive and negative sentiment present in them. Thus, each record has it is a positive score and as well as a negative score. The following rule is used to assign a class label to each document X:

$$T(X_i) = \begin{cases} Positive & \text{if +ve score}_{X_i} \geq \text{+ve score}_{X_i} \\ Negative & \text{otherwise} \end{cases} \tag{4}$$

Where i is the document index $1 \geq i \leq n$.

The following preprocessing steps are applied to the data before running any of the classifiers.

1) Punctuation marks are removed by replacing them by empty string using the following regular expression (R) `[\@\#\$\%\^\&\*\_\.?\!\:\,\;\+\=\-\|\<\>\{\}\(\)\[\]\"\/]`.
2) Case folding is performed to transform whole document to lower case.
3) Documents are transformed into a term-incidence matrix (for ML approaches only).

In addition to the above two steps, SL methods also have their data cleaning steps. E.g., SenticNet replaces characters `.!?,` with a single whitespace, followed by tokenization of text to be processed with simple white space (line 2, Algorithm 1).

IV. RESULTS AND EVALUATIONS

Before we can compare SL approaches with ML and the proposed hybrid (or combined) approach, it is important to note that SL approaches do not assign a label to every example in a dataset, a problem known as the coverage problem. It occurs when no word of the example to be classified is present in the SL. For such examples, the SL cannot give any decision regarding their sentiment. The coverage of the five sentiment lexicons used in this study is given in Fig. 3. SentiWordNet with more than 117,000 words and phrases in its dictionary has the second-highest coverage ≈ 95%, while Happiness Index with 1,032 words only, attained the lowest coverage of 63.12%. However, SentiStrength demonstrates full coverage, which is not a surprise as it was created from the five datasets that are used in this study. Therefore, in order to compare the performance of the SL approaches to ML, and the proposed hybrid approach, it is necessary to have them tested on the same set of examples. Therefore, if a particular SL covers only 600 examples out of the 1000 examples, then the ML and the proposed approach classifiers are trained over these 600 examples only using a ten-fold cross-validation approach.

TABLE II. Performance comparison of all the approaches dataset wise. All values are in percentages.

Approach	Dataset	Precision	Recall	Specificity	F-Measure	Accuracy
Sentiment Lexicons	BBC	0.3903	0.7547	0.2271	0.5039	0.5055
	Digg	0.4941	0.7634	0.3239	0.5904	0.5360
	MySpace	0.8987	0.8735	0.6733	0.8848	0.8048
	Runner World	0.8132	0.8534	0.6335	0.8304	0.7296
	Twitter	0.8035	0.8307	0.5830	0.8147	0.7155
	YouTube	0.8011	0.8480	0.5740	0.8273	0.7300
Avg. Sentiment Lexicon		0.7018	0.8206	0.5025	0.7419	0.6702
Pure Machine Learning	BBC	0.4411	0.1837	0.8694	0.7038	0.6305
	Digg	0.5802	0.4450	0.7172	0.6603	0.5923
	MySpace	0.8881	0.9733	0.1231	0.7712	0.8638
	Runner World	0.8043	0.9400	0.1300	0.7479	0.7710
	Twitter	0.8133	0.9329	0.1903	0.7566	0.7692
	YouTube	0.8242	0.9161	0.2415	0.7504	0.7635
Avg. Pure Machine Learning		0.7252	0.7318	0.3785	0.7317	0.7317
Proposed Hybrid Approach	BBC	0.3838	0.1553	0.9337	0.2000	0.6790
	Digg	0.4654	0.2655	0.8820	0.2993	0.6123
	MySpace	0.8839	0.9852	0.0948	0.9313	0.8731
	Runner World	0.8000	0.9797	0.0638	0.8799	0.7886
	Twitter	0.7882	0.9511	0.1686	0.8599	0.7671
	YouTube	0.7911	0.9652	0.1434	0.8678	0.7746
Avg. Proposed Hybrid Approach		0.6854	0.7170	0.3810	0.6730	0.7491

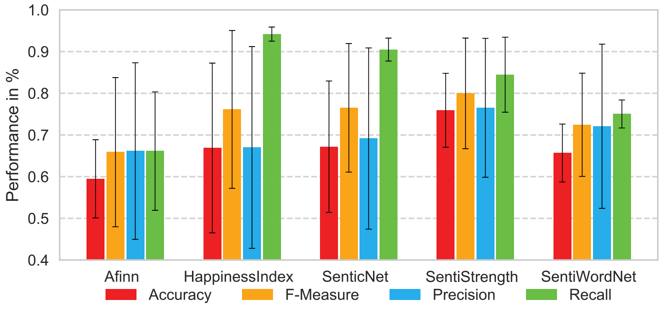


Fig. 4. Average Performance and Standard Deviation of Lexicon-based Approaches.

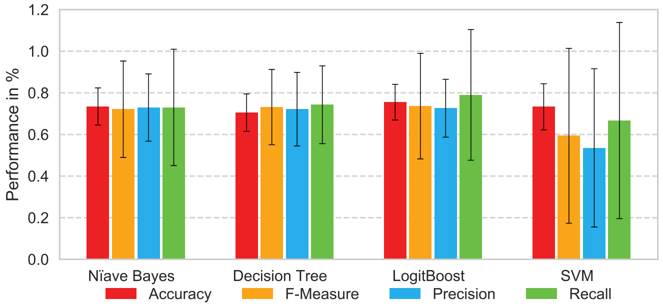


Fig. 5. Average Performance and Standard Deviation of (Pure) Machine Learning Approaches.

Therefore all the results for ML and the hybrid approaches are based on only those comments that are covered by that particular lexicon. The detailed comparative results for the SL, ML, and proposed hybrid classifiers is presented in Fig. 4, 5, and 6, respectively. It is to be noted at each value bars in these figures correspond to the values of the corresponding approach averaged over the six datasets mentioned in III-B.

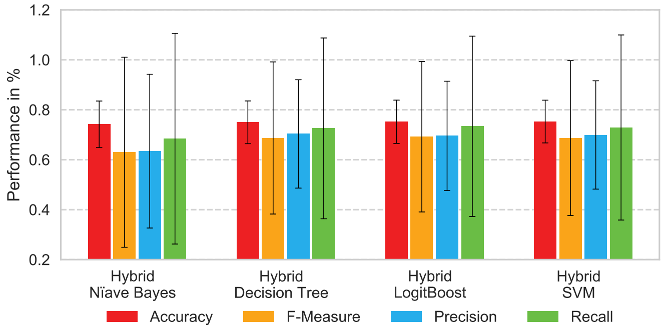


Fig. 6. Average Performance and Standard Deviation of the Proposed Hybrid Approach.

Not surprisingly SentiStrength outperforms the rest of the SL approaches (Fig. 4) because its dictionary was built upon the datasets used in this research work. Therefore, SenticNet seems to be the winner (after excluding SentiStrength) among SL approaches with the highest accuracy and F-measure, but it has a high standard deviation. Happiness Index is not far behind SentiStrength and has a higher recall but at the expense of lower precision. This can be attributed to a bias towards the positive class; furthermore, it has even more variance than SentiStrength. Overall the performance of the SL is not encouraging as the accuracy is below than 70% even when only those comments are used which are covered by these lexicons.

As expected, ML-based approaches perform better than SL approaches because they have a training phase (Fig. 5). All the four classifiers achieve an accuracy of more than 70% with LogitBoost outperforming the rest with more than 75% accuracy. This is a bit surprising as NB and SVM are thought as better classifiers for text classifiers. It is to be noted that each value bar in this figure corresponds to the average performance of a classifier over all the datasets using only the examples that are covered by the respective SL. E.g., the NB's precision of

71.01% is calculated from the examples of the six datasets that were covered by Afinn, and from the examples of six datasets that were covered by SenticNet, and so. Therefore each value in this table is an average of the performance of thirty different NB classifiers. Surprisingly LogitBoost outperforms NB and SVM.

As hypothesized, the proposed hybrid classifiers outperform the SL and ML classifiers (Fig. 6). The accuracy of all the four SL approaches increases when used with the proposed approach with decision tree benefiting the most with an average accuracy improvement of more than 4%. The comparison of all the approaches on individual datasets is presented in Table II. Performance gain over the SL approach was expected as SL is an unsupervised approach, whereas the ML and the proposed approach is a supervised one. The dataset to benefit the most from the proposed approach is the BBC dataset with an increase of 17.25% and 4.85% in average accuracy over the SL and ML approaches, respectively. For the Twitter data only does the ML approach beat the proposed hybrid approach but by only a margin of 0.21%. Overall the proposed approach outperforms the SL and ML approaches by a margin of 7.88% and 1.70%, respectively. Average accuracy improvement of 1.70% in the accuracy of the hybrid approach is significant, considering that the average accuracy of the ML approaches is about 73% only. The detailed performance gains are reported in Tables III and IV, respectively. SVM seems to be the major benefactor of the proposed approach.

TABLE III. Average improvement in percentage accuracy by the proposed hybrid approach over SL approaches. HI, SN, SS, SWN, and LB refer to the Happiness Index, SenticNet, SentiStrength, SentiWordNet, and LogitBoost, respectively.

	H-NB	H-DT	H-LB	H-SVM	Average
Afinn	14.68%	15.50%	15.71%	15.81%	15.42%
HI	7.28%	8.10%	8.31%	8.41%	8.02%
SN	6.98%	7.80%	8.01%	8.11%	7.72%
SS	-1.75%	-0.93%	-0.72%	-0.62%	-1.01%
SWN	8.49%	9.31%	9.52%	9.62%	9.23%
Average	7.13%	7.95%	8.17%	8.26%	7.88%

TABLE IV. Average improvement in percentage accuracy by the proposed hybrid approach over their ML counterparts.

	H-NB	H-DT	H-LB	H-SVM	Average
NB	0.71%	1.53%	1.75%	1.84%	1.46%
DT	3.64%	4.46%	4.68%	4.77%	4.38%
LB	-1.39%	-0.57%	-0.35%	-0.26%	-0.64%
SVM	0.87%	1.69%	1.90%	2.00%	1.61%
Average	0.96%	1.77%	1.99%	2.09%	1.70%

A. Scalability and Complexity Analysis

In terms of time and space requirements, the proposed approach is highly efficient than directly learning a supervised ML classifier. The two-dimensional sentiment score space is generated in a single pass over the labelled data. The time required to generate this space is $O(\|L\| \cdot a)$, where $\|L\|$ is the total number of labelled documents in the training data, $\|A\|$ as defined earlier is the number of unique terms in the dataset, and a is the average number of terms in a document. Since a document vector is sparse in the $\|A\|$ dimensional space, therefore, $a \ll \|A\|$, thus making the $O(\|L\| \cdot a)$ asymptotic running time linear in terms of the size of the dataset. The next major step on which the running time is dependent is the training of the supervised ML model is this two-dimensional sentiment score space. Thus the total time to generate the proposed model is $O((\|L\| \cdot a) + MTT)$, where MTT is the time taken to train the ML model. Depending on which ML model is used, MTT can also be linear in terms of the size of the dataset. Therefore the proposed algorithm is trained in linear time, and it is the fastest (asymptotically) running time for a binary class classification algorithm.

The proposed approach is more scalable than an approach that learns the ML algorithm directly. It is because the first step in generating the two-dimensional sentiment space requires hash table lookups to retrieve the score of a word and sum them up. A hash table lookup requires $O(1)$ time as the size of the SL is already known and is static. Furthermore, the size of the hash table of SL is much lesser than $\|A\|$. In the second step, the ML model is learned in a space with two attributes only which are very efficient as compared to when the ML model is learned directly in a $\|A\|$ dimensional term space because $\|A\|$ can easily reach hundreds of thousands of words.

Classification model of the proposed approach requires lesser space than the ML approaches by order of magnitude. E.g., NB calculates probabilities of each of the $\|A\|$ words for each class C, thus making it space complexity as $O(\|A\| \cdot \|C\|)$. Whereas for the proposed approach, in addition to the SL whose size is significantly lesser and as well as independent of the size of the data, only $O(\|C\|)$ probabilities are stored because of $\|A\|$ is equal to two in the two-dimensional sentiment score space.

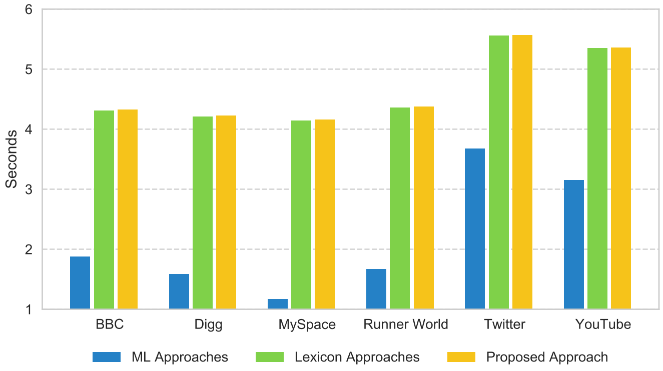


Fig. 7. Average Running Time of Algorithms on the Datasets.

Being a lazy learning approach, the SL approaches have the advantage of not having a training phase. However, the prediction of labels for the unseen data can be a bit slow as a dictionary lookup is required for each word in the document. Fig. 7 plots the average running times of the SL, ML, and the proposed approach to predict the labels for the various datasets. Since our approach uses the SL approach at the first step, its prediction is therefore deemed to be slower than the SL approach. However, the figure suggests that the overhead is very low as there is almost an overlap between the running time curves of the SL and the proposed approach.

Document representation as described in Section III corresponds to a sparse vector in the $\|A\|$ dimensional vector space. Using this representation, the whole data is represented as a document incidence matrix having a size of $\|L\| \cdot \|A\|$ dimensions. Since $\|A\|$ is a large number for text classification problems, therefore building such a big matrix requires much memory and computational cost. Approaches like NB do not require a document incidence matrix for computing its probabilities; instead, NB is efficiently implemented using a hash table data structure that would only require $O(\|L\| \cdot a)$ space. Since a is significantly less than $\|A\|$, it is a big improvement and makes NB one of the fastest classifiers. Like NB, the proposed approach is implemented using the hash table data structure, thus having a low memory footprint. Since we do not need to access terms and their scores in any specific order, therefore we retrieve, store and update the scores of each term in constant time using a hash table. This makes our approach very fast and memory efficient.

V. Conclusion and Future Work

Sentiment mining of textual content on social media can give insights for targeted advertisement, product reviews, and much more. In this article, we have proposed a hybrid sentiment analysis technique that uses sentiment lexicons (SL) to transform the input term space into a sentiment score space of only two dimensions where a supervised machine learning (ML) algorithm is learned to output the final decisions. The proposed approach demonstrates significant performance gain over the original SL and ML approaches when evaluated for three ML algorithms and five SL over six social media datasets. It also takes less time and memory to train than the ML approaches. Thus our approach is suitable for scenarios where training data is scarce, and more balanced classification is required. In the future, we plan to identify the terms in the SL that result in the classification bias & devise a mechanism to penalize them for reducing the bias of the proposed hybrid approach further.

References

[1] A. Jurek, M. D. Mulvenna, and Y. Bi, "Improved lexicon-based sentiment analysis for social media analytics," *Security Informatics*, vol. 4, no. 1, p. 9, 2015.

[2] J.-Y. Chen, H.-T. Zheng, Y. Jiang, S.-T. Xia, and C.-Z. Zhao, "A probabilistic model for semantic advertising," *Knowledge and Information Systems*, pp. 1–26, 2018.

[3] M. M. Mostafa, "More than words: Social networks' text mining for consumer brand sentiments," *Expert Systems with Applications*, vol. 40, no. 10, pp. 4241–4251, 2013.

[4] D. Liu and L. Lei, "The appeal to political sentiment: An analysis of donald trump's and hillary clinton's speech themes and discourse strategies in the 2016 us presidential election," *Discourse, Context & Media*, 2018.

[5] W. Wang, H. Wang, and Y. Song, "Ranking product aspects through sentiment analysis of online reviews," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 29, no. 2, pp. 227–246, 2017.

[6] W. Muhammad, M. Mushtaq, K. N. Junejo, and M. Y. Khan, "Sentiment analysis of product reviews in the absence of labelled data using supervised learning approaches," *Malaysian Journal of Computer Science*, vol. 33, no. 2, pp. 118–132, 2020.

[7] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002, pp. 79–86.

[8] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of LREC*, vol. 6. Citeseer, 2006, pp. 417–422.

[9] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Recognition of affect, judgment, and appreciation in text," in *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 806–814.

[10] S. Mohammad, "From once upon a time to happily ever after: Tracking emotions in novels and fairy tales," in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Association for Computational Linguistics, 2011, pp. 105–114.

[11] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, "Emotion detection from text & speech: a survey," *Social Network Analysis & Mining*, vol. 8, no. 1, p. 28, 2018.

[12] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in *Proceedings of the first ACM conference on Online social networks*. ACM, 2013, pp. 27–38.

[13] Z. Hailong, G. Wenyan, and J. Bo, "Machine learning and lexicon based methods for sentiment classification: A survey," in *Web Information System and Application Conference (WISA), 2014 11th*. IEEE, 2014, pp. 262–265.

[14] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[15] J. Eisenstein, "Unsupervised learning for lexicon-based classification." in *AAAI*, 2017, pp. 3188–3194.

[16] M. Araújo, P. Gonçalves, M. Cha, and F. Benevenuto, "ifeel: A system that compares and combines sentiment analysis methods," in *23rd international conference on World wide web companion*. International World Wide Web Conferences Steering Committee, 2014, pp. 75–78.

[17] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," *arXiv preprint arXiv:1103.2903*, 2011.

[18] P. S. Dodds and C. M. Danforth, "Measuring the happiness of large-scale written expression: Songs, blogs, and presidents," *Journal of Happiness Studies*, vol. 11, no. 4, pp. 441–456, 2010.

[19] M. M. Bradley and P. J. Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, Tech. Rep., 1999.

[20] M. Thelwall, "Heart and soul: Sentiment strength detection in the social web with sentistrength," *Proceedings of the CyberEmotions*, pp. 1–14, 2013.

[21] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[22] E. Cambria, C. Havasi, and A. Hussain, "Senticnet 2: A semantic & affective resource for opinion mining & sentiment analysis." in *FLAIRS*, 2012, pp. 202–207.

[23] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," *arXiv preprint arXiv:1103.2903*, 2011.

[24] C. Biever, "Twitter mood maps reveal emotional states of america," *New Scientist*, vol. 207, no. 2771, p. 14, 2010.

[25] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[26] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[27] J. Song, K. T. Kim, B. Lee, S. Kim, and H. Y. Youn, "A novel classification approach based on naïve bayes for twitter sentiment analysis," *KSII Transactions on Internet and Information Systems*, vol. 11, no. 6, pp. 2996–3011, 2017.

[28] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[29] A. Ortigosa, J. M. Martín, and R. M. Carro, "Sentiment analysis in facebook and its application to e-learning," *Computers in Human Behavior*, vol. 31, pp. 527–541, 2014.

[30] R. A. Berk, "Classification and regression trees (cart)," in *Statistical learning from a regression perspective*. Springer, 2016, pp. 129–186.

[31] A. S. Manek, P. D. Shenoy, M. C. Mohan, and K. Venugopal, "Aspect term extraction for sentiment analysis in large movie reviews using gini index feature selection method and svm classifier," *World wide web*, vol. 20, no. 2, pp. 135–154, 2017.

[32] Y. Liu, J.-W. Bi, and Z.-P. 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*, vol. 394, pp. 38–52, 2017.

[33] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[34] S. Kotsiantis, E. Athanasopoulou, and P. Pintelas, "Logitboost of multinomial bayesian classifier for text classification," *International Review on Computers and Software (IRECOS)*, vol. 1, no. 3, pp. 243–500, 2006.

[35] J. Friedman, T. Hastie, R. Tibshirani *et al.*, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[36] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining lexicon based and learning-based methods for twitter sentiment analysis," *HP Laboratories, Technical Report HPL-2011*, vol. 89, 2011.

[37] M.-T. Martín-Valdivia, E. Martínez-Cámara, J.-M. Perea-Ortega, and L. A. Ureña-López, "Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches," *Expert Systems with Applications*, vol. 40, no. 10, pp. 3934–3942, 2013.

[38] E. Cambria and A. Hussain, *Sentic computing: a common-sense-based framework for concept-level sentiment analysis*. Springer, 2015, vol. 1.

[39] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," *Jrn. of Informetrics*, vol. 3, no. 2, pp. 143–157, 2009.

[40] P. J. Stone, D. C. Dunphy, and M. S. Smith, "The general inquirer: A computer approach to content analysis." 1966.

[41] W. W. Cohen, "Fast effective rule induction," in *Proceedings of the twelfth international conference on machine learning*, 1995, pp. 115–123.

[42] M. Z. Asghar, F. M. Kundi, S. Ahmad, A. Khan, and F. Khan, "T-saf: Twitter sentiment analysis framework using a hybrid classification scheme," *Expert Systems*, vol. 35, no. 1, p. e12233, 2018.

[43] O. Appel, F. Chiclana, J. Carter, and H. Fujita, "Successes and challenges in developing a hybrid approach to sentiment analysis," *Applied Intelligence*, vol. 48, no. 5, pp. 1176–1188, 2018.

[44] J. Wiebe and E. Riloff, "Creating subjective and objective sentence classifiers from unannotated texts," in *Computational Linguistics and Intelligent Text Processing*. Springer, 2005, pp. 486–497.

[45] H. Ghorbel and D. Jacot, "Sentiment analysis of french movie reviews," in *Advances in Distributed Agent-Based Retrieval Tools*. Springer, 2011, pp. 97–108.

[46] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, "Sentiment analysis leveraging emotions and word embeddings," *Expert Systems with Applications*, vol. 69, pp. 214–224, 2017.

[47] S. Xiong, H. Lv, W. Zhao, and D. Ji, "Towards twitter sentiment classification by multi-level sentiment-enriched word embeddings," *Neurocomputing*, vol. 275, pp. 2459–2466, 2018.

[48] E. Cambria, S. Poria, D. Hazarika, and K. Kwok, "Senticnet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings," in *AAAI*, 2018.

[49] X. Fu, X. Sun, H. Wu, L. Cui, and J. Z. Huang, "Weakly supervised topic sentiment joint model with word embeddings," *Knowledge-Based Systems*, vol. 147, pp. 43–54, 2018.

[50] A. Mudinas, D. Zhang, and M. Levene, "Combining lexicon and learning based approaches for concept-level sentiment analysis," in *Proceedings of the First Int. Workshop on Issues of Sentiment Discovery and Opinion Mining*. ACM, 2012, p. 5.

[51] D. Mumtaz and B. Ahuja, "A lexical and machine learning-based hybrid system for sentiment analysis," in *Innovations in Computational Intelligence*. Springer, 2018, pp. 165–175.

[52] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk a new source of inexpensive, yet high-quality, data?" *Perspectives on psychological science*, vol. 6, no. 1, pp. 3–5, 2011.

Capsule Network for Cyberthreat Detection

Sahar Altalhi[1], Maysoon Abulkhair[2], Entisar Alkayal[3]
Information Technology Department
King Abdulaziz University

***Abstract*—In cybersecurity, analyzing social network data has become an essential research area due to its property of providing real-time updates about real-world events. Studies have shown that Twitter can contain information about security threats before some specialized sites. Thus, the classification of tweets into security-related and not security-related can help with early warnings for such attacks. In this study, the use of a capsule network (CapsNet), the new deep learning algorithm, is investigated for the first time in the field of security attack detection using Twitter. The aim was to increase the accuracy of tweet classification by using CapsNet rather than a convolutional neural network (CNN). To achieve the research objective, the original implementation of CapsNet with dynamic routing is adapted to be suitable for text analysis using tweet data set. A random search technique was used to tune the model's hyperparameters. The experimental results showed that CapsNet exceeded the baseline CNN on the same data set, with accuracy of 92.21% and a 92.2% F1 score; also, word2vec embedding performed better than a random initialization.**

***Keywords*—Capsule network; dynamic routing; deep learning; Twitter; text analysis; attack detection**

I. Introduction

Security monitoring and attack detection are essential parts of any organization's management for protection against cyber-attacks. These attacks can cause service disruption, asset damage, data breaches, or data loss. To avoid such dangerous effects, a number of official security data sources are available, including the National Vulnerability Database (NVD) [1], which contains a security analysis of discovered vulnerabilities, and the ExploitDB [2], which provides a user-friendly interface for all discovered exploits targeting known vulnerabilities. These traditional data sources provide trusted security information, but it comes at a cost, which is the delay of reporting the information [3]. Not all reported vulnerabilities will be exploited in the real world, and some have a higher probability of being exploited and thus need to be patched first [4].

For system administrators, the time between the detection of a cyberattack plan and the actual occurrence is critical. They need up-to-date information about current or imminent attacks to analyze them, study their impact, and be aware of new attack types and hacking tools in real time [5]. One of the new solutions for this problem is utilizing social network data to extract real-time notifications about the security situation of the organization or software and hardware used in its infrastructure.

As one of the most popular social networks, Twitter is considered a rich source of information about different security threats. This claim is supported by studies showing that Twitter contained information about security threats before some specialized sites [6]–[8]. This observation attracted researchers to analyze Twitter data and extract knowledge to be used in the detection and prediction of security attacks. The objectives when using Twitter data in the security field vary, from vulnerability and exploit detection [4], [9], [10] to attack detection by linking a sentiment score to a specific target with real security events [11]–[13], and trying to determine the threshold of tweet sentiment that predicts the probability of the attack occurring [14].

Text classification using different machine learning (ML), neural network (NN), and deep learning (DL) algorithms has been widely investigated for detecting cyber-attacks using Twitter data. One of the most advanced techniques for this purpose is the convolutional neural network (CNN) [15]. As one of the DL algorithms, a CNN overcomes the traditional ML technique limitation by providing automation for the learning process [16]. However, the CNN comes with limitations that are mainly related to the use of the pooling layer [17], which will be described in detail in Section II.

In 2017, the godfather of DL, Geoffrey Hinton, proposed the capsule network (CapsNet), which was first examined using the Modified National Institute of Standards and Technology's (MNIST) data set [18]. CapsNet outperforms its predecessor, the CNN, in many image classification tasks [19], but it is still in early stages for text classification [20]. This study aimed to use Twitter to examine CapsNet's capability for providing accurate classifications of security tweets with the goal of cyberthreat detection. The CapsNet is implemented by building an NN model to classify tweets as security-related or not security-related. Then, the CapsNet model was evaluated in terms of classification accuracy and F1 score, and using the CNN as a baseline model, compared the performance of CapsNet in tweet classification for the security field.

The rest of this paper is organized as follows: In Section II, an overview of CapsNet's improvements in comparison to CNNs will be given. Section III covers the main recent work done in the field using Twitter for cyberthreat detection. This is followed by Section IV, which describes the implementation of the proposed model. In Section V, the details of the experiments conducted is described, and Section VI includes the results and discussion. Finally, the paper's conclusion and future works are discussed in Section VII.

II. Background

Until recently, CNNs achieved state-of-the-art results for many natural language processing (NLP) tasks [21]. However, CNNs have limitations and drawbacks, such as with pooling. Pooling, one of the building blocks of CNNs, is used to reduce the complexity and the number of parameters in the CNN while preserving the main features [20]. This makes CNNs

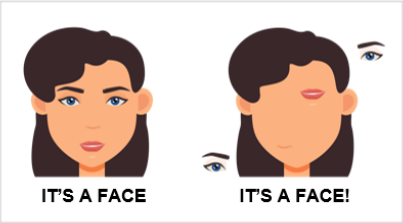


Fig. 1. Face recognition in CNN.

particularly efficient with classification tasks, but it causes a loss of valuable information such as the precise location of an object or the relationships between the object's parts [18]. Fig. **??** shows the way that the CNN works. Even when parts of the face are not arranged correctly, the CNN will classify it as a face regardless of the location and relationships between the parts [19]. For better modeling of spatial relationships among parts, CapsNet is proposed [22].

The architecture of the CapsNet overcomes CNN's drawbacks by different properties [18]. First, the basic unit in the CapsNet is the capsule (vector), where each one is a set of neurons representing an object or an object part. CapsNet transforms vector inputs into vector outputs; thus, it can learn more complex transformations than CNNs, which operate on scalars. The output of a capsule is an activity vector, where its length represents the probability of the existence of the object, and its coordinates (dimensions) encode the object's attributes (pose information), which preserves the spatial relationship between features. Second, CapsNet uses a routing-by-agreement technique to replace the routing by max pooling used in the CNN. In simple terms, instead of extracting the most important features by using max pooling and ignoring the less important ones, propagation between the layers will be based on routing-by-agreement. This means that the output of each capsule will be forwarded to the next layers' capsules with different weights that are based on the agreements between the capsules.

In NLP, CapsNet has a greater ability to efficiently learn the spatial relationships between words, such as the local order of words and their semantic representations [22]. Many researchers have investigated the use of CapsNet for NLP tasks like sentiment analysis [23], [24], fake news detection [25], stock performance prediction using Twitter [26], implicit emotion detection [27], and offensive posts on social media [28]. The results of these studies showed that CapsNet outperformed the CNN in text classification, which was part of the motivation for the present study.

III. Related Work

In this section, some studies that used Twitter data for the detection of security attacks are reviewed. For each work, the specific problem that was solved by each of these studies is summarized, the analysis techniques used, and the results obtained to give an overview of the research already conducted in the field of interest.

In order to discover Twitter discussions about emerging attacks against a specific target, the authors of [29] proposed an approach to security event detection that learned with positive and unlabeled data based on user-provided expectations. Expectation regularization (ER) was used to find the ratio between positive and negative examples in the training process. The study's security events included denial of service (DoS) attacks, data breaches, and account hijacking represented in the form of (Entity, Date) as training examples. Two sets of manually extracted features were used to find new events. Using the logistic regression (LR) classifier, the proposed solution was able to detect new events automatically in real time for each predefined category.

The use of simple discrete features may suffer limitations in representing subtle semantic differences between true event mentions and false cases with similar word patterns. To overcome this limitation, the researchers in [16], based on [29], modified the method to be more semantically based by using a long short-term memory (LSTM) based neural embedding model that learns tweet-level features automatically. This change improved the detection accuracy as compared to the previous method because of the NN's ability to represent deep semantic information, which is more difficult to capture through discrete features.

As an end-to-end solution for cyberthreat detection, SYNAPSE [30] provided a real-time extraction of security events from Twitter with high-level abstraction. A data set of more than 195,000 tweets was collected from security-related accounts and filtered by keywords related to the monitored infrastructure. The statistical method called term frequency-inverse document frequency (TF-IDF) was used to extract the tweets' features. Support vector machine (SVM) algorithm was used for feature learning, and it achieved a minimum true positive rate (TPR) and a true negative rate (TNR) of 90% in classifying tweets. For more informative extractions, the model proposed in [30] included stream tweet clustering using a dynamic clustering algorithm and summarization of each cluster with the exemplar tweet. The model was able to detect important actionable threats by verifying them with threats reported in the Common Vulnerability Scoring System (CVSS).

With the same objectives as the previous work [30], the authors in [31] proposed an event detection model with joint phases that performed the filtering, clustering, and summarization with shared tweet representation. Features were extracted using skip-gram and LSTM to obtain vector representations, and a multi-layer perceptron (MLP) classifier was used for tweet classification identifying security-related tweets. The tweets were clustered in groups, and each cluster was summarized with the most informative tweet provided. All these phases were conducted jointly based on features extracted at the beginning. The collaborative event detection and summarization model was more effective than solutions that used discrete or neural models for new event detection, clustering, and summarization.

The authors in [32] proposed a model that consists of three steps: data collection and pre-processing, feature extraction, and class prediction. They collected two balanced sets of tweets. The first set, which represented the positive class, contained tweets retrieved from cybersecurity accounts, while the negative set was tweets retrieved from non-security specialized accounts, such as health, news, and magazines. Next, for feature extraction, they used TF-IDF. In the class prediction step, the binary naïve Bayes (NB) classifier was trained using a 10-fold validation approach, which resulted in an average accuracy of 77.90% for tweet classification into security-related or not.

The authors in [33] proposed a DL classification model based on domain-specific and contextual embeddings to extract features from raw tweets. These features are convolved using a meta-encoder and then combined to be sent to the CNN, LSTM, and contextual encoder for feature learning in parallel. The resultant feature maps were concatenated with contextual embeddings in a fusion layer. A softmax classifier was used for the final prediction for each tweet as security-related or not. Compared to a set of ML and NN baseline models, the proposed model performed better with accuracy of 82%, precision of 79%, 72% recall, and an F1 score equal to 76%.

The authors in [34] used a CNN to classify tweets containing security keywords as security-related or not. All tweets were retrieved from security-specialized Twitter accounts that mentioned three predefined organizations or their assets. The results obtained were using GloVe and word2vec embedding and also used random initialization trained on the classification task. The embedded tweets were fed into three CNN layers in parallel to be convolved with a different number of filters and filter sizes. The researchers suggested a named entity recognition (NER) step to extract the main entities in the tweet using bidirectional LSTM (BiLSTM). The results confirmed that the CNN model performed better than traditional ML techniques. The classification performance achieved 94% recall and 91% TNR, while the NER achieved a 92% F1 score with specifying appropriate entities.

Recent studies that reviewed in this section used a CNN, which opened the door for more investigation and encouraged more studies to improve the accuracy of detecting potential security attacks. The present study aimed to implement the new CapsNet algorithm for the first time in the field of attack detection using Twitter.

IV. Implementation

Before describing the model implementation, a general representation of the tweet classification model that has the purpose of cyberthreat detection is illustrated. As shown in Fig. 2, it contains three layers: an input layer, a classification algorithm, and an output layer. The input layer holds the tweets to be classified, which pass to the classification algorithm in the second layer, the CapsNet in this work. The final goal of this architecture is the label's prediction of each tweet, which is the task of the output layer in labeling each tweet as security-related or not security-related.

The CapsNet model that is proposed for classification purpose is the main contribution of this work. The input of the model is a tokenized tweet of n words, and the output is the predicted class of this tweet. The same principles used in [18] for MNIST handwritten digit data set classification will be followed and adapted to be compatible with the tweet data set. The architecture of the model is shown in Fig. 3 and described in the following sections.

A. Embedding Layer

This layer acts as a link between the input layer and the NN because the NN does not understand the textual input. If n indicates the number of words in a tweet, then each tweet is represented by an array of length n. Thus, there is a need to convert each word into a numeric representation using a word embedding model. Each word will be mapped to its corresponding numeric representation in the embedding model. According to this description, the embedding layer converts the tokenized tweet from an n-dimensional vector to an $n \times d$ dimensional tensor of a floating points matrix to be sent to the next layer.

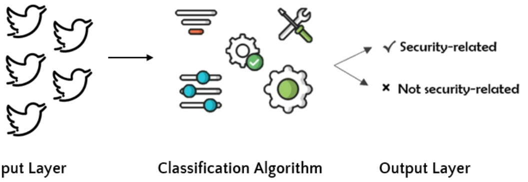


Fig. 2. The general architecture of the cyberthreat detection model using Twitter.

B. Convolutional Layer

The first convolutional layer is a regular convolutional layer that the embedded tweet is fed to before passing to the primary capsule layer. It convolves the embedding matrix with a set of filters f and a kernel size $k \times d$. This means it processes k words at a time, which results in a tensor with size $f(n-k+1)$.

C. Primary Capsule Layer

The primary capsule layer is fully connected to the next layer and consists of three transformations performed sequentially:

- Second convolutional layer: Similar to the previous convolutional layer. It performs a convolution operation on its input with a kernel size k, which reduces the input by $k+1$.
- Reshape: As mentioned previously, one of the significant contributions of CapsNet is that it deals with vectors. Scalar is a quantity with magnitude only, while a vector is a quantity with the magnitude as well as direction. This layer is added to reshape the input feature maps, scalar values, to an output vector map of the desired dimensions to get a set of vectors (capsules) instead of scalars.
- Squashing function: To ensure that all vectors' lengths, which represent a probability, are between 0 and 1 while preserving the orientations of the vectors (features detected) as the following equation [18]:

$$v_j = \frac{\|s_j\|^2}{1+\|s_j\|^2}\frac{s_j}{\|s_j\|} \tag{1}$$

where v_j is the vector output of capsule j and s_j is its total input.

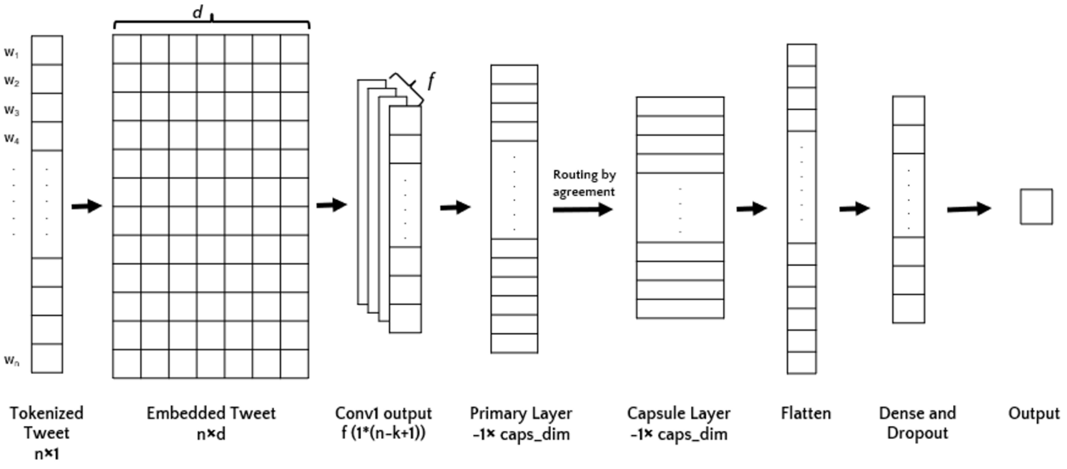


Fig. 3. CapsNet architecture for cyberthreat detection.

D. Capsule Layer

At the point, between the primary and the capsule layer, the novel routing algorithm sits. The goal of routing-by-agreement is to send the output of the lower-level capsule (output of the primary capsule) with high weights to the capsule in the next layer (capsule layer) that it agrees with. To do that, it calculates the predicted output of the next layer by learning routing coefficients in multiple routing rounds. In other words, it strengthens routing weights where predictions made by primary capsules match secondary capsule outputs based on the routing algorithm proposed in [18]. Between the CNNs, which usually implement routing by max pooling that results in loss of some information and the fully connected layers, routing-by-agreement reduces the noise forwarded to the next layer while keeping all the desired information for accurate classification.

E. Flatten, Dense, and Dropout Layers

The output of the previous layer (capsule layer) is a two-dimensional array/matrix, and the next layer is a dense layer that expects a one-dimensional array. The flatten layer is responsible for transforming the two-dimensional matrix of features into a vector by stacking the rows next to each other in a way that can be fed into a fully connected layer for prediction. Then, instead of using the decoder proposed in the work by [18], dropout is used as a regularization method against overfitting [35], which will drop a percentage of the neurons in the flatten layer randomly [36].

F. Output Layer

Since the problem of this work was a binary classification, the final layer of this architecture is a dense layer that predicts the class of each input tweet. Many activation functions can be used to accomplish the aim of this layer, such as softmax or sigmoid, which labels the tweet to be security-related (positive) or not security-related (negative).

TABLE I. Data Set Statistics.

	Training	%	**Validation**	%	**Testing**	%
Positive	2,134	50.14	300	50	712	50
Negative	2,122	49.86	300	50	712	50
Total	4,256	100	600	100	1424	100

V. Experiments

In this section, the hardware and software configurations used in the experiments is reviewed. In addition, the data set that were used, the embedding layer specifications, the baseline CNN model that was proposed for comparison purposes, and the optimization process that was conducted to tune the models' hyperparameters will be described.

A. Tools

The experiments were run on Google Colab [37], a free cloud-based service, with a Tesla P4 GPU and 25 GB of RAM. The code was implemented in Python 3.6.9 with Keras 2.2.4 [38], using TensorFlow 1.15.2 [39] as a backend.

B. Data Set

The data set that satisfied the model requirements was the data set created in the work [34]. It contains tweets that were retrieved from predefined Twitter security-related accounts that mention the infrastructures being monitored or its assets and denoted as A, B, and C. The use of specialized Twitter accounts eliminated the retrieval of tweets containing the desired keyword without security context, such as the words "apple, windows, network, virus, worm, root." The data set was already divided into three sets: training, validation, and testing. Two sets of security specialist accounts, denoted as S1 and S2, were used. The training and validation sets contained the tweets that were retrieved from the S1 accounts, while the testing set was compiled from the S1 and S2 Twitter accounts. The goal of using different sets of Twitter accounts was to give us insights about the models' performances on not only unseen tweets but also tweets retrieved from a different set of Twitter accounts. Another property for the data set was the

TABLE II. Convolutional neural network baseline random search specification and results.

Layer	Parameters	Search Values	Best Value	
			Random	Word2vec
Convolutional Layer 1	Number of filters	[128, 192, 256, 320, 384, 448, 512]	256	384
	Filter size	[3,5,7,9]	7	3
Convolutional Layer 2	Number of filters	[128, 192, 256, 320, 384, 448, 512]	128	128
	Filter size	[3,5,7,9]	7	7
Convolutional Layer 3	Number of filters	[128, 192, 256, 320, 384, 448, 512]	384	256
	Filter size	[3,5,7,9]	9	3
Dense Layer 1	Number of neurons	[64]	64	64
Dropout	Dropout rate	[0.1, 0.2, 0.3, 0.4, 0.5]	0.1	0.3

TABLE III. CapsNet Random Search Specifications and Results.

Layer	Parameters	Search Values	Best Value	
			Random	Word2vec
Convolutional Layer	Number of filters	[128, 256, 384, 512	384	384
	Filter size	[3, 5, 7, 9]	7	5
Primary Capsule Layer	Number of filters	[16, 24, 32, 40, 48, 56, 64]	48	24
	Filter size	[3, 5, 7, 9]	3	9
	Primary capsule dimensions	[8, 12, 16, 20, 24]	24	12
Capsule Layer	Secondary capsule dimensions	[8, 12, 16, 20, 24]	16	12
	Number of capsules	[64]	64	64
	Number of routing iterations	[3, 4, 5, 6]	6	5
Dense Layer	Number of neurons	[64, 128, 192, 256]	64	256
Dropout	Dropout rate	[0.1, 0.2, 0.3, 0.4, 0.5]	0.3	0.3

time interval of the tweets, where the validation and testing sets were retrieved from time intervals following the training set. This means that the obtained results would simulate the real deployment of the model. Then, the collected tweets were filtered based on a set of keywords describing the selected organizations and labeled as security or not.

C. Data Set Retrieval and Statistics

Because of Twitter's policy, which prevents publishing tweets in plain text, the data set was only available in the form of (tweet_ID, label). Thus, a Twitter developer account was created to retrieve the text of the tweets knowing the IDs using Python and Tweepy library. At the time of retrieving the tweets, some were missed due to deletion by the user or the user being suspended. The data set was manipulated to serve the work objectives as follows: the tweets from the three infrastructures were merged and duplicates were deleted since division by the organization was out of the scope of this study. In addition, to work with a balanced data set, the validation and testing tweets were merged, and 300 tweets from each class were specified as a validation set and the remaining as testing tweets, while keeping the classes balanced. The resultant data set statistics are shown in Table I.

D. Data Set Pre-Processing

The raw tweets were cleaned in a pre-processing step, the approach used by [34] for the same data set was followed. In detail, each tweet was converted into lowercase, special characters other than "." and "-" were removed and replaced with a dot and hyphen, respectively. These symbols were needed because they could exist in the software versions and common vulnerabilities and exposures (CVE) numbers. Then, all numbers were converted into its textual representation to be analyzed as text, which resulted in tokenized tweets that were the input for the embedding layer, as described in Section IV-A.

E. Word Embedding

The embedding layer received the tokenized tweet results from the data pre-processing and converted each word into a high-dimensional vector. A widely used NLP technique for feature extraction that represents the semantic meaning of words is word embedding [40]. In this work, two ways of initializing the embedding matrix were examined: using Keras embedding [38] and word2vec pre-trained word embeddings [41], both with 300 dimensions.

F. Baseline Convolutional Neural Network (CNN) Model

In order to choose the appropriate architecture for the baseline model that was used for purposes of comparison, the CNN model used as a baseline in the CapsNet-MNIST proposal [18] was manipulated to be suitable for the text data set. In [18], the CNN and CapsNet models were not architecturally similar, but the authors designed them with similar computational efforts that served the work's objectives. Similarly, the baseline CNN model of this work was built with three convolutional layers, flatten layer, dense layer, and dropout layer, and then added a last dense layer for final prediction.

G. Hyperparameter Tuning

Hyperparameters are the model's parameters that were not included in the training. These parameters should be set carefully because they affect how the model will learn from the data. The manual selection of these values could be less than optimal, and the solution for that problem is hyperparameter tuning. This step was performed because one of the work objectives was finding the optimal values that would lead to the best model performance and generate acceptable results. The random search is used for optimization [42]. For a fair comparison, 100 combinations of each model were tested, the CNN and CapsNet in 200 epochs and early stopping after five epochs. Table II lists all the layers that were included in the

TABLE IV. ACCURACY AND F1 SCORE RESULTS.

Model	Embedding	Accuracy	F1 Score
CNN baseline pooling	Random	91.01	90.84
	Word2vec	91.15	90.92
CNN baseline	Random	91.43	91.44
	Word2vec	91.64	91.46
CapsNet	Random	91.85	91.78
	Word2vec	**92.21**	**92.20**

search process, the values to be examined for each hyperparameter, and in the final column, the optimal values based on the validation set results. Similarly, Table III shows the hyperparameter tuning specification for the proposed CapsNet model. To reduce the total number of combinations, hyperparameters such as the optimizers or the activation functions were not included because we fixed them in both models, and for batch size and learning rate, we kept the default values.

VI. RESULTS AND DISCUSSION

This study was conducted to investigate the use of CapsNet for cyberthreat detection by classifying tweets into security-related or not security-related. The model was built based on hypothesizing and aiming at proving that CapsNet could classify security tweets more accurately than a CNN and that routing-by-agreement is more efficient than pooling. In order to verify the correctness of the work hypothesis, the final architectures generated from the hyperparameter tuning process was trained to minimize the validation loss using binary cross-entropy loss function, and evaluated them using the classification accuracy and F1 score. Classification accuracy is the ratio of correct predictions (positive and negative) to the total number of samples and is computed as in [43]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

while the F1 score is calculated as:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

For the first hypothesis, mentioned above, the CapsNet model was compared with a CNN model that did not include a pooling layer. However, the second hypothesis was tested by comparing the CapsNet with a CNN model that had a pooling layer to prove the efficiency of the routing over the pooling. As can be seen in Table IV, the proposed model achieved competitive results over the strong baseline models. In addition, the accuracy and F1-score comparisons are presented in Fig. 4 and Fig. 5 respectively. In the three models, using the pre-trained word embedding word2vec gave better results than the randomly initialized ones. In general, CapsNet models' results were better than the CNN, followed by the CNN with a pooling layer. The CapsNet model with word2vec achieved the best results, with 92.21% accuracy and 92.20% F1 score, while the worst result was related to the CNN model with a pooling layer at 91.01% accuracy and an F1 score of 90.84%. By comparing the CNN models to each other, it became clear that the use of a pooling layer in the text classification tasks would not be a wise choice, at least in the context of this work.

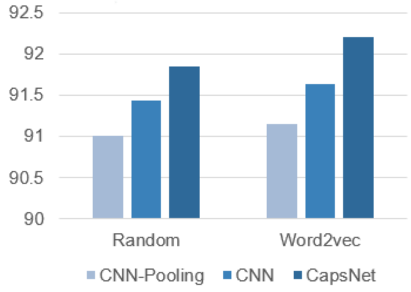


Fig. 4. Accuracy Comparison.

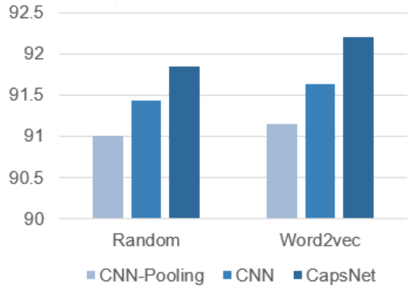


Fig. 5. F1-score Comparison.

VII. CONCLUSION

Securing software, hardware, data, and services has become a crucial part of any organization's management due to the increasing numbers of emerging attacks that threaten its security. In this study, the novel DL algorithm CapsNet was utilized along with Twitter data to provide accurate classification of tweets for the purposes of cyberthreat detection. A random search was used for hyperparameter tuning for random and word2vec embeddings. CapsNet model was built based on the hyperparameters was found after the random search, then the model was compared with two CNN architectures: a CNN baseline model without a pooling layer and a CNN baseline pooling model that included a pooling layer. The model was evaluated using accuracy and F1 score, and from the results, multiple remarks were gleaned. First, it was proved the better efficiency of routing-by-agreement compared to pooling. Second, in the three models, the pre-trained word embedding word2vec achieved better results than random embedding. Third, the proposed CapsNet model outperformed the strong competitor of CNN, with 92.21% accuracy and 92.2% F1 score.

Because there is always room for improvement, we plan to compare the obtained results with a recurrent neural network (RNN) model given that CapsNet introduces improvements on its architecture. In addition, we aim to examine replacing the CNN layers in the CapsNet with RNN to take advantage of dealing with tweets word-by-word rather than the whole tweet at once. In addition, we aim to test more embeddings, such as GloVe, Fasttext, BERT, and Elmo.

ACKNOWLEDGMENT

This work was supported by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under grant No. (DG-1441-60-612). The authors, therefore, gratefully acknowledge the DSR technical and financial support.

References

[1] H. Booth, D. Rike, and G. Witte, "The national vulnerability database (nvd): Overview," National Institute of Standards and Technology, Tech. Rep., 2013.

[2] O. Security, *Offensive Security's Exploit Database Archive*, (accessed May 31, 2020). [Online]. Available: https://www.exploit-db.com/

[3] S. Trabelsi, H. Plate, A. Abida, M. M. B. Aoun, A. Zouaoui, C. Missaoui, S. Gharbi, and A. Ayari, "Mining social networks for software vulnerabilities monitoring," in *2015 7th International Conference on New Technologies, Mobility and Security (NTMS)*. IEEE, 2015, pp. 1–7.

[4] C. Sabottke, O. Suciu, and T. Dumitraș, "Vulnerability disclosure in the age of social media: exploiting twitter for predicting real-world exploits," in *24th {USENIX} Security Symposium ({USENIX} Security 15)*, 2015, pp. 1041–1056.

[5] S. Samtani, R. Chinn, H. Chen, and J. F. Nunamaker Jr, "Exploring emerging hacker assets and key hackers for proactive cyber threat intelligence," *Journal of Management Information Systems*, vol. 34, no. 4, pp. 1023–1053, 2017.

[6] L. G. A. Rodriguez, J. S. Trazzi, V. Fossaluza, R. Campiolo, and D. M. Batista, "Analysis of vulnerability disclosure delays from the national vulnerability database," in *Anais do I Workshop de Segurança Cibernética em Dispositivos Conectados*. SBC, 2018.

[7] C. Sauerwein, C. Sillaber, M. M. Huber, A. Mussmann, and R. Breu, "The tweet advantage: An empirical analysis of 0-day vulnerability information shared on twitter," in *IFIP International Conference on ICT Systems Security and Privacy Protection*. Springer, 2018, pp. 201–215.

[8] R. Campiolo, L. A. F. Santos, D. M. Batista, and M. A. Gerosa, "Evaluating the utilization of twitter messages as a source of security alerts," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, 2013, pp. 942–943.

[9] D. Kergl, R. Roedler, and G. D. Rodosek, "Detection of zero day exploits using real-time social media streams," in *Advances in Nature and Biologically Inspired Computing*. Springer, 2016, pp. 405–416.

[10] M. Almukaynizi, E. Nunes, K. Dharaiya, M. Senguttuvan, J. Shakarian, and P. Shakarian, "Proactive identification of exploits in the wild through vulnerability mentions online," in *2017 International Conference on Cyber Conflict (CyCon US)*. IEEE, 2017, pp. 82–88.

[11] A. Hernández, V. Sanchez, G. Sánchez, H. Pérez, J. Olivares, K. Toscano, M. Nakano, and V. Martinez, "Security attack prediction based on user sentiment analysis of twitter data," in *2016 IEEE international conference on industrial technology (ICIT)*. IEEE, 2016, pp. 610–617.

[12] A. Sliva, K. Shu, and H. Liu, "Using social media to understand cyber attack behavior," in *International Conference on Applied Human Factors and Ergonomics*. Springer, 2018, pp. 636–645.

[13] A. Rodriguez and K. Okamura, "Generating real time cyber situational awareness information through social media data mining," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2. IEEE, 2019, pp. 502–507.

[14] A. Hernandez-Suarez, G. Sanchez-Perez, K. Toscano-Medina, V. Martinez-Hernandez, H. Perez-Meana, J. Olivares-Mercado, and V. Sanchez, "Social sentiment sensor in twitter for predicting cyber-attacks using 1 regularization," *Sensors*, vol. 18, no. 5, p. 1380, 2018.

[15] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[16] C. Y. Chang, Z. Teng, and Y. Zhang, "Expectation-regulated neural model for event mention extraction," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 400–410.

[17] G. Hinton., *Machine Learning, reddit*, (accessed May 31, 2020). [Online]. Available: https://www.reddit.com/r/MachineLearning/comments/2lmo0l/ama_geoffrey_hinton/clyj4jv/

[18] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in neural information processing systems*, 2017, pp. 3856–3866.

[19] M. K. Patrick, A. F. Adekoya, A. A. Mighty, and B. Y. Edward, "Capsule networks–a survey," *Journal of King Saud University-Computer and Information Sciences*, 2019.

[20] J. Kim, S. Jang, E. Park, and S. Choi, "Text classification using capsules," *Neurocomputing*, vol. 376, pp. 214–221, 2020.

[21] W. Zhao, J. Ye, M. Yang, Z. Lei, S. Zhang, and Z. Zhao, "Investigating capsule networks with dynamic routing for text classification," *arXiv preprint arXiv:1804.00538*, 2018.

[22] H. W. Fentaw and T.-H. Kim, "Design and investigation of capsule networks for sentence classification," *Applied Sciences*, vol. 9, no. 11, p. 2200, 2019.

[23] Y. Du, X. Zhao, M. He, and W. Guo, "A novel capsule based hybrid neural network for sentiment classification," *IEEE Access*, vol. 7, pp. 39 321–39 328, 2019.

[24] Y. Wang, A. Sun, J. Han, Y. Liu, and X. Zhu, "Sentiment analysis by capsules," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 1165–1174.

[25] M. H. Goldani, S. Momtazi, and R. Safabakhsh, "Detecting fake news with capsule neural networks," *arXiv preprint arXiv:2002.01030*, 2020.

[26] J. Liu, H. Lin, X. Liu, B. Xu, Y. Ren, Y. Diao, and L. Yang, "Transformer-based capsule network for stock movement prediction," in *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, 2019, pp. 66–73.

[27] P. Rathnayaka, S. Abeysinghe, C. Samarajeewa, I. Manchanayake, and M. Walpola, "Sentylic at iest 2018: Gated recurrent neural network and capsule network based approach for implicit emotion detection," *arXiv preprint arXiv:1809.01452*, 2018.

[28] H. Hettiarachchi and T. Ranasinghe, "Emoji powered capsule network to detect type and target of offensive posts in social media," in *Proceedings of RANLP*, 2019.

[29] A. Ritter, E. Wright, W. Casey, and T. Mitchell, "Weakly supervised extraction of computer security events from twitter," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 896–905.

[30] F. Alves, A. Bettini, P. M. Ferreira, and A. Bessani, "Processing tweets for cybersecurity threat awareness," *arXiv preprint arXiv:1904.02072*, 2019.

[31] Z. Wang and Y. Zhang, "A neural model for joint event detection and summarization." in *IJCAI*, 2017, pp. 4158–4164.

[32] Y. Erkal, M. Sezgin, and S. Gunduz, "A new cyber security alert system for twitter," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2015, pp. 766–770.

[33] S. Yagcioglu, M. S. Seyfioglu, B. Citamak, B. Bardak, S. Guldamlasioglu, A. Yuksel, and E. I. Tatli, "Detecting cybersecurity events from noisy short text," *arXiv preprint arXiv:1904.05054*, 2019.

[34] N. Dionísio, F. Alves, P. M. Ferreira, and A. Bessani, "Cyberthreat detection from twitter using deep neural networks," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[35] A. Khikmatullaev, J. Lehmann, and K. Singh, "Capsule neural networks for text classification," Ph.D. dissertation, 04 2019.

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[37] Google, "Colaboratory: frequently asked questions." (accessed May 31, 2020). [Online]. Available: https://research.google.com/colaboratory/faq.html

[38] F. Chollet *et al.*, "Keras documentation," *keras. io*, 2015.

[39] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[40] S. Wang, W. Zhou, and C. Jiang, "A survey of word embeddings based on deep learning," *Computing*, vol. 102, no. 3, pp. 717–740, 2020.

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of

word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[42] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of machine learning research*, vol. 13, no. Feb, pp. 281–305, 2012.

[43] J. M. Torres, C. I. Comesaña, and P. J. García-Nieto, "Machine learning techniques applied to cybersecurity," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 10, pp. 2823–2836, 2019.

Estimating the Causes of Poor Academic Performance of Students: A Case Study

Juwesh Binong
Department of Electronics & Communication Engineering
North-Eastern Hill University
Shillong, India 793 022

***Abstract*—Poor academic performance of students is not the only concern for parents and teachers, but also a concern for the country as a whole. This paper makes an attempt to identify the cause(s) of poor academic achievement. This paper presents a method of identifying the most influencing factor on academic performance. The proposed method capable of using qualitative ratings as input for the factors considered and find the correlation of each factor with academic performance, and finally rank the influences of the factors on performance to sort out the most influencing one. The study was carried out on the academic performance of 189 students of B.Tech for five academic semesters. The results indicate the degree influences of various factors on performance, with the most influencing one being the academic ability of students.**

Keywords*—Academic performance; qualitative rating; factors; correlation coefficient; analytic hierarchy process

I. Introduction

Good academic performance is very important for developing country like India [1, 2]. By understanding the cause of poor academic performance, the concerned authority can take appropriate decision to improve academic performance or achievement of students [3].

The majority of the literature is, however, based on surveys [4, 5, 6] and self-reports, methods which have well-known systematic biases that lead to limitations on conclusions and generality as well as being costly to implement.

Academic performance by a student is the result of various contributing factors. Literature lists multiple factors that affect academic performance of students. Academic performance is a complex equation with multiple angles. Effect of various factors on academic performance of students can not be denied.

Identifying the factors that influence academic performance is an essential part of educational research [7]. kassarnig et al. have shown that important indicators of academic performance are based on social ties. They confirm that class attendance stands as the most important predictor, but other factors like peer effect also influence on academic performance.

Attendance as a factor influencing a student's academic achievement is seen in literature. In [8] authors have demonstrated that consistent class attendance strongly correlates with academic achievement. They also demonstrated that their dataset allowed them to determine that attendance among social peers was substantially correlated (> 0.5), suggesting either an important peer effect or homophily with respect to attendance.

Credé et al. [9] have shown that class attendance and grades reveal a strong relationship with both class grades and GPA. According to their meta-analysis class attendance stands as a better predictor of college grades than any other known predictors of academic performance.

Influence of peers on academic performance cannot be denied. The peer effects start out from the assumption that human behavior is affected not only by personal and demographic features, but also the surrounding environment and to the individuals with whom he/she interacts [10].

The place of residence also plays an important role in students' academic achievement. Snyder et al. [11] concluded that living in an on campus or off campus environment had no statistical relationship with the academic performance of freshman student athletes. The study of Etikan et al. [12] suggested that there is no significant difference in the academic performance of the students residing on campus and outside the school environment, but found some influence on the choice of student accommodation preference.

The academic ability or intelligence is seen as an important factor on students' academic achievement. Intelligence is considered as the strongest predictor of academic achievement with correlations ranging from 0.30 to 0.70 [13]. They investigated the correlation between standardized intelligence tests and school grades using psychometric meta-analysis. The study results of Colom and Flores-Mendoza [14] indicate that socioeconomic status factors do not predict children differences in scholastic achievement, whereas children intelligence tests' scores predict their scholastic differences. These results underscore personal intelligence as a genuine predictor of individual differences in scholastic achievement. The association between intelligence and academic performance is well established in [15]. In [16], the authors showed the existence of a strong correlation between a latent intelligence trait and a latent trait of educational achievement.

Gbollie and Keamu [17] explored the motivational beliefs and learning strategy use by Liberian junior and senior high school students in connection with their academic performance.

Most of the factors influencing academic performance take qualitative data as input. A popular method that can accept qualitative input and ranks a set of factors is the analytic hierarchy process (AHP) [18], a popular Multiple-criteria decision-making (MCDM) method. AHP is a flexible but well-structured methodology for organizing and analyzing complex decisions [19], originally developed by Prof. Thomas

L. Saaty [20]; and it is widely used in a wide variety of decision making situations, such as government, business, industry, shipbuilding, education, health-care, etc. [21, 22, 23].

AHP can be realized by three steps: 1) constructing the AHP hierarchy; 2) making a pairwise comparison of the elements of the hierarchical structure, and 3) aggregation of an overall priority rating to select the best candidate. In AHP human judgments can be used in performing the evaluations.

AHP has also been extended making it suitable to use with other mathematical methods including fuzzy logic. Yager and Kelman [24] extended the AHP by integrating the fuzzy linguistic ordered weighted averaging operators and thereby enhancing the capabilities of AHP as a comprehensive tool for decision-making. Li et al. [25] used the AHP with fuzzy inference technique in the dynamic route guidance system to provide dynamic routing advice based on real-time traffic information.

A. Contribution

The main contribution of this paper is a method that utilizes qualitative rating to study the influence of various factors on students' academic performance. As the method proposed follows the steps of AHP, it can accept approximate values as input, where getting an exact quantitative value is difficult or impossible.

The effectiveness of the proposed method is established using a dataset of more than one hundred students containing more than one three hundred records.

The rest of the paper is organized as follows. Section 2 formulates the problem. The data collection approach and the proposed method is discussed in Section 3. Experimental results and discussion on the outcomes are shown in Section 4. Finally, conclusion and future work are discussed in Section 5.

II. Problem Formulation

Factors affecting students academic performance can be categories based on various factors.

Considering the factors as variables the problem of students academic performance can be formulate as follows:

For a given dataset D containing a parameter R which is influenced by a set of features F=$\{f_1, f_2, \cdots f_n\}$, the problem is to select an optimal feature f_x that influence the most or a set of features with their amount of influences on R.

III. Materials and Methods

A. Candidates for the Study

The candidates for this case study comprises of 312 academic records (percentage of marked obtained) of 189 students of B.Tech in Electronics & Communications Engineering of North Eastern Hill University (NEHU), Shillong, India, covering academic years 2012-13, 2013-14, 2014-15 and 2015-16. Majority of the students belong to the Indian states of Assam, Meghalaya, other north eastern (Indian) states, and other parts of the country including West Bengal, Bihar, Uttar Pradesh etc. More than 80% of the students belong to Schedule Tribe and Schedules Class categories. Of the population, 70% of students were male and 30% females, and in the age group of 17-21 years.

B. Factors Considered for the Study

Poor academic performance may be the result of one or more factors. As seen from literature a number factors influence on academic performance. In this work only four factors are targeted; they are : academic ability, attendance, peers' effect and residential effect.

1) Academic Ability: Students' academic ability is taken as an important factor for this work, as the students considered for the research are from different parts of the neighbouring regions and they possess a wide range of past academic records.

2) Attendance: The percentage of course attendance of the students under consideration varies a wide range : very poor attendance as low as below 50% to a high of 100%. This makes course attendance a possible cause of poor performance.

3) Peers' Effect: Peer effect is considered as a possible cause of poor academic performance as two or more students forming group of social life seem to achieve nearly same type of academic performance. It is observed that students sharing a common desk in the class or sharing social life inside and outside the campus generally obtain similar or nearly equal grades.

4) Residence Effect: The place of residence and its environment is taken as an influencing factor on academic performance. Students mostly reside in hostel inside the university campus and his/her social peer does not always share the same room. But it is seen that students sharing same room performing the same way in some cases.

C. Data Collection

Data of students academic achievements as well as of other factors were collected for the B.Tech students of Electronics and Communication Engineering, North Eastern Hill University (NEHU), Shillong, India.

The percentage obtained by the students of B.Tech in Electronics & Communications Engineering in their end semester examinations was collected from the department. Data field for hostel residential information was collected from hostel wardens and roommate quality was assigned based on his/her academic achievement. Information about peer was collected from their friend circle and teachers. Collection of peers' information was not very difficult as the students belong to our department.

At the university, course attendance of 75% is mandatory to be eligible to appear in the end semester examination.

In terms of privacy of candidates considered for the study their identity have been changed.

D. Assigning Qualitative Rating

Other than the percentage of attendance, most of the factors considered for the study possesses no quantitative value. What one can get for these factors is qualitative value as input. For example, peer of a student can be bad or good. Similarly a hostel mate may a good or a bad one. Qualitative rating for the factors considered was assigned as given in the following:

1) Academic Achievement: The academic achievement of a student was considered in term of the CGPA (overall percentage of marks) obtained for each subject in the end semester examination.

The percentage of marks obtained by the students were converted to equivalent percentage range. Based on the performance in each subject in a semester, qualitative ratings were assigned to a student for each subject using Table I.

TABLE I. QUALITATIVE RATING FOR ACADEMIC PERFORMANCE

Percentage Secured	Qualitative Rating
90-100	Very Good performance
70-89	Good performance
50-69	Average performance
30-49	Poor performance
less than 30	Very Poor performance

Scores were assigned to the qualitative rating obtained based on the intensity of achievement. The use of such score has advantages over grade obtained by students because a student who fails get 'F' grade which equals 4 in 10 point scale and can not consider lower grade than 4.

The qualitative rating used to measure the qualities and the corresponding score for various factors are presented in Table II.

TABLE II. QUALITATIVE RATING & RATING SCORE USED

Qualitative Rating	Rating Score
Very Good	9
Good	7
Average	5
Poor	3
Very Poor	1

2) Qualitative Rating for Academic Ability: An absolute judgment of academic excellence of a student in term of some number always does not give a correct judgment. Students with the same percentage but from different universities may be not of the same academic ability or excellence. Hence, students' academic ability given in term of percentage are rated using qualitative rating. Students' academic ability was rated based on their past performance as shown in Table III.

Past academic achievement by a student was considered as a factor to judge the academic ability of a student. The grade obtained by a student in the lower semester were used as past academic performance. Though grade obtained in class-X and class-XII would be the ideal choice, were not use because of non-availability of data.

TABLE III. QUALITATIVE RATING ON ACADEMIC ABILITY OF STUDENTS

Qualitative Rating	Meaning
Very Good academically	Student securing marks 90% or above
Good academically	Student securing marks between 80-89%
Average academically	Student securing marks between 70-79%
Poor academically	Student securing marks between 60-69%
Very Poor academically	Student securing marks between 45-59%

3) Qualitative Rating for Attendance: Attendance percentage acquired by students for a subject is taken as an indicator of self-motivation. Attendance roughly represents a parameter of motivation. As two equal percentage of attendance generally do not represent the same of amount of motivation of two students, percentage of attendance acquired by students was transformed into grading scale as shown in Table IV. The rating has been taken a non-equal division approach as generally academic achievement does not follow a linear division.

TABLE IV. QUALITATIVE RATING ON ATTENDANCE

Qualitative Rating	Meaning
Very Good attendance	Securing attendance 95% or above
Good attendance	Securing attendance between 85-94%
Average attendance	Securing attendance between 75-84%
Poor attendance	Securing attendance between 60-74%
Very Poor attendance	Securing attendance below 60%

4) Qualitative Rating for Peers' Effect: Assigning quantitative values to measure the quality of a student's peer is not logical. Generally we humans use qualitative terms like 'good', 'very good', 'bad', 'very bad' etc to refer the quality of one's friend. So qualitative terms were used to refer to the quality of a student's peer. A qualitative score was awarded to a student base on the performance of his/her peer with whom the student spends most of the times; and rating is shown in Table V. For example, if 'Y' is the peer of 'X', and academic performance of 'Y' is good, 'X' receives higher score.

5) Qualitative Rating for Residence Effect: Assigning quantitative values to measure the quality of a student's roommate is not logical as in the case of peer effect. Hence, a qualitative score was awarded to a student base on the academic performance of his/her roommates. The rating is shown in Table VI.

Students residing with parents were assigned 'very good' rating assuming a favourable condition at home; and students residing in shared accommodation in rented house were

TABLE V. Qualitative Rating on Peers' Effect

Qualitative Rating	Meaning
Very Good Peer	Very good academic performance by his/her peers
Good Peer	Good academic performance by his/her peers
Average Peer	Average academic performance by his/her peers
Poor Peer	Poor academic performance by his/her peers
Very Poor Peer	Very poor academic performance by his/her peers

TABLE VI. Qualitative Rating on Residential Effect

Qualitative Rating	Meaning
Very Good Residence	Very good academic performance by his/her roommates
Good Residence	Good academic performance by his/her roommates
Average Residence	Average academic performance by his/her roommates
Poor Residence	Poor academic performance by his/her roommates
Very Poor Residence	Very poor academic performance by his/her roommates

awarded rating using Table VI to consider the effect of roommates.

E. Ranking of Factors

Getting true quantitative values for the listed factors are difficult and almost impossible. There are chances of being erroneous or biases of these values. AHP hold promises in such situation by accepting qualitative values as input. The following steps were performed in determining the most important one among the factors.

1) Calculating Correlation Coefficients: Correlation coefficient of two random variables, say X and Y shows how strongly the values of these variables are related to one another. The Pearson's correlation coefficients ρ_{XY} of two random variables *X* and *Y*, denoted by $Corr(X,Y)$ or $\rho_{X,Y}$ is given by [26] as:

$$\rho_{XY} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} \tag{1}$$

where x_i and y_i are the values of random variables *X* and *Y* for $i = 1, 2, \cdots, n$; and $\overline{x}$ and $\overline{y}$ are the means of x_i and y_i respectively.

If $\{f_1, f_2, \cdots f_n\}$ are factors influencing the academic performance γ, the Pearson's correlation coefficients $\rho_{1,\gamma}$,$\rho_{2,\gamma}$, $\cdots \rho_{n,\gamma}$ for each of the factors with the academic performance γ is calculated using equation 1.

2) Decision Matrix and Ranking of Factors: Correlation coefficient obtained for four different factors were used to access the strength of the judgments. Based on the rating obtained from the rating assignment process, decision matrix [27] was formed and the factors were prioritized by calculating their normalized scores using the three steps as follows:

1) Making of the decison matrix
2) Normalization of each column of the decision matrix
3) Row-wise summation and normalization of row-sums.

Step 1: The decison matrix M is constructed as shown in Equation 2:

$$M = \begin{array}{c} \\ \rho_1 \\ \rho_1 \\ \vdots \\ \rho_n \end{array} \begin{array}{c} \begin{array}{ccccc} \rho_1 & \rho_2 & \rho_3 & \cdots & \rho_n \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nn} \end{bmatrix} \end{array} \tag{2}$$

where each element of the matrix is a ratio of the two correlation coefficients, such as $x_{11} = \frac{\rho_1}{\rho_1}$, $x_{12} = \frac{\rho_1}{\rho_2}$, $\cdots, x_{nn} = \frac{\rho_n}{\rho_n}$.

Step 2: The column sum of the decision matrix (Equation 2) is calculated for each column and each element of the matrix is normalized to get the matrix shown in Equation 3.

$$M = \begin{array}{c} f_1 \\ f_2 \\ \vdots \\ \\ f_n \end{array} \begin{bmatrix} \hat{x_{11}} & \hat{x_{12}} & \hat{x_{13}} & \dots & \hat{x_{1n}} \\ \hat{x_{21}} & \hat{x_{22}} & \hat{x_{23}} & \dots & \hat{x_{2n}} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \hat{x_{n1}} & \hat{x_{n2}} & \hat{x_{n3}} & \dots & \hat{x_{nn}} \end{bmatrix} \tag{3}$$

Step 3: The row sum of the matrix (Equation 3) is calculated to get a column matrix as shown in Equation 4.

$$M = \begin{array}{c} f_1 \\ f_2 \\ \vdots \\ \\ f_n \end{array} \begin{bmatrix} \sum \hat{x_{1n}} \\ \sum \hat{x_{2n}} \\ \cdots \\ \cdots \\ \sum \hat{x_{nn}} \end{bmatrix} = \begin{bmatrix} S(\rho_1) \\ S(\rho_1) \\ \cdots \\ \cdots \\ S(\rho_n) \end{bmatrix} \tag{4}$$

$S(\rho_1), S(\rho_1), \cdots, S(\rho_n)$ are the normalized scores and these values specify the intensity of influence of factors on academic performance, and hence the ranking of the factors are obtained.

Algorithm 1 describes the process of computing the ranks of the factors based on the amount of its influencing on the academic performance.

```
Algorithm 1: The proposed cause estimation method
  Input: Students Academic Dataset
  Output: Ranked factors based on correlation
          coefficients
  forall factor $f_i$ do
      Select a factor $f_i$.
      Compute correlation coefficient $r_i$ for $f_i$.
  end
  Compute rank for all $r_i$.
```

IV. Results and Discussion

In the following the influence of various factors on the academic achievement of students are analysis. The results of the expereimentation are discussed first, followed by the discussion on the results.

A. Results

The correlation coefficients of each factors with the academic performance score of the students were calculated using Equation 1, and are shown in Table VII.

TABLE VII. Correlation Coefficients obtained from the dataset

factors	Academic Ability	Attendance	Peers' Effect	Residence Effect
ρ	0.5480	0.3305	0.3297	0.3364

1) Academic Ability: Result showed that correlation between student academic ability and student academic performance has a strong relation. Academic performance almost linearly increases with student's academic ability (see Fig. 1).

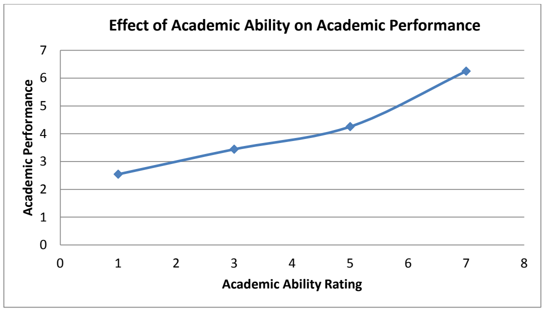


Fig. 1. Effect of academic ability on academic performance

2) Peers Factor: Effect of peers reveals a strong relation on student academic performance as shown in the result (see Fig. 2). The performance of students shows an almost linear effect on the performance of students. Quality of peer shows significant impact on one's academic performance.

3) Residence Factor: Effect of residence on academic performance of students shows impact under certain condition. The poor residential condition shows poor in academic performance, but after certain level, this factor shows no effect as indicated by an almost horizontal line for better residential conditions (see Fig. 3).

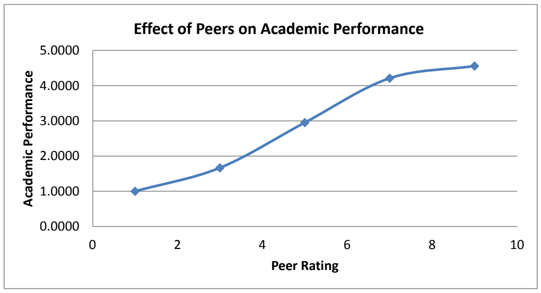


Fig. 2. Effect of peers on academic performance

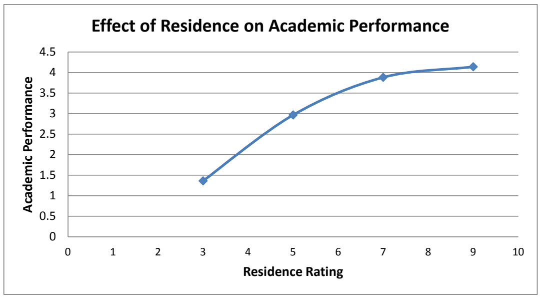


Fig. 3. Effect of residence on academic performance

4) Attendance Factor: Correlation is strongly significant with very low attendance ($\leq$ 50 %). As reported in literature the impact of attendance on the academic performance of students is clearly visible from the plot (see Fig. 4).

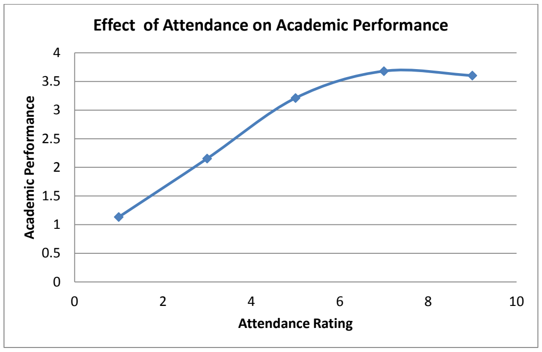


Fig. 4. Effect of attendance on academic performance

5) Decision Matrix: The final normalized scores of the factors as obtained from the decision matrix are shown in Table VIII. These scores specify the intensity of influences of the factors on the academic performance.

TABLE VIII. Normalized scores obtained from the decision matrix

Academic Ability	Attendance	Peers' Effect	Residence Effect
0.3550	0.2157	0.2151	0.2142

B. Discussion

Based on the experimentation, it is observed that academic ability of a student influence strongly in his/her academic achievement with score 0.3550. For achieving good academic performance a student with good academic ability is reqired. The result indicates that the effect of academic ability on performance shows a complex pattern: for poor and good performing students it shows a strong relation, while for average students the effect is less pronounced.

Influence of peers' quality is found to have an almost linear effect on academic achievement. A very bad peer contributes to a very poor academic achievement. A poor residential environment also shows a significant negative effect on one's academic achievement. This factor shows less or no influence under a good residential environment, but this effect holds strong for the low residential condition. Interestingly, this effect vanishes under a good and a very good residential environment.

A general decreasing trend of attendance curve indicates a strong effect of attendance on academic achievement. Lower the attendance is higher the effect on the achievement and higher the attendance is lower the effect. The attendance percentage shows a strong relation on fail; the effect becomes less pronounced at higher percentage of attendance.

The most influencing factor can not be the universal one. Other factor may be the most influencing one in different conditions and in different institute.

V. Conclusion and Future Work

The paper has demonstrated the connection between academic performance and various factors, and shown how different factors influence on the performance, and also the effect of the factors are ranked to arrived at the most influencing one. The proposed approach can be enhanced by considering more factors effecting on academic performance.

The proposed method has shown that this method can be utilized to access students' performance factors for cases having difficulty in getting quantitative data as input. But, more into inside students social behaviours and daily life are required to get detailed knowledge regarding the individuals students so as to able to assign a correct qualitative rating.

Results with different factors indicate that students' academic ability is an important factor for students' academic achievement, but other factors influence on academic performance as well. Further research is may be done to understand to what extent these factors influence students' academic success.

The study can be performed on a bigger dataset with more information.

References

[1] Bhise, R., Thorat, S., and Supekar, A. Importance of data mining in higher education system. *IOSR Journal Of Humanities And Social Science (IOSR-JHSS)*, pages 2279–0837, 2013.

[2] Osmanbegović, E. and Suljić, M. Data mining approach for predicting student performance. *Economic Review*, 10(1):3–12, 2012.

[3] Romero, C. and Ventura, S. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010.

[4] Banerjee, P. A. A systematic review of factors linked to poor academic performance of disadvantaged students in science and maths in schools. *Cogent Education*, 3(1):1178441, 2016.

[5] Al-Zoubi, S. M. and Younes, M. A. B. Low academic achievement: causes and results. *Theory and Practice in Language Studies*, 5(11):2262, 2015.

[6] Wilder, S. Effects of parental involvement on academic achievement: a meta-synthesis. *Educational Review*, 66(3):377–397, 2014.

[7] Kassarnig, V., Mones, E., Bjerre-Nielsen, A., Sapiezynski, P., Dreyer Lassen, D., and Lehmann, S. Academic performance and behavioral patterns. *EPJ Data Science*, 7:1–16, 2018.

[8] Kassarnig, V., Bjerre-Nielsen, A., Mones, E., Lehmann, S., and Lassen, D. D. Class attendance, peer similarity, and academic performance in a large field study. *PloS one*, 12(11):e0187078, 2017.

[9] Credé, M., Roch, S. G., and Kieszczynka, U. M. Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics. *Review of Educational Research*, 80(2):272–295, 2010.

[10] Celant, S. The analysis of students' academic achievement: the evaluation of peer effects through relational links, 2013. *Quality & Quantity*, 47(2):615–631, 2013.

[11] Snyder, E. M., Kras, J. M., Bressel, E., Reeve, E. M., and Dilworth, V. The relationship of residence to academic performance in ncaa division i freshman athletes. *Journal of issues in intercollegiate athletics*, 4:105–119, 2011.

[12] Etikan, I., Bala, K., Babatope, O., Yuvalı, M., and Bakır, I. (2017). Influence of residential setting on student outcome. *Biom Biostat Int J*, 6(4):00177, 2017.

[13] Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., and Spinath, F. M. Intelligence and school grades: A meta-analysis. *Intelligence*, 53:118–137, 2015.

[14] Colom, R. and Flores-Mendoza, C. E. Intelligence predicts scholastic achievement irrespective of ses factors: Evidence from brazil. *Intelligence*, 35(3):243–251, 2007.

[15] Soares, D. L., Lemos, G. C., Primi, R., and Almeida, L. S. The relationship between intelligence and academic achievement throughout middle school: The role of students' prior academic performance. *Learning and Individual Differences*, 41:73–78, 2015.

[16] Deary, I. J., Strand, S., Smith, P., and Fernandes, C. Intelligence and educational achievement. *Intelligence*, 35(1):13–21, 2007.

[17] Gbollie, C. and Keamu, H. P. Student academic performance: The role of motivation, strategies, and perceived factors hindering liberian junior and senior high school students learning. *Education Research International*, 2017.

[18] Saaty, T. L. *The Analytic Hierarchy Process*. New York: McGraw-Hill, 1980.

[19] Saaty, T. L. What is the analytic hierarchy process? In *Mathematical Models for Decision Support*, pages 109–121. Springer, 1988.

[20] Saaty, T. L. A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, 15(3):234–281, 1977.

[21] Saaty, T. L. Decision making with the analytic hierarchy process. *International journal of services sciences*, 1(1):83–98, 2008.

[22] Fong, P. S.-W. and Choi, S. K.-Y. Final contractor selection using the analytical hierarchy process. *Construction Management & Economics*, 18(5):547–557, 2000.

[23] Saracoglu, B. O. Selecting industrial investment locations in master plans of countries. *European Journal of Industrial Engineering*, 7(4):416–441, 2013.

[24] Yager, R. R. and Kelman, A. An extension of the analytical hierarchy process using owa operators. *Journal of Intelligent & Fuzzy Systems*, 7(4):401–417, 1999.

[25] Li, C., Anavatti, S. G., and Ray, T. Analytical hierarchy process using fuzzy inference technique for real-time route guidance system. *IEEE Transactions on Intelligent Transportation Systems*, 15(1):84–93, 2014.

[26] Lee Rodgers, J. and Nicewander, W. A. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[27] Yang, J.-B. and Xu, D.-L. On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 32(3):289–304, 2002.

Sensing of Environmental Variables for the Analysis of Indoor Air Pollution

Jaime Xilot[1], Edgard Benítez-Guerrero[3]
Faculty of Statistics and Informatics
Universidad Veracruzana
Xalapa, Veracruz, Mexico

Guillermo Molero-Castillo[2], Everardo Bárcenas[4]
Engineering Faculty
National Autonomous University of Mexico
Mexico City, Mexico

***Abstract*—Ambient intelligence systems try to perceive the environment and react, proactively and pervasively to improve people's environmental conditions. A current challenge in Ambient intelligence is trying to mitigate environmental risks that affect global public health, such as increasing air pollution. This paper presents the analysis of some environmental variables related to indoor air pollutants, such as CO, $PM_{2.5}$, PM_{10}, humidity and temperature; all of these captured in a university environment. The environmental measurements were carried out through a wireless sensor network consisting of two nodes. The cloud computing service, that is, ThingSpeak, was used as the storage medium. With this network, the presence of pollutants in the study area were detected with concentration levels within the permitted ranges, as well as its correlation with the atmospheric variables of temperature and humidity. The implementation of the sensor network allowed the capture of data in a transparent and non-intrusive way, and the analysis allowed the understanding of the behavior of pollutants in indoor spaces, where air circulation is limited, which in the face of high levels of pollution can be harmful to human health.**

Keywords*—Air pollution; Ambient Intelligence (AmI); indoor air quality; wireless sensor network

I. Introduction

In recent years, Ambient Intelligence (AmI) has gained notoriety due to the need to capture and analyze environmental conditions with the idea of improving the environment of the inhabitants and society in general. A feature of Ambient Intelligence systems is the ubiquitous and pervasive computation that makes them sensitive, responsive and adaptive, without the need to interact with users [1].

The idea of ubiquitous and pervasive computing, in Ambient Intelligence systems, is the presence of communication devices and technologies such as wireless sensor networks (WSN), which work together to monitor environmental conditions, like sound, temperature, pressure, movement, pollution, among others. Each node in the network can perform sensing, processing and wireless communication functions [2], [3], [4].

At present, WSNs are implemented in various areas, for instance health care and monitoring the environment, sports, transportation, entertainment, smart spaces, among others. However, they face various problems and challenges, highlighting the limitation in storage capacity, communication due to its short-range, security, privacy and limited processing resources. One solution for storing data is through cloud computing [5].

Cloud technology has been developed to provide quick and easy access to a set of configurable computing resources, such as networks, servers, applications, services and storage [5]. The latter is offered as an infrastructure service, where the user does not need to worry about the location of the servers, nor about the available resources due to the high scalability of the resources, as they can be increased on demand [2].

Sensor integration and cloud storage have led to the Sensor-Cloud architecture, which is used to capture environmental conditions and accumulate transmitted data for later analysis. In this way, Sensor-Cloud allows users to collect, process, visualize, analyze and share sensor data in an efficient and easy way [2], [5]. Currently, there are several providers of Sensor-Cloud platforms, such as Amazon, Google, Microsoft, ThingSpeak, among others.

Therefore, today it is possible to monitor the conditions of the environment through a wireless sensor network, storing the measurements in the cloud and applying certain algorithms to obtain inferences about these measurements. Precisely one of the growing environmental problems at present is air pollution, which has become a concern of society in general due to the environmental impact and serious damage to the health of people and ecosystems [6]. Various organisms in the world warn of the adverse and devastating effects of air pollution with serious consequences for humanity and the sustainable development of the world [7], [8], with children under five and older adults being the most susceptible [9], [10].

According to the World Health Organization [7], in 2016 alone, 3 million deaths per year in open spaces and another 3 million in indoor spaces are attributed to air pollution. While for 2019 the same organization estimated 4.2 million premature deaths from outdoor pollution and 3.8 million deaths from domestic exposure to smoke from stoves and dirty fuels [11]. Furthermore, amidst the main types of deaths attributed to air pollution are [10]: stroke, heart disease, lung cancer.

In this sense, this work presents the sensing and analysis of environmental variables related to indoor air pollutants, which were captured through a wireless sensor network that was connected to a cloud service, where the measurements were stored. This wireless sensor network was composed of two nodes, one that captures atmospheric pollutants PM_{10} and $PM_{2.5}$ (Particulate Matter), and the other that captures carbon monoxide, temperature and humidity in the environment. This wireless sensor network was installed in a university environment in Mexico City.

II. Background

A. Air Pollution

Poor air quality means that there are particles or gases contaminating the environment in a certain period, either in open or closed spaces and in different quantities. These pollutants cause various damages to health and the environment, also contributing negatively to climate change and depletion of the ozone layer [6], [10].

It is important to highlight that at present, there is a wide diversity of factors that contribute to the emission of air pollutants: a) those of natural origin, such as erosion, volcanic activity, forest fires and biological material; b) those that are produced by human activities, especially in metropolitan areas by industry, transport and power generation; and c) those derived from activities in rural areas by agriculture and livestock. Undoubtedly, the combination of these pollutants, associated with meteorological and environmental conditions, potentiates the deterioration of air quality, which in addition to the effects on health and ecosystems, leads to a brake on the development of the economy and the advancement of social welfare due to the damage it causes in agriculture, livestock, tourism, and other areas [6].

From a health perspective, the most important air pollutant is particulate matter (PM), which has an aerodynamic diameter less than or equal to 10 micrometers (PM_{10}) and particles less than or equal to 2.5 micrometers ($PM_{2.5}$) [9]. PM_{10} is caused by the disintegration of larger particles and the incomplete combustion of fossil fuels. They also contain materials from the Earth's crust and biological materials, such as pollen, spores, viruses and bacteria. Other emitters are unpaved roads, agriculture, firewood burning, the food industry, electric power generation, among others. While $PM_{2.5}$ is a complex mixture of inorganic materials, metals, inert species and carbonous material from combustion. Among the activities that generate this pollutant are: housing combustion, electric power generation, diesel vehicles [6], [9], [12].

On the other hand, the gases that contribute to deteriorating air quality are Carbon Monoxide (CO), Carbon Dioxide (CO_2), Ozone (O_3), Nitrogen Oxides (NO_x) and Sulfide Dioxide (SO_2). Of these, CO is one of the most important pollutants generated by incomplete combustion, that is, the components of the fuel that do not fully oxidize, these are called unburned, being motor vehicles and the generation of electrical energy some of their most important generating sources [6].

B. Indoor Air Quality

Indoor air is that which is found in closed places and its circulation is limited such as homes, hotels, banks, offices, schools, hospitals, among others. A large percentage of the population spends 90% of their time inside these places or any other space, where the air quality may be contaminated [13]. This may be due to various factors such as cleaning products used, building materials, maintenance activities, electronic equipment, textiles, combustion when cooking, use of heaters and so on. Therefore, new research emerges on how to prevent poor air quality, especially in indoor spaces[13], [14], [15].

Indoor Air Quality (IAQ) is the relationship between temperature, humidity, ventilation and chemical and biological pollutants in indoor spaces, not counting industrial buildings. Poor indoor air quality causes many diseases, which can sometimes be fatal [13], [14]. The indoor pollutants considered by the World Health Organization [14] are radon, carbon monoxide, benzene, particulate matter, formaldehyde, naphthalene, nitrogen dioxide, trichloroethylene, tetrachlorethylene, polycyclic aromatic hydrocarbons, among others.

C. Health and Environmental Care

From different edges of technological development, it is important to promote environmental care with the purpose of improving the quality of life without compromising that of future generations [10]. This implies obtaining information through monitoring networks with the purpose of knowing the concentration levels of atmospheric pollutants. This information allows for air quality diagnostics and air pollution modeling [6].

Obtaining real-time information on air quality is of great importance for the control of air pollution, either to make the population aware of possible risks, as well as to eliminate or minimize the emission of pollutants. Furthermore, it is important to make predictions through valid models to identify trends in order to determine future pollutant concentrations or locate sources that originate them. Thus, the purpose is to have useful information to support decision-makers in order to prevent various levels of contamination [16], [17].

III. Literature Review

At present, there are works that used data from regulated monitoring stations, also known as certified, with which various models were implemented for the analysis and forecast of air pollution. On the other hand, there are works in which low-cost sensor networks were used for data acquisition, with which analyzes were performed to determine the reliability of their measurements, as well as the implementation of models for forecasting concentration levels of pollutants in the air.

One of these works is [18], where they installed 17 sensor nodes in playgrounds in Oslo, Norway. The purpose was the monitoring of nitrogen dioxide (NO_2) levels as an air pollutant in open spaces, in order to provide to the staff of these children's rooms with updated information to prepare an adequate plan for outdoor activities, thus reducing children's exposure to pollutants. As part of the work, to improve the precision of the data obtained by the sensors, data fusion techniques were used. With these data, detailed maps of pollutant concentrations throughout the area were generated. To compare the measurements obtained, they used captures from a nearby regulated atmospheric station, thus finding correlations with the data obtained by the sensors. As later work, the authors suggest using supervised learning techniques to reduce measurement error.

In [19] they presented an air quality monitoring system (AQM) in Qatar. For this, they used low-cost sensors, whose measurements were stored, processed and converted into forecasts of air pollutants of O_3, NO_2 and SO_2. For the forecast, they used machine learning algorithms such as Support Vector Machine (SVM), M5P Model Trees (M5P) and Artificial Neural Networks (ANN), which were evaluated by Prediction Trend Accuracy (PTA) and Root Mean Square Error (RMSE).

The results obtained showed that M5P achieved a better prediction performance. These results were later distributed via a mobile application and text messages. As future work, they suggest implementing changes in data capture for real-time forecasting.

Moreover, according to [20], the United States Environmental Protection Agency (EPA) established the Community Air Sensor Network (CAIRSENSE) project, in order to assess the feasibility of various sensor networks installed in an area of two kilometers from the suburban area of Decatur, Georgia. The pollutants measured between August 2014 and May 2015 were NO, O_3, CO, SO_2 and $PM_{2.5}$, which were compared with measurements from regulated air monitoring stations. The work showed that low-cost sensors provide variable performance, and in some cases had a high correlation. This can be a consequence of various environmental factors that contribute to the performance of the measurements obtained by the sensors. Therefore, the authors recommend testing these types of sensors in different climates in order to identify the air pollutants that predominate in a certain area or place.

In the case of Mexico, according to the Air Quality program [21], the Ministry of the Environment of Mexico City, in collaboration with the National Supercomputing Center of Barcelona (CNS), in 2017 developed a forecasting system of air quality that integrates three models: meteorological, emissions and chemical transport. The data used comes from stations of the Atmospheric Monitoring System (SIMAT). Data from global systems were used for the meteorological model, while information on the quantity and distribution of NOx, CO, SO_2, NH_3 (ammonia), $PM_{2.5}$ and volatile organic compounds were used for the emissions model. Thus, based on the Megalopolis Environmental Commission [28], if one of the stations reports 151 ozone points, then the forecast system is consulted for the next 24 hours. However, in May 2019 an atmospheric contingency was presented that questioned the performance of the Air Quality Forecasting System.

Table I summarizes the main characteristics of the analyzed works emphasizing the key points that characterize them such as author, type of monitoring, pollutants analyzed, place and type of application.

Nowadays, continuous and systematic environmental monitoring requires technological resources that hinder its implementation [18], [22]. In response, compact low cost and easy to implement sensors have been developed that allow monitoring of environmental conditions with high spatial and temporal resolution [23]. These devices, in recent years, have become an important part of active environmental monitoring with a remarkable performance [18], [24], which makes them a viable option to obtain space-time measurements with high resolution.

For this reason, obtaining real-time information on the concentrations of pollutants in the air is of great importance for controlling pollution levels and protecting people from adverse health impacts. Therefore, this work presents the sensing and analysis of environmental variables related to indoor air pollutants, captured through a wireless sensor network connected to the cloud service through ThingSpeak, which is an open source Internet of Things application to store and retrieve data over the Internet.

IV. Method

Given the purpose of capturing and analyzing environmental variables related to indoor air pollutants, as a method for this research three work stages, exploratory and applied, were defined: a) design and installation of the network of wireless sensors, b) data acquisition in the observation area, and c) analysis of air pollutants in the observation area.

A. Wireless Sensor Network

For data capture, a wireless sensor network was designed and implemented, consisting of a NodeMCU ESP8266 board; temperature, humidity, CO, PM_{10} and $PM_{2.5}$ sensors; and ThingSpeak cloud storage service. The NodeMCU board is a development kit for Internet of Things which integrates various components: ESP8266 microcontroller, ESP12E module, CP2102 serial-USB converter chip, microUSB port for power supply and programming, firmware, status led, a button for reset, flash button, among others. While the sensors used were:

- DHT11, it is a digital sensor that measures the humidity and temperature of the environment in which it is located. Humidity is measured by means of a capacitive sensor and the temperature by means of a thermistor.
- DHT22, it is a digital sensor that measures humidity and temperature, which has better accuracy than DHT11.
- MQ-7, it is a sensor to detect CO concentrations in parts per million (ppm). It is highly sensitive and fast responsive, used in industry, homes and portable detectors.
- SDS011, it is a sensor used to obtain the concentration of particles in the air between 0.3 and 10 micrometers in diameter. It has a digital output and a built-in fan that makes it stable and reliable.

These sensors were physically connected to the ESP8266 node, which in turn was wirelessly connected to Internet to create a permanent communication channel with the ThingSpeak platform. This platform allows the processing and storage of 8200 messages per day [25]; enough quantity for this project. In addition, it provides real-time data stream visualization. Fig. 1 shows an overview of wireless sensor network design.

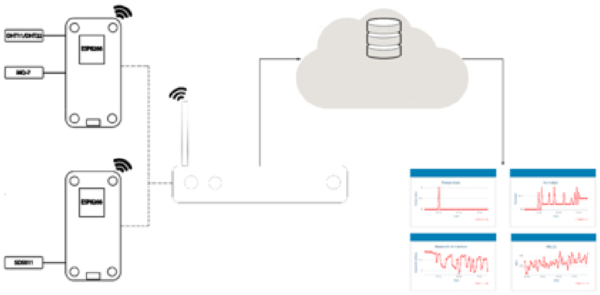


Fig. 1. Overview Diagram of Wireless Sensor Network Design

For the reading, import and export of data, a communication channel was enabled, which can be accessed through

TABLE I. RELATED WORKS

Author	Monitor type	Pollutants	Place	Application	Limitations
Castell *et al.*, 2018	17 low-cost sensor nodes	NO$_2$	Oslo, Norway	Monitoring of pollution levels	– Does not use wireless sensors. – Does not use cloud services.
Jiao *et al.*, 2016	Low cost sensor network	NO$_x$, O$_3$, CO, SO$_2$ y PM$_{2.5}$	Decatur, Georgia, EUA	Monitoring	– Does not use cloud services.
Bashir, Kadri and Rezk, 2016	Low cost sensor network	O$_3$, NO$_2$, SO$_2$	Qatar	Machine Learning (SVM, M5P, ANN)	– Does not use cloud services.
Athira, Geetha, Vinayakumar and Soman, 2018	1498 stations of the National Environmental Monitoring Center in Chine	PM$_{10}$, PM$_{2.5}$, NO$_2$, CO, O$_3$ y SO$_2$	China	Neural Networks (RNN, LSTM y GRN)	– They use regulated monitoring stations.
Air quality, 2019	Stations of the Atmospheric Monitoring System	NO$_x$, CO, SO$_2$, NH$_3$, PM$_{2.5}$ and Volatile compounds	Mexico City, Mexico	Weather and air quality forecast	– They use regulated monitoring stations.

HTTP calls, REST API and the MQTT API, or downloaded directly in the Comma-Separated Values (CSV) file format. The process begins with the capture of data through the sensors, these measurements are sent through the communication channel to the ThingSpeak platform, where the data was stored in the cloud in a structured way, that is, each column represents a captured variable and each record represents one row. Subsequently, with the data history, it is possible to make visualizations and analyzes of particular interests.

B. Data Acquisition in the Observation Area

The acquisition of data was performed in a university environment, specifically in an Artificial Intelligence Laboratory of a public university in Mexico City. The laboratory is a workspace for students and teachers, where academic activities, theoretical classes, and guided practices are held. This space has an area of $70m^2$ and a capacity for 32 people. Fig. 2 shows the workspace, which has furniture, computer equipment, and an air conditioning system.

It is important to note that in Mexico City, throughout 2019, the Atmospheric Environmental Contingencies Program (PCAA) in its phase I has been activated on three occasions and lasting from one to three days, this due to the high O$_3$ and PM$_{2.5}$ concentration levels [26].

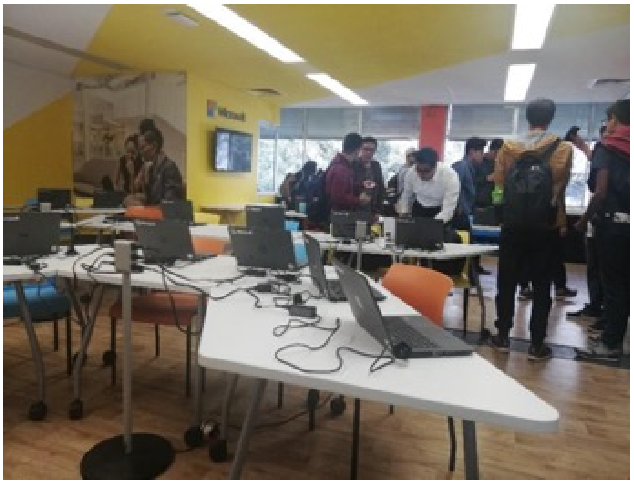


Fig. 2. Observation Space for the Acquisition of Data on Air Pollutants

In this sense, based on the implemented wireless sensor network, the pollutants CO, PM$_{10}$ and PM$_{2.5}$, and environmental variables of humidity and temperature were measured in the observation area.

- Carbon Monoxide (CO), is an odorless, colorless, and tasteless gas that is produced by the combustion of organic substances and pollutants emitted by vehicles, and on a smaller scale by industry. It is considered one of the six most criteria pollutants and harmful to health and the environment [27]. In addition, it is toxic due to its ability to combine with hemoglobin, impeding it from capturing and transporting oxygen and thus making it difficult to deliver to tissues, affecting different organs such as the heart and brain. This causes confusion, difficulty concentrating, and low reflexes [28], [29]. It also causes visual problems, reduces mental capacity, manual dexterity and making it difficult to perform complex tasks. At extremely high levels it produces severe poisoning and can lead to death [30].
- Particulate matter PM$_{10}$ is a mixture of substances in liquid or solid state that remain suspended in the atmosphere for varying periods [31]. This type of particle has 10 micrometers in diameter (PM$_{10}$) and its unit of measurement is μg/m^3 (microgram per cubic meter). They are mainly composed of materials from the Earth's crust and biological material, such as pollen, spores, viruses and bacteria. They originate mostly from the process of disintegration of larger particles. They are mainly deposited in the nose, mouth, and laryngopharynx, and can reach the thoracic region of the respiratory tractors such as the trachea, pharynx, and lungs [9], [30].
- Particulate matter PM$_{2.5}$ is composed of fine particles with diameter fewer than 2.5 micrometers, and just like PM$_{10}$, its unit of measurement is μg/m^3. The main emitters are carbon, organic carbon, biological material, which in turn includes endotoxins, bacteria, spores, allergens and pollen, and inorganic material, such as sulfates, ammonia, nitrates, transition metals, and earth metals [32]. Other sources of emission are sea salts, erosion, forest fires, volcanic activity, plant fragments, microorganisms, housing combustion,

power generation, transportation, among others [9]. They can reach alveoli and even blood, generating more serious and dangerous problems and diseases than PM_{10} [30].

- Atmospheric temperature is a magnitude that measures the thermal level of the atmosphere, in this case in degrees centigrade (°C). The human being is designed to intuitively perceive the notion of cold (lower temperature) and heat (higher temperature). Furthermore, the physical properties of matter depend on temperature.
- Relative humidity is the amount of water vapor that exists in a given environment. It is expressed as a percentage (%) and it is dependent on environmental conditions, the more water in the environment, the higher the degree of humidity. In addition, when humidity in the air cools due to temperature, it condenses and becomes liquid.

The capture period of these atmospheric variables was 15 consecutive days, from November 18 to December 2, 2019, with a capture timing scheduled at each node of one minute. The purpose was to detect possible air pollutants in the observation area, where students and teachers interact daily for academic activities. Tables II and III show extracts from the capture measurements of temperature, humidity and carbon monoxide (node 1); and PM_{10} and $PM_{2.5}$ (node 2), respectively.

TABLE II. Node data extract that captures temperature, humidity and CO

Date	Temperature (°C)	Humidity (%)	CO (ppm)
2019-11-18 00:00:12	22.0	41	0.81926
2019-11-18 00:01:14	22.0	41	0.81926
2019-11-18 00:02:15	22.0	41	0.77062
...	...	...	...
2019-12-02 23:57:32	21.6	45	1.25169
2019-12-02 23:58:34	21.7	45	1.24067
2019-12-02 23:59:35	21.7	45	1.25169

TABLE III. Node data extract that captures PM_{10} and $PM_{2.5}$

Date	PM_{10} ($\mu g/m^3$)	$PM_{2.5}$ ($\mu g/m^3$)
2019-11-18 00:00:50	52.400	32.020
2019-11-18 00:01:50	52.500	32.760
2019-11-18 00:02:51	52.440	32.889
...	...	...
2019-12-02 23:57:20	35.729	15.289
2019-12-02 23:58:21	35.739	15.500
2019-12-02 23:59:21	36.129	15.700

C. Analysis of Pollutants in the Observation Area

For the analysis of possible pollutants in the observation area an exploration of the data was carried out, mainly through the deployment of graphs in order to identify, depending on the temporality, some kind of trend, seasonality and correlation between variables. In addition, for the analysis, the permitted levels in each of the evaluated air pollutants and their risk level were taken as a reference.

- CO is measured in parts per million (ppm); 1 ppm equals 1145 $\mu g/m^3$. The average levels of CO in households without gas stoves range from 0.5 to 5 ppm [33]. NOM-021-SSA1-1993 [34] establishes that the maximum permissible value of exposure to this pollutant, for a susceptible person, is 11 ppm on average for a maximum of 8 hours once a year. However, the EPA [35], through the National Ambient Air Quality Standards (NAAQS) establishes that the exposure to this pollutant should not exceed 9 ppm during an average of 8 hours or 35 ppm for one hour once a year.
- Regarding PM_{10}, the limit values established by the Ministry of Health [9] for the exposure of a vulnerable person to this pollutant is 75 $\mu g/m^3$ as an average in 24 hours and 40 $\mu g/m^3$ as an annual average. On the other hand, the World Health Organization [36] suggests an annual average of 20 $\mu g/m^3$ and an average of 50 $\mu g/m^3$ for 24 hours. However, the EPA [35] sets the limit at 150 $\mu g/m^3$ for 24 hours without exceeding this exposure more than once a year for an average of 3 years.
- Given that, $PM_{2.5}$ causes more serious and dangerous problems and diseases than PM_{10} [9], and since they can reach alveoli and even blood the Ministry of Health [11] establishes that 45 $\mu g/m^3$ is the limit of exposure to this pollutant for an average of 24 hours for a vulnerable person and 12 $\mu g/m^3$ as an annual average.

V. Results

With respect to the CO concentration levels (Fig. 3), it was observed that during the first 10 days, values between 0.6 and 1.3 ppm were presented, which are within the range of the permitted average levels, that is, values between 0.5 and 5 ppm. However, for day 11 a peak of 2.5 ppm was observed which may be associated with an increase in pollution levels during that day. Then, it decreased reaching values below 2.0 ppm in the last days of sensing. In general, the obtained values were within the allowed average levels, that is, less than 9 ppm. Furthermore, there is no evidence of a clear cause and effect relationship of the CO measurement with respect to the passage of days.

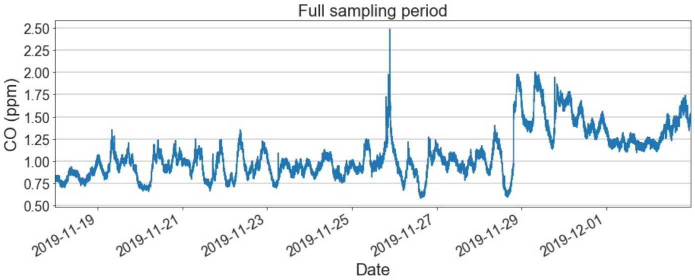


Fig. 3. CO Behavior

For the case of particulate matter of 2.5 and 10 micrometers in diameter (Fig. 4), a similar behavior was observed between $PM_{2.5}$ and PM_{10}, reaching peaks of maximum concentration in two days of measurement with 84.72 and 76.80 μg/m^3, respectively and a lower peak that coincides with the maximum CO peak of day 11, shown in Fig. 3. On the remainder of the days evaluated, concentrations were lower reaching minimum measurements of 13.75 and 33.62 μg/m^3, respectively. These values, in accordance with the parameters defined by the Ministry of Health [9] are at the permissible average levels, that is, 45 μg/m^3 for the case of PM$_{2.5}$ and 75 mg/m3 for PM$_{10}$ with exception of the two maximum concentration peaks mentioned.

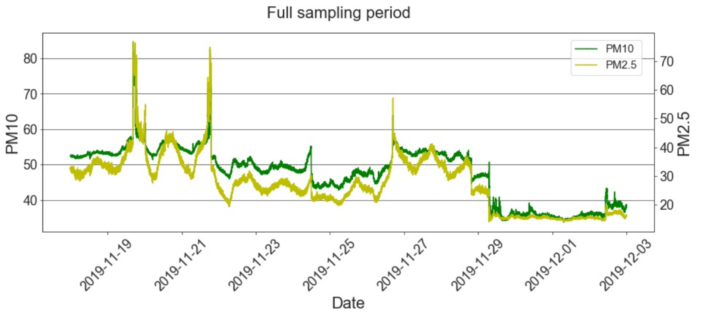


Fig. 4. PM$_{10}$ and PM$_{2.5}$ Behavior

Regarding the air quality affected by the CO concentration levels and its relationship with temperature and humidity, the presence of a certain degree of observable correlation between these variables was perceived. In the case of temperature (Fig. 5), through its seasonality, it was observed that it does not condition the increase or decrease in CO concentration levels as time elapses, except in the last part of the sampling period, whereas the temperature increased slightly there was also a slight increase in the CO concentration levels. While in the case of humidity (Fig. 6) there is clear evidence that higher percentages of water concentration in the environment were lower levels of CO contamination in the air.

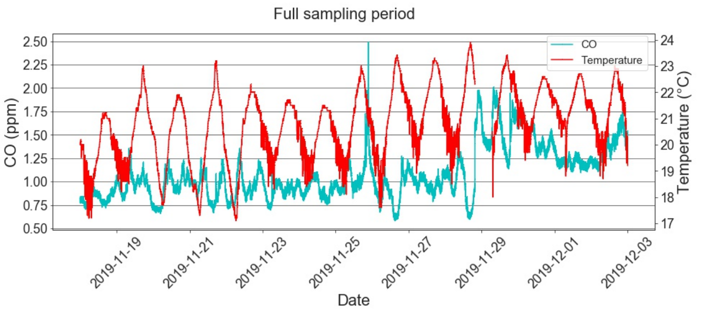


Fig. 5. CO and Temperature Relationship

About the effect that the atmospheric variables of temperature and humidity in the environment had on the PM$_{2.5}$ and PM$_{10}$ particulate matter, it was observed, Fig. 7 and 8, that these also did not condition the increase or decrease in air pollution levels either, that is, there is no observable pattern or degree of correlation between these variables. Was observed that the particulate matter suspended in the environment had varying behavior during the data capture period.

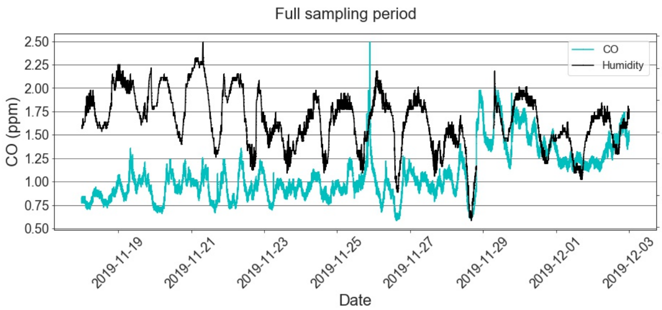


Fig. 6. CO and Humidity Relationship

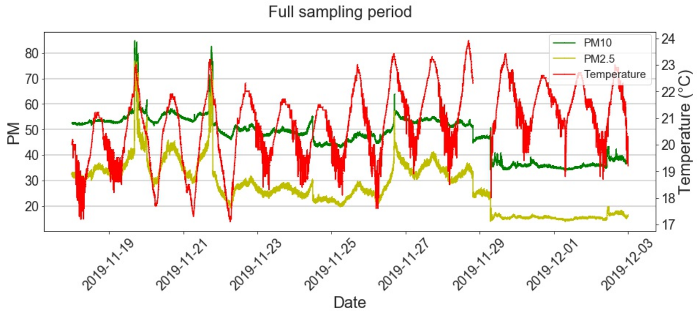


Fig. 7. Relationship of PM10 and PM2.5 with Temperature

Based on these results, regarding CO, the levels recorded to coincide with the values established by the United States Environmental Protection Agency for indoor spaces without stoves; as is the case of the laboratory under study. Given that the concentrations were lower than the limits set by the various national and international organizations that determine health standards, it can be affirmed that the presence of CO is not harmful to human health. For PM$_{10}$ and PM$_{2.5}$ particles, levels were detected at the margin or above-established limits that, coupled with the presence of outliers, could warn of errors in data collection.

Without a doubt, environmental conditions and the presence of pollutants in the air can affect the performance of people's activities and their quality of life. This presence of air pollutants, and specifically in indoor spaces, is also conditioned by other characteristics of the environment, such as lack of ventilation, days of working hours, a vehicular influx in parking lots, among others.

VI. Conclusions

Since indoor air quality can be affected by outdoor air, at present it seeks to promote environmental care from different edges of technological development in order to prevent diseases and health effects caused by air pollution and thus not compromising future generations.

One of the initial steps in dealing with exposure to air pollutants is learning about the conditions in which the environment is located. Therefore, it is important to quantify and analyze air quality as much as possible. This implies obtaining data through several monitoring systems, as support for the

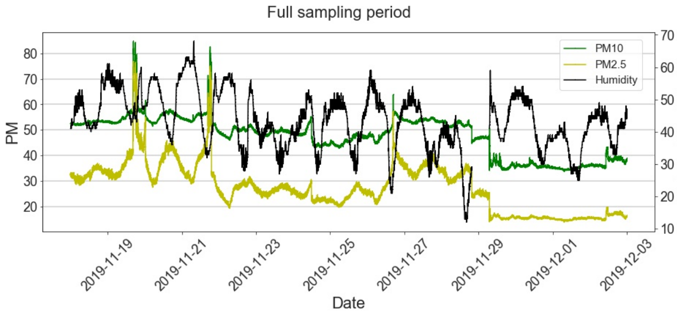


Fig. 8. Relationship of PM10 and PM2.5 with Humidity

definition of actions for the prevention, control, and mitigation of air pollutants such as CO, $PM_{2.5}$ and PM_{10}.

On the other hand, it is not enough to have regulated monitoring stations to measure pollutant concentration levels, but it is also useful to have other data capture mechanisms, for example, through wireless sensor networks with which you can measure specific observation areas, whether in closed or open spaces.

In this work, a wireless sensor network was designed and installed for monitoring air pollutants in an indoor space of a public university in Mexico City. The main feature of this sensor network is the persistence of data in the cloud and high temporal resolution. The sensor network consists of two nodes, one to measure the concentrations of particular matter PM_{10} and $PM_{2.5}$ and another to measure Carbon Monoxide and atmospheric variables of humidity and temperature in the environment.

Through the analysis, it was observed that the concentration levels of the contaminants are within the permissible parameters. In general, the level of pollution increased during working hours, meaning that the air quality was better during the first hours of the day. Likewise, it was determined that humidity and temperature are physical factors that do not condition variations in pollutants. However, no trends o seasonality were presented and a cause-effect relationship was not established over time.

It could be said, then, that low-cost wireless sensor networks obtain the environmental parameters of humidity and temperature, and levels of air pollution of CO, PM_{10} and $PM_{2.5}$ in a closed space, in a pervasive and ubiquitous manner, that is, transparent and non-intrusive way, with cloud storage, high temporal resolution, high space coverage, in real time, and which can be customized based on requirements. Compared to regulated monitoring stations they are an option that redeems economic and practical. However, knowledge is required in the programming of microcontrollers, taking into account that there are no official programming guides and mainly calibration guides on such sensors that leads to uncertainty in data reliability.

In addition, the analysis of the data allowed the understanding of the behavior of environmental variables in indoor spaces. Based on the results obtained, as future work, it is planned to develop models, based on deep learning, for the forecast of pollutants in closed places where air circulation is limited, which given high levels of pollution can be harmful to human health.

References

[1] D. Cook, J. Augusto, and V. Jakkula, "Ambient intelligence: Technologies, applications, and opportunities," *Pervasive and Mobile Computing*, vol. 5, no. 4, pp. 277–298, 2009.

[2] A. Alamri, A. Shadab, M. Hassan, M. Hossain, A. Alelaiwi, and M. Hossain, "A survey on sensor-cloud: Architecture, applications, and approaches," *International Journal of Distributed Sensor Networks*, vol. 2013, p. 18, 02 2013.

[3] H. Elazhary, "Internet of things (iot), mobile cloud, cloudlet, mobile iot, iot cloud, fog, mobile edge, and edge emerging computing paradigms: Disambiguation and research directions," *Journal of Network and Computer Applications*, vol. 128, pp. 105–140, 2019.

[4] D. Sumano, E. Domínguez, M. Lopez Trinidad, and H. Tapia-McClung, "A cloud based virtualisation protocol for wireless sensor networks," *International Journal of Sensor Networks*, vol. 31, pp. 119–132, 07 2019.

[5] A. Flammini and E. Sisinni, "Wireless sensor networking in the internet of things and cloud computing era," *Procedia Engineering*, vol. 87, pp. 672 – 679, 2014, eUROSENSORS 2014, the 28th European Conference on Solid-State Transducers. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877705814026927

[6] SEMARNAT, "Programa de gestión para mejorar la calidad del aire en el estado de veracruz de ignacio de la llave (proaire)," *Secretaría del Medio Ambiente y Recursos Naturales*, 2018, last view: 24-oct-2019.

[7] WHO, "Ambient air pollution: A global assessment of exposure and burden of disease," *World Health Organization*, 2016, last view: 25-oct-2019.

[8] A. Pruss, J. Wolf, C. Corvalán, R. Bos, and M. Neira, Eds., *Preventing disease through healthy environments. A global assessment of the burden of disease from environmental risks*, 2016, last view: 27-oct-2019.

[9] NOM-025-SSA1, *Valores límite permisibles para la concentración de partículas suspendidas PM10 y PM2.5 en el aire ambiente y criterios para su evaluación*, Secretaría de Salud, 2014, 20 de agosto de 2014.

[10] P. Ekins, P. Boileau, and J. Gupta, Eds., *Global Environment Outlook GEO-6, Healthy Planet, Healthy People*. United Kingdom: Cambridge: United Nations Environment Programme, 2019.

[11] WHO, "Air pollution," *World Health Organization*, 2019, last view: 24-oct-2019.

[12] Y. Macilla, A. Mendoza, P. Herckes, and M. Fraser, "Source apportionment of pm2.5 based on molecular organic markers in monterrey, mexico," *107th Annual Conference & Exhibition*, 2014.

[13] N. Brown, *Indoor air quality*, Cornell University, Ithaca, NY, 2019.

[14] WHO, "Who guidelines for indoor air quality: Selected pollutants," *World Health Organization*, 2010, last view: 7-nov-2019.

[15] ——, "Who guidelines for indoor air quality: Household fuel combustion," *World Health Organization*, 2014, last view: 7-nov-2019.

[16] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," *Environmental Science and Pollution Research*, vol. 23, 2016.

[17] Y. Akin, Z. Cansu, and H. Oktay, "Air pollution modelling with deep learning: A review," *Int. J. of Environmental Pollution & Environmental Modelling*, vol. 1, no. 3, pp. 58–62, 2018.

[18] N. Castell, F. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, and A. Bartonova, "Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?" *Environment International*, vol. 99, pp. 293–302, 2017.

[19] K. Bashir, A. Kadri, and E. Rezk, "Urban air pollution monitoring system with forecasting models," *IEEE Sensor Journal*, vol. 16, no. 8, 2016.

[20] W. Jiao, G. Hagler, R. Williams, R. Sharpe, R. Brown, D. Garver, R. Judge, M. Caudill, J. Rickard, M. Davis, L. Weinstock, S. Zimmer, and K. Buckle, "Community air sensor network (cairsense) project: evaluation of low-cost sensor performance in a suburban environment in the southeastern united states," *Atmospheric Measurement Techniques*, vol. 9, pp. 5281–5292, 2016.

[21] C. del Aire, “Pronóstico de calidad del aire y meteorológico para la cdmx,” http://www.aire.cdmx.gob.mx/pronostico-aire/sobre-modelo.php, Secretaría del Medio Ambiente, Ciudad de México, Tech. Rep., 2019, last view: 29-oct-2019.

[22] C. Malings, R. Tanzer, A. Hauryliuk, S. Kumar, N. Zimmerman, L. Kara, A. Presto, and R. Subramanian, “Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring,” *Atmospheric Measurement Techniques*, vol. 12, no. 2, pp. 903–920, 2019.

[23] S. Munir, M. Mayfield, D. Coca, S. Jubb, and O. Osammor, “Analysing the performance of low-cost air quality sensors, their drivers, relative benefits and calibration in cities – a case study in sheffield,” *Environmental Monitoring and Assessment*, vol. 191, no. 2, 2019.

[24] A. Lewis and P. Edwards, “). validate personal air-pollution sensors. nature,” *International Journal of Science*, vol. 535, pp. 29–31, 2016.

[25] I. The Mathworks, “Learn more about thingspeak,” The MathWorks, Inc, Tech. Rep., 2019.

[26] C. Calidad del Aire, “Activación del programa para contingencias ambientales atmosféricas (pcaa) en la zona zmvm,” http://www.aire.cdmx.gob.mx/descargas/ultima-hora/calidad-aire/pcaa/pcaa-historico-contingencias.pdf, Gobierno de la ciudad de México, Tech. Rep., 2019, last view: 31-oct-2019.

[27] E. EPA, “Criteria air pollutants,” https://www.epa.gov/criteria-air-pollutants, Environmental Protection Agency, Tech. Rep., 2018, last view: 2-dic-2019.

[28] NOM-041-SEMARNAT, *Que establece los límites máximos permisibles de emisión de gases contaminantes provenientes del escape de los vehículos automotores en circulación que usan gasolina como combustible*, Secretaría del Medio Ambiente y Recursos Naturales, 2015, 10 de junio de 2015.

[29] M. Kampas and E. Castanas, “). human health effects of air pollution,” *Environmental Pollution*, vol. 152, no. 2, pp. 362–367, 2008.

[30] CEMDA, “Recomendaciones de política pública para mejorar la calidad del aire en méxico,” https://www.cemda.org.mx/wp-content/uploads/2013/02/calidadelaire.pdf, Tech. Rep., 2013, last view: 27-nov-2019.

[31] SEMARNAT-INE, “Guía metodológica para la estimación de emisiones de pm2.5,” https://bit.ly/2UNNJmo, Secretaría del Medio Ambiente y Recursos Naturales – Instituto Nacional de Ecología, Tech. Rep., 2011, last view: 27-nov-2019.

[32] INECC-SEMARNAT, “valuación de partículas suspendidas pm2.5 en el Área metropolitana de monterrey,” Instituto Nacional de Ecología y Cambio Climático, Secretaría del Medio Ambiente y Recursos Naturales, Tech. Rep., 2015, last view: 28-nov-2019.

[33] E. EPA, “Carbon monoxide’s impact on indoor air quality,” https://www.epa.gov/indoor-air-quality-iaq/carbon-monoxides-impact-indoor-air-quality, United States Environmental Protection Agency, Tech. Rep., 2018, last view: 2-dic-2019.

[34] NOM-021-SSA1, *Criterio para evaluar la calidad del aire ambiente con respecto al monóxido de carbono (CO). Valor permisible para la concentración de monóxido de carbono (CO) en el aire ambiente como medida de protección a la salud de la población*, Secretaría de Salud, 1993, 23 de diciembre de 1994.

[35] E. EPA, “Naaqs table,” https://www.epa.gov/criteria-air-pollutants/naaqs-table, Environmental Protection Agency, EU, Tech. Rep., 2018, last view: 2-dic-2019.

[36] WHO, “Air pollution,” urlhttps://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health, Tech. Rep., 2018, last view: 27-nov-2019.

Future of the Internet of Things Emerging with Blockchain and Smart Contracts

Mir Hassan[1], Chen Jincai[2]
Wuhan National Laboratory for Optoelectronics
Huazhong University of Science and Technology
Wuhan, China

Adnan Iftekhar[3], Xiaohui Cui[4]
Key Laboratory of Aerospace Information Security
and Trusted Computing, Ministry of Education
School of Cyber Science and Engineering
Wuhan University, Wuhan, China

Abstract—**The Internet of Things (IoT) has the potential to change the way the world works from home automation to smart cities, from improved healthcare to an efficient management system in supply chains to industry 4.0 revolution. IoT is increasingly becoming an essential part of the homes and industrial automation; nevertheless, there are still many challenges that need to fix. IoT solutions are costly and complicated, while issues regarding security and privacy must be addressed with a sustainable plan. Support the growing number of connected devices; the IoT is in dire need of a reboot. Blockchain technology might be the answer. Starting as a decentralized financial solution in the form of Bitcoin, Blockchain technology has expanded to diverse areas and Information Technology applications. Blockchain technology and Smart Contracts can address the outstanding security and privacy issues that impede further development of the IoT. Blockchain is a decentralized system with no central governance, facilitates interactions, promotes new and improved transaction models, and allows autonomous coordination of the devices using enhanced encryption techniques. The primary reason for this paper is to showcase the challenges and problems we are facing with the current internet of things solutions and analyze how the use of Blockchain and Smart Contracts can help achieve a new, more robust internet of things system. Finally, we examine some of the many projects using the Internet of Things together with Blockchain and Smart Contracts, to create new solutions that are only possible by integrating these technologies.**

Keywords*—*Internet of Things (IoT); blockchain; smart contracts; peer-to-peer security

I. Introduction

The Internet of Things is a development of portable, home, and installed applications that are being associated with the web incorporating the more prominent computational abilities and information investigation to scramble important data on the internet. Numerous gadgets are presently associated with the web and, later on, several billions of gadgets. As related gadgets associate, they can turn into a canny arrangement of frameworks; When these keen gadgets and techniques for frameworks share and examine information over the cloud, they can fundamentally change our organizations, our lives, and our reality in interminable ways. They enhance therapeutic results, make better items quicker with lower advancement expenses, and make shopping progressively agreeable, or by upgrading vitality age and utilization [1]. Smart devices are monitored from every aspect of usage as they perform their work efficiently. Imagine a brilliant device, for example, a shrewd traffic camera; this camera can screen the road for an obstruct, mishaps, and climate conditions and imparts that status to a portal that joins it with information from different cameras, making a wise citywide transportation framework.

Envision that this smart traffic framework is associated with other citywide traffic frameworks. That whole astute traffic framework is related to another citywide transportation framework that gets information from their very own keen gadgets, making a much progressively extensive system of frameworks. On the off chance that a city's astute traffic framework recognizes huge clog because of a mishap, at that point, that understanding can be sent to the citywide transportation framework, that can examine the effect of the disaster on other city frameworks. Perceiving that the disaster is close to the airplane terminal and two city schools, the framework could tell those different frameworks so they can change flight and school plans. Such a keen city framework additionally breaks down and determines ideal courses around the mishap and sends directions to the computerized signage frameworks to control drivers around the disaster. That is only one case of the potential advantages that can happen when canny gadgets share bits of knowledge with different frameworks, shaping consistent extending systems of frameworks.

In the IoT, different devices independently trade relevant data, focusing on data streams to enhance our life, further obscuring the limits between the digital and the physical universes [2]. Unfortunately, the Internet of Things faces challenges to perform with performance efficiency.

In the following section, an existing Blockchain and Smart Contract technology are introduced with which we can overcome these challenges mentioned in Section 1. Then, in Section 3, we take a look on how the blockchain and smart contract technology can help take IoT forward and analyze the already existing projects and experiments that are using these two technologies side by side to make the IoT world more safe, secure, affordable and more accessible.

II. Challenges for the Future of the Internet Of Things

The development of this captivating innovation has additionally made difficulties that can't be unraveled by utilizing advances intended for the conventional web. Beating these difficulties, in any case, is a determining factor that may decide if the IoT will result in the long run win and to what degree. Some prominent challenges for the Internet of Things present in Fig. 1. IoT solutions require a high cost in the deployment of infrastructure to enhance the privacy and security of data.

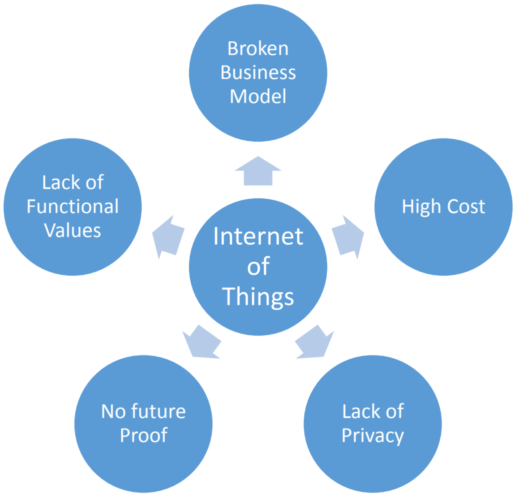


Fig. 1. Challenges for Internet of Things

Unfortunately, they have no future proof of protected data due to public authorities' functional values in their business model.

A. The Cost of Connectivity

The high cost of extensive infrastructure and maintenance associated with large server farms and centralized clouds result in prohibitively expensive IoT solutions. It is unlikely that companies will have the right profit margin because of several years of support and maintenance required for even the cheap IoT devices. This cost of serving and supporting billions of smart devices - even something as simple as maintaining servers and releasing software updates [3]. According to the survey report IoT application cost high in the U.S. [4] for developing in a different field.

B. The Security and Privacy Challenges

Most IoT solutions these days are provided by centralized authorities, whether it is government, manufacturers, or service providers. It can allow these authorities to gain unauthorized access to collect and analyze user's data. Closed source (often described as security through obscurity) approaches are being built in the current system. However, these solutions are obsolete, and the newer path of open-source (security through transparency) is required to scale IoT to the next level. Although the open-source systems may be susceptible to exploitation and accidents, it is unlikely for governments or other targeted institutions to collect unauthorized users' data.

C. The Sustainability Challenge

Among information communication technologies involved in the Smart Cities movement, IoT is believed to be an essential method, especially in the field of sustainable development. Since the application of IoT is deeply integrated into Smart Cities, which serves as a paradigm for the development of IoT technology, planners should be able to link smart Cities to the concept of sustainability. The APA Smart Cities and sustainability Initiative views Smart Cities as an extension of sustainability in that Smart Cities aims to maximize the benefits for most people with minimal costs and impacts, which reflects the very goal of sustainability.

III. Blockchain and Smart Contract

A decentralized way to deal with IoT systems administration can explain the inquiries brought up in the last segment. The appropriation of an institutionalized shared correspondence model to process the several billions of exchanges between devices will altogether lessen the expenses related to the establishment and upkeep of vast concentrated server farms and conveyed calculation and capacity needs to disseminate more than billions of devices that frame IoT systems. It keeps the disappointment of any single hub in a network from backing off the whole system. The blockchain technology is a potential candidate to organize and control it in a decentralized manner as ilustrated in Fig. 2.

A. The Blockchain Technology

Blockchain, an underlying technology powers bitcoin. It was the brainchild of a person or group of people known by the pseudonym, Satoshi Nakamoto. Blockchain technology has created the backbone of a new type of internet by allowing digital information to be distributed but not copied. The tech community is now finding other potential uses for this technology, such as connecting systems in the IoT world [5].

The network of untrusted nodes maintains the transactions of distributed ledger is called a blockchain. Every block of the blockchain contains a rundown of transactions sorted out in a Merkle tree; new blocks are added to the blockchain. Blockchains are frequently called a majority rule approach to keep transactions as they depend on accord to confirm transactions and do not require a central authority. We recognize two types of blockchains: public blockchain and private blockchain [6].

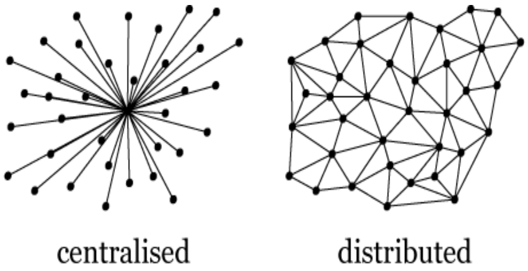


Fig. 2. Centralized and Distributed Network

B. Public and Private Blockchain

A public blockchain network is entirely open, and anyone can join and participate in the system. The network typically has an incentivizing mechanism to encourage more participants to join the network. One of the drawbacks of a public blockchain is the substantial amount of computational power that is necessary to maintain a distributed ledger on a large scale. More specifically, to achieve consensus, each node in a network must solve a complex, resource-intensive cryptographic problem called a proof of work to ensure all are in sync. Two well-known implementations of this kind of blockchains are Bitcoin and Ethereum. [7]

A private blockchain network requires an invitation and must be validated by either the network starter or by a set of rules put in place by the network starter. Businesses who set up a private blockchain will generally set up a permission network. It places restrictions on who is allowed to participate in the network, and only in certain transactions. Participants need to obtain an invitation or permission to join. A well-known implementation of this type of blockchain is IBM's Hyperledger.

C. Blocks in the Blockchain

Every block is a structure of a header and a body. The header incorporates the hash values of the previous, current, and nonce block. The block information is locked into the database utilizing the index method illustrated in Fig. 3). Since the hash values stored in each peer in the block are influenced by the benefits of the previous blocks, it is challenging to falsify and alter the registered data. In spite of data, alteration is possible if 51 percent of peers are hacked at the same time, the assault situation is convoluted.

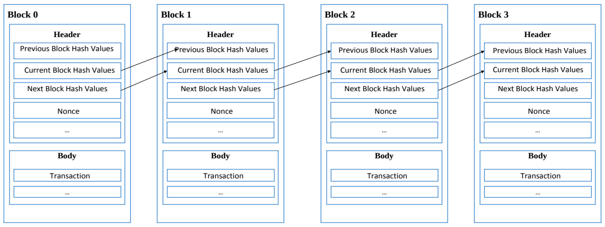


Fig. 3. Blockchain Connection Structure

D. Block Mining

Block Mining is the system that permits the blockchain to decentralize security. It anchors the bitcoin framework and empowers a system without a central authority. Miners will approve new transactions and record them on the worldwide ledger also known as Blockchain. In Bitcoin Blockchain, a block (the structure containing transactions) is mined every 10 minutes interval. Miners compete to resolve a problematic mathematical problem based on a cryptographic hash algorithm. The solution found is called the Proof-Of-Work. This proof indicates that a miner spent a lot of time and resources to solve the problem. As an incentive, Miners who tackle a cryptographic riddle are compensated with bitcoins or transaction fees [8].

E. Transaction Processing Lifecycle

The Fig. 4 is explaining the transaction processing lifecycle. It is a set of multiple processes. In the blockchain, someone initiates an operation. The transaction is then broadcast to all nodes of the blockchain. The miners validate and verify the transaction; 51 percent of the miners in the blockchain have to approve the transaction for the transaction to be added on the block. Once the new block is mined by miners, the blockchain is added to the existing Blockchain, thus making the transaction complete and permanent.

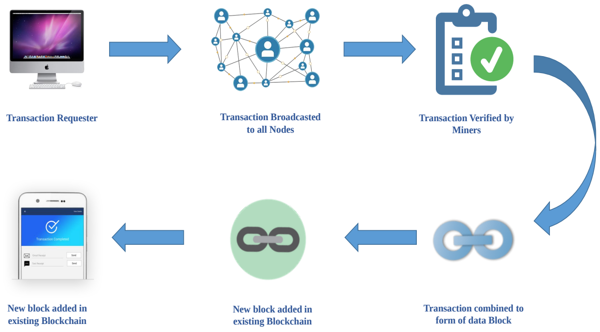

Transaction Requester
Transaction Broadcasted to all Nodes
Transaction Verified by Miners
New block added in existing Blockchain
New block added in existing Blockchain
Transaction combined to form of data Block

Fig. 4. Transaction Processing Lifecycle

F. Smart Contracts Structure

The initial release of smart contracts in Ethereum was intended to give parties who do not trust each other away to conclude an agreement, where they can be confident that the transaction will take place as they intend, and where they can verify the status of the contract or transaction at any time illustrate in Fig. 5. The initial strategy to achieve these design goals smart contract implementation did not follow the typical pattern for the development of the application. In particular, it included the logic, properties, and data in one package, substantially disintegrate the layers of business and data logic layers into a single layer, Then they were written to the blockchain. That provided immutable, deterministic execution, and the transparency required in untrusted environments.

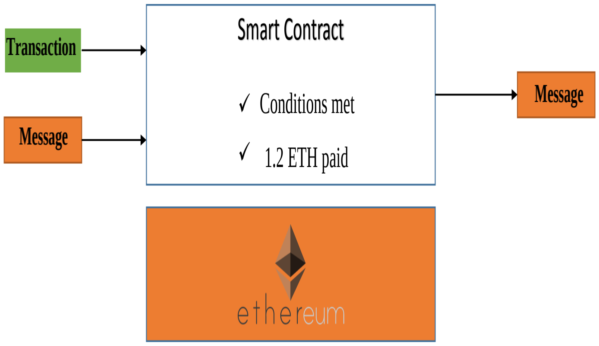


Fig. 5. Smart Contract Structure of Ethereum

G. Blockchain Benefits

Blockchain can revolutionize many industries including banking, education, voting healthcare, and supply chains. Blockchain's key benefits can be defined as:

1) Decentralization: Lack of a central data hub is one of the main reasons why blockchain is so exciting. Individual transactions in the blockchain have validity. Nodes are authorized to enforce constraints. The information in the blockchain is distributed throughout the world on different nodes so that it is near to impossible for an attacker to corrupt the stored data.

2) Efficiency: There is no involvement of intermediaries to carry out the transactions in a blockchain; it is done directly between the two parties. The digitized information allows the prompt process time of a transaction. Adding in the Smart Contracts functionality means that an action is automatically triggered when the established criteria of the contract are satisfied. This reduces the time and the cost of processing transactions.

3) Auditability: Every transaction of the blockchain is recorded in a permanent sequence. This inerasable record of transactions provides the complete audit trail for the life of an asset. This is essential especially in the cases where there is a need to verify the authenticity of an asset.

4) Traceability: With blockchain, it is possible to track the lifecycle of a product, from the manufacturer to the consumer. It is very advantageous to track goods and find out whether the products are counterfeit or real. With blockchain, it is also possible to transfer the ownership of certain goods or products from one person to another. There is no need for a paper trail. For example, a transfer of ownership of car, land, house and many more, can be done within the blockchain without the need for paperwork and intermediaries involved.

5) Transparency: Lack of transparency in a commercial organization or other business entities can sometimes lead to distrust, work delays, and loss for the company as well as consumers. By providing complete transparency regarding transaction details of the commercial relationship, further trust and stability can be developed among the parties based on openness rather than negotiation.

6) Security: The authenticity of the information stored in the blockchain can be assured because multiple nodes of the network verify the transaction using cryptography. Most significant keys to unlock the benefits of the Internet of Things is the assured information, in an autonomous process that links actions to assets.

IV. Emerging IoT with Blockchain Technology

Emerging blockchain technology can be used to improve the security of IoT applications in health care, smart cities, energy grids, public safety, education, supply chain management, education, and other application areas. Some of them use cases, where blockchain technology can benefit IoT, are discussed below:

A. Public Health – Counterfeit Drugs

The distribution and production of counterfeit drugs is a vital and pressing worldwide issue. According to the WHO, currently, in developing countries, the delivery of counterfeit drugs is may be up to 30 percent whereas 10 per cent of all drugs in the world are counterfeit. By using blockchain and IoT elements, the spread of counterfeit drugs can be controlled. Legal drug can be stamped physically at the location of production and an identifier to the stamp can be recorded into the blockchain ledger. Forged drugs can be identified, traced, and eliminated easily because they will not have any record in the blockchain system [9]

B. Smart City – Smart Homes

In today's world, a home can be a powerful computer, with a plethora of household devices related to home security systems (alarms, surveillance cameras, door locks etc.), environmental control (Air conditioner, sensors), home entertainment (audio/video equipment), and household electronics (electronic lights, refrigerators, dishwashers, washing machine, etc.) are present. All these smart home devices allow homeowners to observe and control their home from a remote site. The information coming from the network of devices is first sent to a central server and only then presented to the homeowner on their cell phones, tablets or computers. It is vital to secure this information so attackers cannot use this information. The central server or gateway is not as secure as the blockchain. Privacy and security of smart home systems can be achieved by using blockchain. The communications between devices and control information of the devices can be recorded in the distributed ledger as transactions. Cryptography and hashing functions can protect the confidentiality, integrity, and authenticity of the network of the network of IoT as well as with the addition of time-stamp and proper encapsulation, robust security is ensured.

C. Software Updating of IoT Edge Nodes

Edge devices in an IoT network have low, medium, and high levels of computing power. Due to the increase of system-on-a-chip capabilities, edge devices are becoming smarter. Some large appliances like a refrigerator can be equipped with a powerful computing system whereas a small sensor might have a little chip having adequate computing power. Thus, if all the devices in the network have some form of computing power, the functionality of the devices can be reconfigured easily. Such devices can form peer-to-peer networks and can directly communicate with each other to share IoT service functions. Device parameters are related to functionality, and device management can be downloaded and updated periodically.

D. Supply Chain Management – Smart Supply Contracts

The manufacturers create goods and services and deliver it to the retailer under the rules written in the contract. Such interactions begin with inquiries from the buyer, which leads to contract negotiation between a seller and a buyer. The shipping process begins once the contract is signed. This may involve the use of a local shipping agency, local port of exit, customs officials, a distant port of entry, carriers, customs services, a remote delivery agent, and finally the customers.

At every stage, a sequence of messages and acknowledgments is activated culminating in the customer acknowledging receipt of the shipment. Currently, trading policies on national and international trade provide payment detail process to the supplier. This adapted chain of events is well suited for using blockchain technology for smooth, verifiable, and secure supply chain management. It is possible to record and verify all documents and activities at each stage by entering transactions into or querying the appropriately distributed ledger.

Adoption of blockchain in the IoT space can change the way IoT edge devices exchange data in a trustworthy mechanizing environment and encoding transactions while safeguarding data exchanges and ensuring the security of all devices involved.

V. Case Studies of Blockchian IoT Platforms

We will explore some of the projects that are already combining Blockchain Technology and Internet of Things to develop new and viable solutions to existing problems:

A. *Slock.it*

Slock is where blockchain meets IoT. It is a decentralized platform for renting/selling your physical goods. Airbnb apartments become fully automated; parking spots can be sublet on demand, vehicles can be sold or rented without the involvement of any intermediaries. Slock.it bridges the physical world and the blockchain by making smart contracts enforceable: Slock.it has the potential to be the future infrastructure of a sharing economy [10].

B. *Filament*

Filament builds blockchain hardware, IoT and software solution. The distributed blockchain capabilities of Filament leverage open protocols so that devices can process and record transactions independently ensuring digital trust. The filament built a Blocklet Chip and new trusted application software, currently in beta, designed to communicate with multiple blockchain technologies natively. A secure distributed ledger solution is achieved through its software, and the block chip will allow corporations and enterprises to streamline the process of extracting the value of recording, monetizing data assets on the sensors themselves [11].

C. *Skuchain*

Skuchain is a blockchain technology company catered to the B2B trade =and supply chain finance market. The company aims to solve problems in the USD 18 trillion global finance market, an industry that still relies mostly on paper for many processes [12].

D. *Blockchain of Things*

Blockchain of Things created Catenis Enterprise to facilitate developers and organizations who wish to integrate Bitcoin blockchain capabilities into their devices quickly, be systems, software, machines, sensors, and other enterprise scale applications. The Catenis provides a layer that removes technological difficulties and delivers ease of use enhancements via standard web services for systems messaging and end node security. The Catenis allows organizations to rapidly leverage the bitcoin blockchain for enhanced security and reduced cost for global messaging and device communications. Clients can quickly generate a vast number of Catenis virtual devices. These virtual devices correspond to software applications and real-world physical systems in use [13].

VI. Conclusion

The Internet of Things integrated with blockchain will allow you to live in a smart home, drive smart cars and practice smart medicine. In this emerging world with technology, users connect with smart devices using secure identification and authentication, potentially public/private keys, and they define the rules of engagement, such as privacy, with other devices, rather than going along with the laws of a centralized node or intermediary. Manufacturers can transfer ownership, maintenance, access, and responsibility to a community of self-maintaining devices, future-proofing the IoT and saving infrastructure costs, replacing each equipment exactly when it hits obsolescence. Blockchain and IoT are twin technologies that can benefit from each other. They represent the most significant technological disruption since the usage of processing transaction systems in computing. The significant delipment of devices and software, it is conceivable to convey transaction processing and intelligence to all devices. Although there is some critical adaptability, challenges with the distributed systems, many researchers, institutions, individuals are working tirelessly to solve these issues and build an open source foundation for the development of this technology.

Acknowledgments

The Authors would like to acknowledge the support provided by the National Key R&D Program of China (No.2018YFC1604000).

References

[1] S. Li, L. Da Xu, and S. Zhao, "The internet of things: a survey," *Information Systems Frontiers*, vol. 17, no. 2, pp. 243–259, 2015.

[2] K. Ashton *et al.*, "That 'internet of things' thing," *RFID journal*, vol. 22, no. 7, pp. 97–114, 2009.

[3] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions," *Future generation computer systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[4] H. Singh, "Hemendra singh," Jul 2019. [Online]. Available: https://customerthink.com/how-much-does-it-cost-to-develop-an-iot-application/

[5] M. Walport *et al.*, "Distributed ledger technology: Beyond blockchain," *UK Government Office for Science*, vol. 1, 2016.

[6] M. Swan, *Blockchain*. O'Reilly Media, 2015.

[7] Z. Zheng, S. Xie, H. Dai, X. Chen, and H. Wang, "An overview of blockchain technology: Architecture, consensus, and future trends," pp. 557–564, June 2017.

[8] L. Luu, D.-H. Chu, H. Olickel, P. Saxena, and A. Hobor, "Making smart contracts smarter," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 254–269.

[9] A. Kiayias, E. Koutsoupias, M. Kyropoulou, and Y. Tselekounis, "Blockchain mining games," in *Proceedings of the 2016 ACM Conference on Economics and Computation*. ACM, 2016, pp. 365–382.

[10] "Slock.it, "enabling the economy of things" 1 january 2015. [online]. available: https://slock.it/. [accessed 2 december 2018]."

[11] "Filament, "launching your blockchain project has never been easier" 10 january 2015. [online]. available: https://filament.com/. [accessed 2 2018]."

[12] "Empower my supply chain," 10 january 2015. [online]. available: htts://skuchain.com. [accessed 2 december 2018]."

[13] "Blockchain of things, "the ultimate blockchain technology," 10 january 2015. [online]. available: https://blockchainofthings.com. [accessed december 2018]."

News Aggregator and Efficient Summarization System

Alaa Mohamed[1], Marwan Ibrahim[2], Mayar Yasser[3], Mohamed Ayman[4],
Menna Gamil[5], Walaa Hassan[6]
Faculty of Computer Science
Misr International University
Cairo, Egypt

***Abstract*—News Aggregator is simply an online software which collects new stories and events around the world from various sources all in one place. News aggregator plays a very important role in reducing time consumption, as all of the news that would be explored through more than one website will be placed only in a single location. Also, summarizing this aggregated content absolutely will save reader's time. A proposed technique used called the TextRank algorithm that showed promising results for summarization. This paper presents the main goal of this project which is developing a news aggregator able to aggregate relevant articles of a certain input keyword or key-phrase. Summarizing the relevant articles after enhancing the text to give the reader understandable and efficient summary.**

***Keywords*—News aggregator; text summarization; text enhancement; textRank algorithm**

I. Introduction

In the last few years, the world had incredible and huge growth in the rate of news that is published [1]. People live in a time full of information, data, and news [2]. So, nowadays news has an important part and position within the community. As people read the news daily to keep up with the most recent data and inputs. These data may be about technology, sports, weather, food, and celebrities or many other fields [3].

With the development of the Internet, and lot of websites that provide the same data and information, getting this has become simpler getting to it has become simpler [4]. So, users frequently discover it troublesome to decide which of these websites can provide the specified data within the most valuable and effective way [2].

The conventional commerce model of daily papers has been threatened by the internet to lessening their advertising income and by presenting new online media, such as web-only news, blogs and news aggregators [5]. The online social system is a valuable instrument for collecting, aggregating and expending the specific or common contents for different aims in a certain period of time. Daily papers are in competition with present-day online media as shown in Fig. 1. Among online media sites, feeds aggregators show up to be more significant [5] [6].

An Outsell determination (2009) [5], 57% of feeds media clients go to computerized sites, and they are too more likely to turn to an aggregator 31 % than to a daily paper location 8% or other news sites 18% [5]

Feeds aggregator combines news data, and regularly briefs it in a good format and design for the reader, from various sites, newspapers, and agencies [2]. News aggregators are frequently included in classifications such as "Websites Each Engineer Ought to Visit" [7].

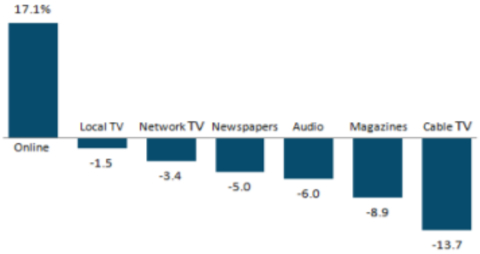


Fig. 1. Online feeds Develops Quickly contrast to Others

Despite the pros of the presence of lots of information to the people through the internet, it will get us another problem which is information overload. There will be too much information that is in front of the user and might not be his interests [8]. This problem can be solved throughout the proposed system. As News Aggregator looks like a gateway that integrates different feeds websites, it organizes feeds by subject [9]. It could be a site that takes data and news from numerous sources and displays it in a single site [10]. Which simplifies readers' search and reading time for news by gathering content based on viewing history [11]. Using news aggregation is one of the best ways to stay on top of the news and topics you want. They offer convenience and time-saving features [12].

News Aggregator system will have a major requirement which is Summarization. Summarizing articles from various sources talking about the same event then writing the content of this event on one summarized page with all perspectives [13]. Summarization is to create a shorter and smaller form of a text by protecting its meaning and the key substance of the initial content [14] [15]. So, summarization has a lot of pros like reducing the time of reading to the user and getting only useful and real news. Content summarization methods can be categorized into Extractive summaries and Abstractive summaries [16]. Extractive Summarization depends on extracting a few parts, such as phrases and sentences, from a piece of text and gather them together to form a summary. Therefore, identifying the right sentences for summarization is of the most extreme importance in an extractive method [17]. But Abstractive Summarization utilizes advanced NLP methods to generate an completely new summary. A few parts of this summary may not indeed appear within the original text [17]. In this paper, we follow extracting summarization technique which gives better output and right sentence for Summarization.

The rest of the paper is structured as follows: Section II reviews related work on News aggregator based on summarization. The proposed system is presented in Section III. Section IV presents experimental results about the summarization and performance analysis of the system. Finally, Section V concludes the paper and discusses some future work.

II. Related Work

In this section, we will discuss other news aggregator websites which based on summarization:

In [18], the authors were focusing on gathering news using matrix-based analysis (MNA) with 5 main steps as follows: the first and second steps are data gathering and extracting the article from the websites and save it in the database. The third step is grouping where they categorize the articles. The last two steps are summarization and visualization that view the important article to the user. Before the grouping step, they added the matrix-based analysis where the matrix has entity as row and the column is the states about the entities. When starting analysis, the user defines what he's looking for where MNA prepare the default values for this purpose. After that, the initialization of the matrix extends a matrix over the two required chosen dimension and look in each cell for the cell documents. The summarization phase is done according to the following steps: topic summary, cell summary and summarizing both by using TF-IDF for each cell in the matrix.

According to [11], the authors were aiming to accumulate the content from diverse websites such as articles fond moreover news headlines from blogs and websites. The belief that Rich Site Summary (RSS) gives us summarized and short data. Which is preferable for the news aggregator that they are still a successful solution for indexing articles. As reducing the time required for visiting some websites, subscribed users can quickly utilize Rich Site Summary feeds without wasting time going to numerous websites. Creating HTTP requests from the web-server is the primary step in the application and these requests are received from clients. At that point, they utilize Python to download Rich Site Summary feeds and extract articles from it according to the input. After periods of time, the web-server gives some requests to the subscribed users and in case there are any upgrades, it'll be stored and downloaded.

Author in [19] was aiming to use Rich Site Summary integrated with HTML by using wrappers (programs) and parser in order to extract the information from a specific source, then adjust them according to news categories and personalized web views via a web-based interface. They explain how they do the content scanner by using HTML and Rich Site Summary. The first step is wrapping (HTML/Rich Site Summary wrapper) which involves identifying the URL address of the new items from the source with category per the news, and the address is stored in the database as for each category pair and also combined with the corresponding wrapper. The second step of wrapping is getting information from the new items, that will be used for getting and indexing the article, for each article they obtain the first sentence and pass it to the corresponding HTML page.
According to [9], the authors were aiming to collect the news from multiple sites, newspapers, magazines, and television and merge them all in one summarized website. It progresses the goodness of results because the contents and data in it are brief and summarized. So, their work based on the Rich Site Summary fetcher for recovering Rich Site Summary reports from specific websites at a certain time. They also use web Crawling (Scraping) besides Rich Site Summary to get more accurate results. Web scraping may be a method utilized to collect huge amounts of information from websites.
From all the above mentioned researches on the news aggregator, the quality of the aggregator system is still an open area to be introduced.

III. Proposed System

In this section, the basic structure of the proposed system is described, which will be able to aggregate online news from cloud service and summarize its content to reduce user's reading time.

Fig. 2 shows the main stages of the proposed system. It consists of four main stages as described below:

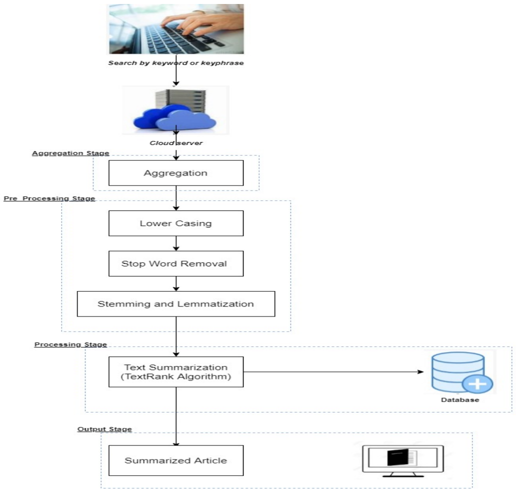


Fig. 2. Proposed System Overview

1) Aggregation stage: Aggregating related articles according to the keyword/phrase that the user enters. The articles are aggregated from different and credible sources like BBC, World Health Organization, etc.

2) Pre-processing stage: Consists of applying specific steps to the aggregated articles such as:
    a) Lowercase: used to reduce the size of the vocabulary in our data that cause multiple copies of the same word meaning.

b) Stop-word Removal: Done to remove small information of a text in order to focus on important words.

c) Lemmatization and Stemming: Remove inflection and map the word into the original/root form.

3) Processing stage: Consists of applying the summarization algorithm on the aggregated articles.
In this paper, we use the TextRank Algorithm [20] for summarizing articles which is a graph-based ranking algorithm that is used for summarization. Fig. 3 shows the steps of the textrank algorithm:

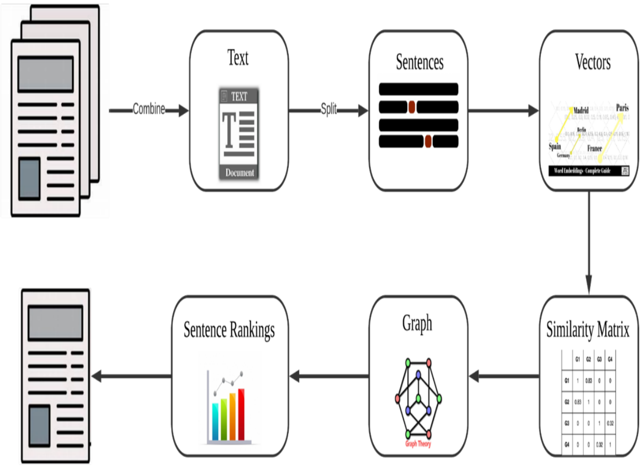


Fig. 3. Steps of TextRank algorithm

The steps can be illustrated below:

a) A set of articles are collected and combined as one one original article.

b) Tokenizing the original text as shown in Fig. 3 into sentences.

c) The Third step is vectorization. In this step, each word is represented by a vector based on the co-occurrence of a word with the others in a single sentence using Global vector algorithm (GloVe) [21]. Then we represent each sentence by a vector calculated from the mean of words vectors in a sentence.

d) In this step, we obtain a similarity matrix for all sentences using cosine similarity [22]. The similarity here refers to common content in sentences.

$$Cos\theta = \frac{\vec{a}.\vec{b}}{||\vec{a}|| * ||\vec{b}||} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i}\sqrt{\sum_{i=1}^{n} b_i}} \quad (1)$$

$$where, a.b = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n$$

is the dot product of the two vectors.

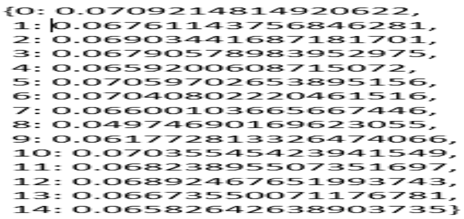
e) After obtaining the similarity matrix in the previous step, we convert it into a Graph where the edges determined by a similarity relation between them. Those edges are used to obtain the vertices weight. The importance of a sentence is based on the number of edges that represented as a score for each vertex as shown in Fig. 4 using PageRank algorithm [23]. Let the directed graph,

$$G = (V, E) \quad (2)$$

where V represents set of vertices and E represents set of edges. The vertex score$V_i$is defined as follows:

$$S(V_i) = (1 - d) + d * \sum_{j \in (V_i)} \frac{1}{|(V_j)|} S(V_j) \quad (3)$$

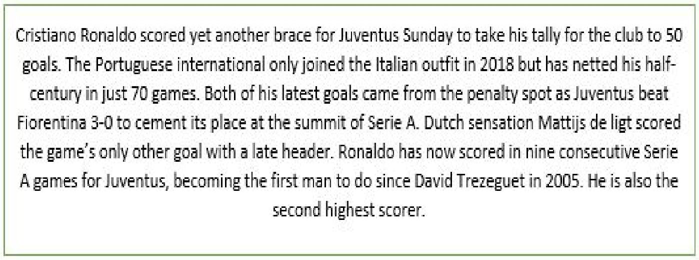
where d is a factor that it's value is between 0 and 1 and usually the value is 0.85, which represents the probability of going to another random vertex from a given vertex in the graph.



```
{0: 0.07092148149206 22,
 1: 0.06761143756846281,
 2: 0.06903441687181701,
 3: 0.06790578983952975,
 4: 0.0659200608715072,
 5: 0.07059702653895156,
 6: 0.07040802220461516,
 7: 0.06600103665667446,
 8: 0.04974690169623055,
 9: 0.061772813326474066,
 10: 0.07035545423941549,
 11: 0.06823895507351697,
 12: 0.06892467651993743,
 13: 0.06673550071176781,
 14: 0.06582642638903735}
```

Fig. 4. Sentence Scores

f) The final step is ranking the scores shown in Fig. 4 in descending order. The highest scores create the final summary shown in Fig. 5.



Cristiano Ronaldo scored yet another brace for Juventus Sunday to take his tally for the club to 50 goals. The Portuguese international only joined the Italian outfit in 2018 but has netted his half-century in just 70 games. Both of his latest goals came from the penalty spot as Juventus beat Fiorentina 3-0 to cement its place at the summit of Serie A. Dutch sensation Mattijs de ligt scored the game's only other goal with a late header. Ronaldo has now scored in nine consecutive Serie A games for Juventus, becoming the first man to do since David Trezeguet in 2005. He is also the second highest scorer.

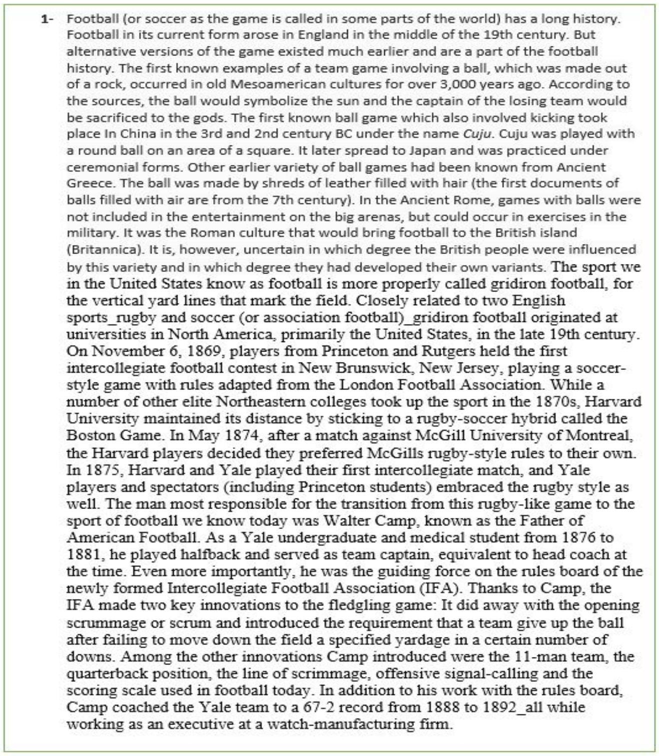
Fig. 5. Illustrative example of TextRank algorithm

4) Output stage: Summary that contains key ideas to the topic is generated.

The proposed system overview will make the user able to explore news from more than one source. The system will also give the user the main feature which is a readable summary to all of the aggregated content of the same topic. In the next section, an experiment will be discussed to compare between two summarization algorithms and to decide which one of them will fulfill the requirements of a good and accurate summary.

IV. Experimental Results

To validate the effectiveness of the proposed News Aggregator system, a set of experiments have been conducted with different keywords and key phrases. Also, the efficiency of the TextRank algorithm is tested against a common used algorithm which is Word Frequency [24]. Fig. 6 shows the original article that will be summarized, Fig. 7 and 8 are the output summary after applying summarization algorithms.

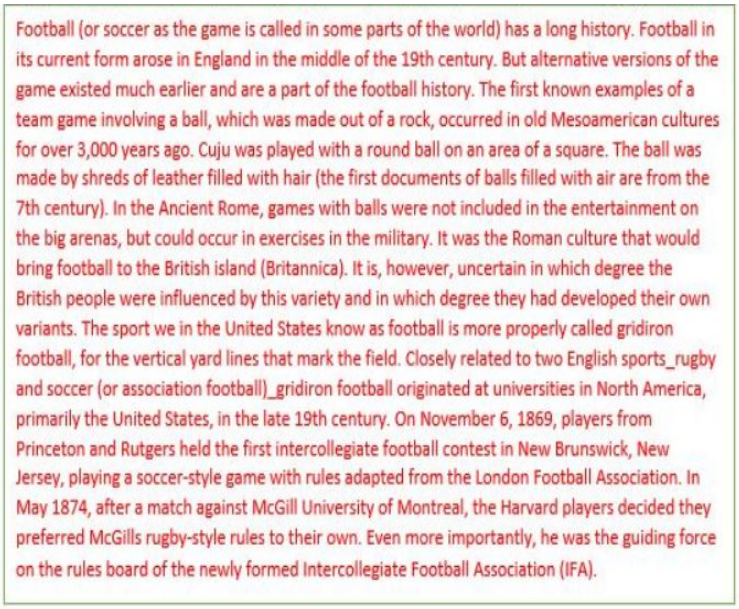


1- Football (or soccer as the game is called in some parts of the world) has a long history. Football in its current form arose in England in the middle of the 19th century. But alternative versions of the game existed much earlier and are a part of the football history. The first known examples of a team game involving a ball, which was made out of a rock, occurred in old Mesoamerican cultures for over 3,000 years ago. According to the sources, the ball would symbolize the sun and the captain of the losing team would be sacrificed to the gods. The first known ball game which also involved kicking took place In China in the 3rd and 2nd century BC under the name *Cuju*. Cuju was played with a round ball on an area of a square. It later spread to Japan and was practiced under ceremonial forms. Other earlier variety of ball games had been known from Ancient Greece. The ball was made by shreds of leather filled with hair (the first documents of balls filled with air are from the 7th century). In the Ancient Rome, games with balls were not included in the entertainment on the big arenas, but could occur in exercises in the military. It was the Roman culture that would bring football to the British island (Britannica). It is, however, uncertain in which degree the British people were influenced by this variety and in which degree they had developed their own variants. The sport we in the United States know as football is more properly called gridiron football, for the vertical yard lines that mark the field. Closely related to two English sports_rugby and soccer (or association football)_gridiron football originated at universities in North America, primarily the United States, in the late 19th century. On November 6, 1869, players from Princeton and Rutgers held the first intercollegiate football contest in New Brunswick, New Jersey, playing a soccer-style game with rules adapted from the London Football Association. While a number of other elite Northeastern colleges took up the sport in the 1870s, Harvard University maintained its distance by sticking to a rugby-soccer hybrid called the Boston Game. In May 1874, after a match against McGill University of Montreal, the Harvard players decided they preferred McGills rugby-style rules to their own. In 1875, Harvard and Yale played their first intercollegiate match, and Yale players and spectators (including Princeton students) embraced the rugby style as well. The man most responsible for the transition from this rugby-like game to the sport of football we know today was Walter Camp, known as the Father of American Football. As a Yale undergraduate and medical student from 1876 to 1881, he played halfback and served as team captain, equivalent to head coach at the time. Even more importantly, he was the guiding force on the rules board of the newly formed Intercollegiate Football Association (IFA). Thanks to Camp, the IFA made two key innovations to the fledgling game: It did away with the opening scrummage or scrum and introduced the requirement that a team give up the ball after failing to move down the field a specified yardage in a certain number of downs. Among the other innovations Camp introduced were the 11-man team, the quarterback position, the line of scrimmage, offensive signal-calling and the scoring scale used in football today. In addition to his work with the rules board, Camp coached the Yale team to a 67-2 record from 1888 to 1892_all while working as an executive at a watch-manufacturing firm.

Fig. 6. Original Article

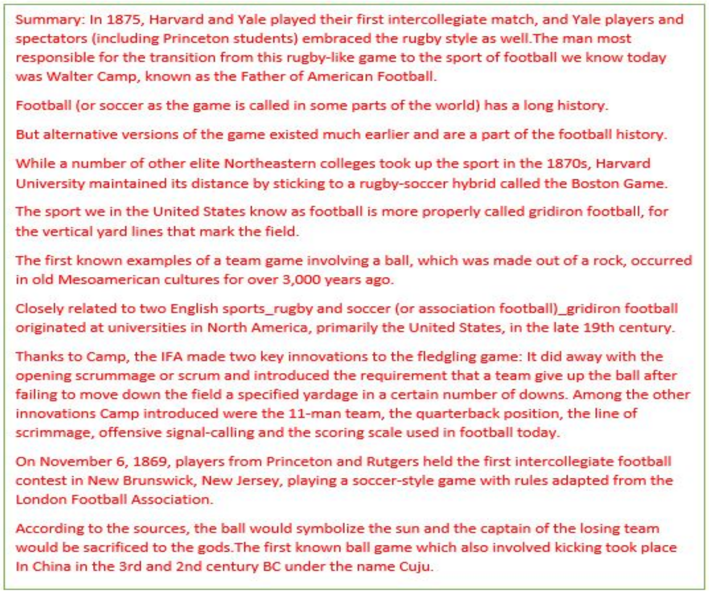
"Word Frequency" algorithm depends on more than one factor to summarize the input text as:

- The existence of the frequency table is the first step towards executing the algorithm, then every sentence will be tokenized.
- After tokenizing, we will have separated sentences, scoring every sentence will be the next step, which its formula is a division of every non-stop word in the sentence by the total number of words in the sentence.
- Getting average score of the sentences is the next step. A comparison will be done between every sentence and this average score, if the score sentence is larger, then this sentence will be considered as a summarized part of the input article.



Football (or soccer as the game is called in some parts of the world) has a long history. Football in its current form arose in England in the middle of the 19th century. But alternative versions of the game existed much earlier and are a part of the football history. The first known examples of a team game involving a ball, which was made out of a rock, occurred in old Mesoamerican cultures for over 3,000 years ago. Cuju was played with a round ball on an area of a square. The ball was made by shreds of leather filled with hair (the first documents of balls filled with air are from the 7th century). In the Ancient Rome, games with balls were not included in the entertainment on the big arenas, but could occur in exercises in the military. It was the Roman culture that would bring football to the British island (Britannica). It is, however, uncertain in which degree the British people were influenced by this variety and in which degree they had developed their own variants. The sport we in the United States know as football is more properly called gridiron football, for the vertical yard lines that mark the field. Closely related to two English sports_rugby and soccer (or association football)_gridiron football originated at universities in North America, primarily the United States, in the late 19th century. On November 6, 1869, players from Princeton and Rutgers held the first intercollegiate football contest in New Brunswick, New Jersey, playing a soccer-style game with rules adapted from the London Football Association. In May 1874, after a match against McGill University of Montreal, the Harvard players decided they preferred McGills rugby-style rules to their own. Even more importantly, he was the guiding force on the rules board of the newly formed Intercollegiate Football Association (IFA).

Fig. 7. Summary using Word Frequency



Summary: In 1875, Harvard and Yale played their first intercollegiate match, and Yale players and spectators (including Princeton students) embraced the rugby style as well.The man most responsible for the transition from this rugby-like game to the sport of football we know today was Walter Camp, known as the Father of American Football.

Football (or soccer as the game is called in some parts of the world) has a long history.

But alternative versions of the game existed much earlier and are a part of the football history.

While a number of other elite Northeastern colleges took up the sport in the 1870s, Harvard University maintained its distance by sticking to a rugby-soccer hybrid called the Boston Game.

The sport we in the United States know as football is more properly called gridiron football, for the vertical yard lines that mark the field.

The first known examples of a team game involving a ball, which was made out of a rock, occurred in old Mesoamerican cultures for over 3,000 years ago.

Closely related to two English sports_rugby and soccer (or association football)_gridiron football originated at universities in North America, primarily the United States, in the late 19th century.

Thanks to Camp, the IFA made two key innovations to the fledgling game: It did away with the opening scrummage or scrum and introduced the requirement that a team give up the ball after failing to move down the field a specified yardage in a certain number of downs. Among the other innovations Camp introduced were the 11-man team, the quarterback position, the line of scrimmage, offensive signal-calling and the scoring scale used in football today.

On November 6, 1869, players from Princeton and Rutgers held the first intercollegiate football contest in New Brunswick, New Jersey, playing a soccer-style game with rules adapted from the London Football Association.

According to the sources, the ball would symbolize the sun and the captain of the losing team would be sacrificed to the gods.The first known ball game which also involved kicking took place In China in the 3rd and 2nd century BC under the name Cuju.

Fig. 8. Summary using TextRank

In this paper, we used a summary evaluation tool named 'Rouge" for our summarization comparisons between TextRank and WordFrequency algorithms, it turns out that this algorithm has drawbacks. Drawbacks of Rouge tool is that its execution depends on the permanent existence of an expert one who knows the actual rules of summarization. Rouge executes by comparing the results of this expert to the system results and his existence might not be always available for us to have his/her consultation. This is the reason to find another evaluation criteria which was applying a survey with our social network, asking to rate each 2 summaries out of 5 which were implemented by 2 different algorithms. 1 refers to a very bad summary and 5 refers to an excellent one.

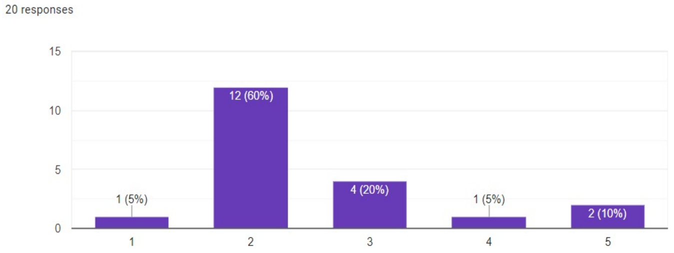


Fig. 9. Word Frequency Chart

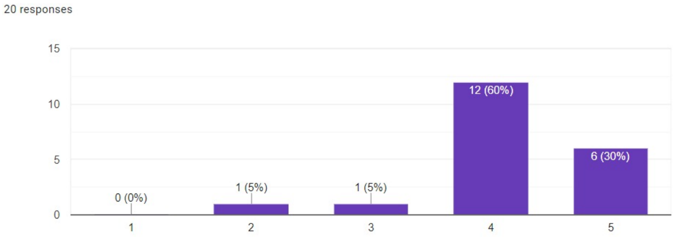


Fig. 10. TextRank Chart

TextRank algorithm was rated as 4 out of 5 from 60% from people who read it and 30% were rating 5 out of 5 as a summary as shown in Fig. 10. While Word Frequency algorithm was having bad rates from the reader as 60% were rating its summary as 2 out of 5 which is of course a bad percentage as shown in Fig. 9. These results ensure that TextRank has more efficient results in summarization.

To further emphasize that, the output summary from the TextRank algorithm was revised by an expert besides the normal readers rating. The feedback considers that the output summary fulfills all the requirements of a good summary such as:

- The length is about 10% of the original.
- Short paragraphs that contain the key ideas and to the point.
- Could be read and clearly understood without referring to the original article.
- Used the appropriate language just like that used in the original article.

V. Discussion

The results of our proposed system presented a very good and understandable summary, as this information was ensured by an expert in this type of fields. TextRank summary was more acceptable in the survey results according to readers perspective and that's an indication that user is more comfortable with reading the summary after applying the experiment. Also, the system fully works online so any type of aggregated articles will be on the spot and will be chosen from trusted and determined sources.

VI. Conclusion

After testing two summarization algorithms, TextRank algorithm was the chosen approach to be applied in the summarization system over Word Frequency algorithm. The reason behind applying TextRank algorithm was simply that TextRank gives more efficient summary for the reader. The system will generate output summary from online sources that contains key ideas to the certain article topic and it could be understood without referring to the original article.

Acknowledgment

We would like to express our special thanks to our advisor Dr. Walaa Hassan for supporting us and for her patience and motivation for our project. As her guidance helped us in writing this paper. Also, we would like to thank and appreciate the efforts of Dr. Sama Dawood (Associate Professor of Faculty of Alsun at Misr International University) for helping us in our summarization phase by defining which summary algorithm has the accurate output.

References

[1] S. Bergamaschi, F. Guerra, M. Orsini, C. Sartori, and M. Vincini, "Relevant news: a semantic news feed aggregator," in *Semantic Web Applications and Perspectives*, vol. 314. Giovanni Semeraro, Eugenio Di Sciascio, Christian Morbidoni, Heiko Stoemer, 2007, pp. 150–159.

[2] S. Chowdhury and M. Landoni, "News aggregator services: user expectations and experience," *Online Information Review*, vol. 30, no. 2, pp. 100–115, 2006.

[3] R. Bahana, R. Adinugroho, F. L. Gaol, A. Trisetyarso, B. S. Abbas, and W. Suparta, "Web crawler and back-end for news aggregator system (noox project)," in *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*. IEEE, 2017, pp. 56–61.

[4] G. Paliouras, M. Alexandros, C. Ntoutsis, A. Alexopoulos, and C. Skourlas, "Pns: Personalized multi-source news delivery," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2006, pp. 1152–1161.

[5] D.-S. Jeon and N. N. Esfahani, "News aggregators and competition among newspapers in the internet (preliminary and incomplete)," 2012.

[6] N. Diakopoulos, M. De Choudhury, and M. Naaman, "Finding and assessing social media information sources in the context of journalism," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2012, pp. 2451–2460.

[7] M. Aniche, C. Treude, I. Steinmacher, I. Wiese, G. Pinto, M.-A. Storey, and M. A. Gerosa, "How modern news aggregators help development communities shape and share knowledge," in *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*. IEEE, 2018, pp. 499–510.

[8] K. Lerman, "Social information processing in news aggregation," *IEEE Internet Computing*, vol. 11, no. 6, pp. 16–28, 2007.

[9] K. Sundaramoorthy, R. Durga, and S. Nagadarshini, "Newsone—an aggregation system for news using web scraping method," in *2017 International Conference on Technical Advancements in Computers and Communications (ICTACC)*. IEEE, 2017, pp. 136–140.

[10] K. A. Isbell, "The rise of the news aggregator: Legal implications and best practices," *Berkman Center Research Publication*, no. 2010-10, 2010.

[11] C. Grozea, D.-C. Cercel, C. Onose, and S. Trausan-Matu, "Atlas: News aggregation service," in *2017 16th RoEduNet Conference: Networking in Education and Research (RoEduNet)*. IEEE, 2017, pp. 1–6.

[12] O. Oechslein, M. Haim, A. Graefe, T. Hess, H.-B. Brosius, and A. Koslow, "The digitization of news aggregation: Experimental evidence on intention to use and willingness to pay for personalized news aggregators," in *2015 48th Hawaii International Conference on System Sciences*. IEEE, 2015, pp. 4181–4190.

[13] R. Barzilay and M. Elhadad, “Using lexical chains for text summarization,” *Advances in automatic text summarization*, pp. 111–121, 1999.

[14] G. Rossiello, P. Basile, and G. Semeraro, “Centroid-based text summarization through compositionality of word embeddings,” in *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, 2017, pp. 12–21.

[15] Y. Li and B. Merialdo, “Multi-video summarization based on video-mmr,” in *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10.* IEEE, 2010, pp. 1–4.

[16] I. Mani, *Automatic summarization*. John Benjamins Publishing, 2001, vol. 3.

[17] O. Tas and F. Kiyani, “A survey automatic text summarization,” *PressAcademia Procedia*, vol. 5, no. 1, pp. 205–213, 2007.

[18] F. Hamborg, N. Meuschke, and B. Gipp, “Matrix-based news aggregation: exploring different news perspectives,” in *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*. IEEE Press, 2017, pp. 69–78.

[19] G. Paliouras, A. Mouzakidis, V. Moustakas, and C. Skourlas, *PNS: A Personalized News Aggregator on the Web*, 01 1970, vol. 104, pp. 175–197.

[20] B. Balcerzak, W. Jaworski, and A. Wierzbicki, “Application of textrank algorithm for credibility assessment,” in *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, vol. 1. IEEE, 2014, pp. 451–454.

[21] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[22] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, “Cosine similarity to determine similarity measure: Study case in online essay assessment,” in *2016 4th International Conference on Cyber and IT Service Management*. IEEE, 2016, pp. 1–6.

[23] P. Chen, H. Xie, S. Maslov, and S. Redner, “Finding scientific gems with google’s pagerank algorithm,” *Journal of Informetrics*, vol. 1, no. 1, pp. 8–15, 2007.

[24] S. Liu, K. Huang, and J. Chai, “Research of news text with word frequency statistics and user information,” in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*. IEEE, 2017, pp. 2633–2637.

Acoustic Frequency Optimization for Underwater Wireless Sensor Network

Emad Felemban
Department of Computer Engineering
Umm Al-Qura University
Makkah, Saudi Arabia

***Abstract*—In recent years, research in Underwater Wireless Sensor Network (UWSN) was the interest of many research groups as it can be used for many important applications such as disaster management, marine environment monitoring, fish farming, and military surveillance. There are many challenges in underwater acoustic communication: strong signal attenuation, limited bandwidth, long propagation delay, high transmission loss, and energy consumption. In this paper, we present a simple flow of mathematical models for the underwater acoustic channel for the underwater acoustic communication channel. We also investigate the influence of different parameters governing the communication channel's performance, such as temperature and wind speed. We also show the importance of selecting the optimal communication frequency to increase communication SNR. We implemented the mathematical model in MATLAB and made it available online for other researchers. We found out that selecting the optimal frequency is very crucial when wind speed is high.**

***Keywords*—Underwater Wireless Sensor Network (UWSN); acoustic signal; mathematical modeling; optimization; noise level; optimal frequency**

I. Introduction

Recent advances in technologies have created many new opportunities to explore underwater resources, which covers about 70% of the planet earth. Unlike terrestrial wireless sensor networks that rely on radio waves for data exchange, UWSN needs a different approach with far more challenges. Wireless communication in an underwater environment can depend on acoustic waves or optical signals to form a communication network. Like terrestrial WSN, a UWSN is a wireless sensor network that works in an underwater environment to collect data, e.g., temperature, pressure, conductivity, turbidity, and dissolved pollutants seldom to provide some control over submerged devices. The main goal is to collect data precisely in a time-efficient and energy-efficient manner and transmit them to a sink node. The only difference, and challenge at the same time, is that RF signals do not work in an underwater environment, requiring the use of another type of signals to transmit data, namely, acoustic signals to provide wireless connectivity.

Underwater Wireless Sensor Networks has many practical applications. In [3], the authors provided a survey on underwater acoustic sensor network applications that have been suggested and studied in the literature for monitoring and controlling. Authors in [4] reviewed recent applications of UWSNs and discussed possible challenges on the implementation of UWSNs. A comprehensive survey is provided in the latest developments in UWSN in [5]. The underwater applications can be classified into five main classes: environmental monitoring, disaster monitoring, military operations, navigation infrastructure, and sports activities. Many of the challenges and opportunities faced by recent deployments of UWSN were also discussed.

UWSN faces lots of challenges and problems that have been discussed thoroughly in [6] and [8]. They include real-time propagation delay, multipath fading, limited battery, bandwidth constraints of communication channels, and high path loss due to noise. In addition, UWSN Node placement in the third dimension, i.e., depth, significantly affects the transmission path loss and operational energy consumption. As a result, transmission loss is also considerably affected by the characterizes of the water body, such as salinity, temperature, and acidity.

Many commercial low-energy underwater acoustic modem is available nowadays to fit UWSN deployment, such as [2]. Usually, acoustic modems come with a range of acoustic operational frequencies, between 15kHz and 30kHz. The selection of the optimal transmission frequency should be run-time adjustable during the operation to achieve optimal communication performance.

In this paper, we present a simple and clear mathematical model that can be used as a mathematical basis for an Acoustic Transmission Frequency Optimizer (ATFO) module, as shown in Fig. 1. The AFTO module will read ambient environmental conditions such as temperature and wind speed from its sensor readings; It will then compute the optimized transmission acoustic frequency based on the mathematical model that will be described later. We assume that the direct sink Node will be responsible for setting the transmission frequency, share it, and synchronize operation with all other nodes. This paper will only focus on how to select the optimal frequency. Although many articles in the literature provided similar mathematical modeling, this paper offers a cleaner version with a shareable source code provided for other researchers to utilize.

In this paper, we assume that a UWSN is being deployed for an arbitrary underwater application, as mentioned in previous surveys. We consider a two-dimensional flat network meaning that all nodes in this network are approximately placed in a plain, including the direct sink node, as shown in Fig. 2. We also assume that the direct sensor node is the only node connected with the ground sensor sink node on the sea level. Based on the application and the deployment environment, the depth of the network is decided during the operation. We assume a fixed network setup i.e., no mobile

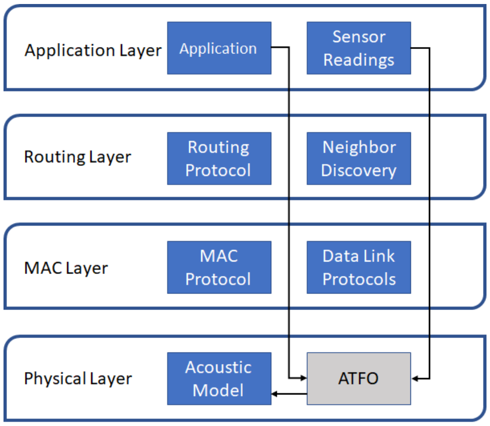


Fig. 1. Underwater Sensor Nodes Networking Layers with AFTO Module

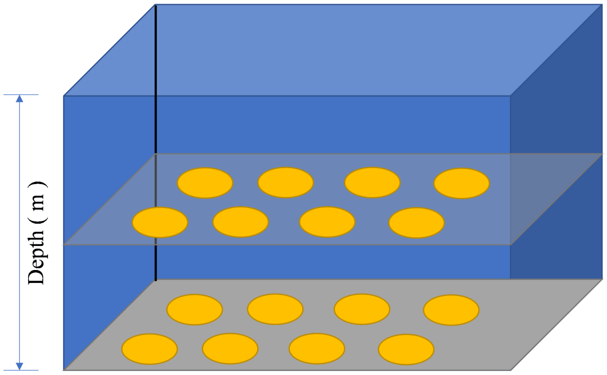


Fig. 2. UWSN Deployment at Various Depths

nodes are considered. We assume that the user can adjust the frequency of the acoustic modems used in UWSN nodes.

II. Related Work

Mathematical modeling of the acoustic channel in underwater communication has been studied widely in the literature. Sozer [19] provided a comprehensive overview of many aspects of Underwater Acoustic Networks, including a summarized form of the mathematical modeling part. In [20], a comprehensive tutorial about channel characterization and properties were introduced, including a detailed graph of noise factors affecting link quality. Another higher view of the mathematical modeling concepts of acoustic channels was introduced by [1].

Previous papers showed that transmission loss in underwater communication systems consists of two main parts Absorption due to water body and noise due to external noise factors. Over the years, three main approximations for the absorption coefficient were introduced, namely Fisher [14], Ainslie [13], and Thorp's [12], which have been used in most underwater acoustic channel modeling literature. Noise sources were also characterized and simplified in many articles in the literature, such as [1], [20]. The paper [21] provided an experimental study to analyze noise factors affecting underwater channels.

Developed mathematical models in the literature have been utilized for different purposes. The authors in [7] have provided a detailed mathematical analysis to find the relation between ambient water conditions and transmission loss. In [9] and [10], the authors provided insights about energy-delay and the energy-hops tradeoff in UWSN. In [18], the energy consumption analysis was provided using mathematical models. The distortion analysis of interference or hindrance from other sensors in the network was evaluated by [11].

The author of [15] used the acoustic channel modeling to find out the relation between link capacity and distance. In [16], the authors provided a detailed analysis of noise affecting the underwater communication profile. While in [17], the combined effect of depth and temperature on available capacity was studied.

One concern about most of the mathematical modelings efforts presented in the literature is ambiguity in certain points of the flow. In particular, we found it very difficult to regenerate similar graphs presented in some papers. We can summarize the causes of this problem into the following points:

- Different approximations for certain parameters. For example, in the literature, there are at least three different approximations for the absorption coefficient that are sometimes being used without proper addressing or referencing, making it very difficult for new researchers to know the difference.
- Importance of Units and Scale. Some equations require input parameters to be in certain units (K Meters vs. Meters), and different scales (Log vs. Linear), which are also, sometimes not very clearly mentioned in the model presented.
- Source Code Availability. Authors of literature assume that new researchers can easily construct or build a direct implementation of the mathematical models presented. These simple tasks took a fairly long time due to the first two points than expected to code and to regenerate similar graphs presented in the literature.

In this paper, we did our best to avoid these concerns. We have provided a clear and concise step-by-step model flow. A table listing all parameters with proper unit and scale is provided. Finally, the source code of the developed model, along with generating graphs, is available in [22].

III. Acoustic Channel Modeling

Practically, it is very well understood that the underwater Acoustic channel is a very challenging media to establish any communication. These challenges can be summarized as follows:

- Bandwidth Limitation: Acoustic signals operate at very low frequency, limiting the available communication band to a minimum. Typical underwater acoustic hydrophone or modems operates in the range of 15 kHz to 40 kHz [2].
- Noise Level and Sources, there are multiple sources of noise in water bodies that degrades the quality of the acoustic signal. Noise intensity measured in Power

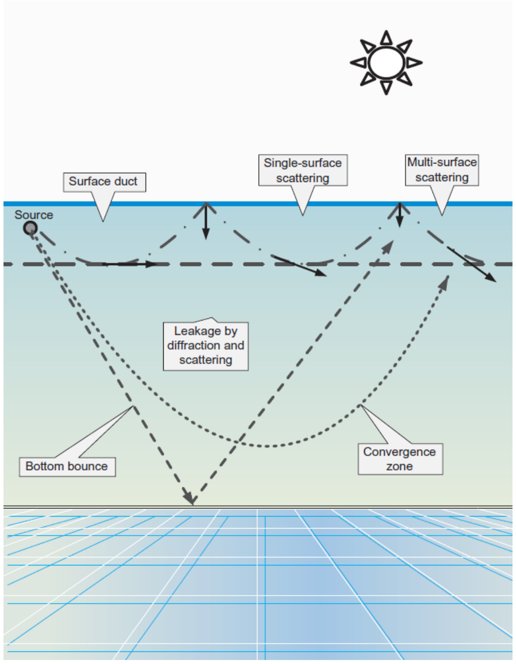


Fig. 3. Acoustic Signal Reflection and Bending Formation from [1]

Spectrum Density with unit dB relative to micro Pascal degrades as frequency increases, as shown in Fig. 5.

- Acoustic Signal Speed and Propagation, Acoustic signals are very slow, $1500\ m/s$. This fact emphasizes the propagation delay, which is usually neglected in the case of a terrestrial wireless network. The high propagation delay also magnifies the multipath problem of acoustic signal radiation. Also, acoustic signals in water bodies have a special form of bending and refraction, shown in Fig. 3, making the multipath problem even more challenging.
- Attenuation Level, Water bodies have more mass than air, making signal propagation through that body more difficult. Acoustic signal suffers from spreading and absorption in the water body. As a rule of thumb and as shown in Fig. 4, attenuation levels increases as frequency increases.
- Affecting parameters; although the frequency is the dominant factor for underwater acoustic signal progation, it still suffers from multiple other factors that have a complicated combined effect. A summarized list of all factors is shown in Table I.

To establish a wireless communication link between two nodes, the received power at the destination node should be higher than a certain threshold called rx Sensitivity Level rx_{Level}. This rule is true regardless of channel and type of carrier wave, i.e., RF vs. Acoustic. Mathematically, this condition can be formulated as

$$rx_{power} \geq rx_{level} \tag{1}$$

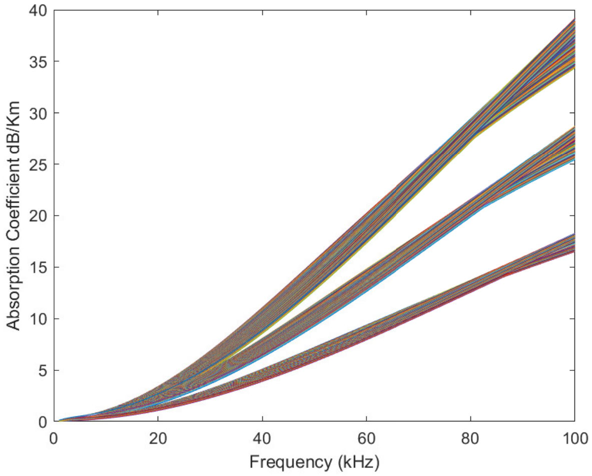


Fig. 4. Absorption Coefficient, α [dB/km] for Different Combination of Input Variables as shown in Table II

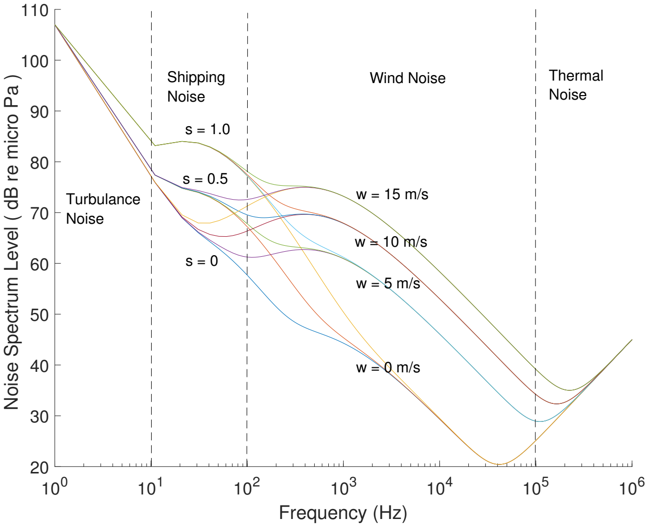


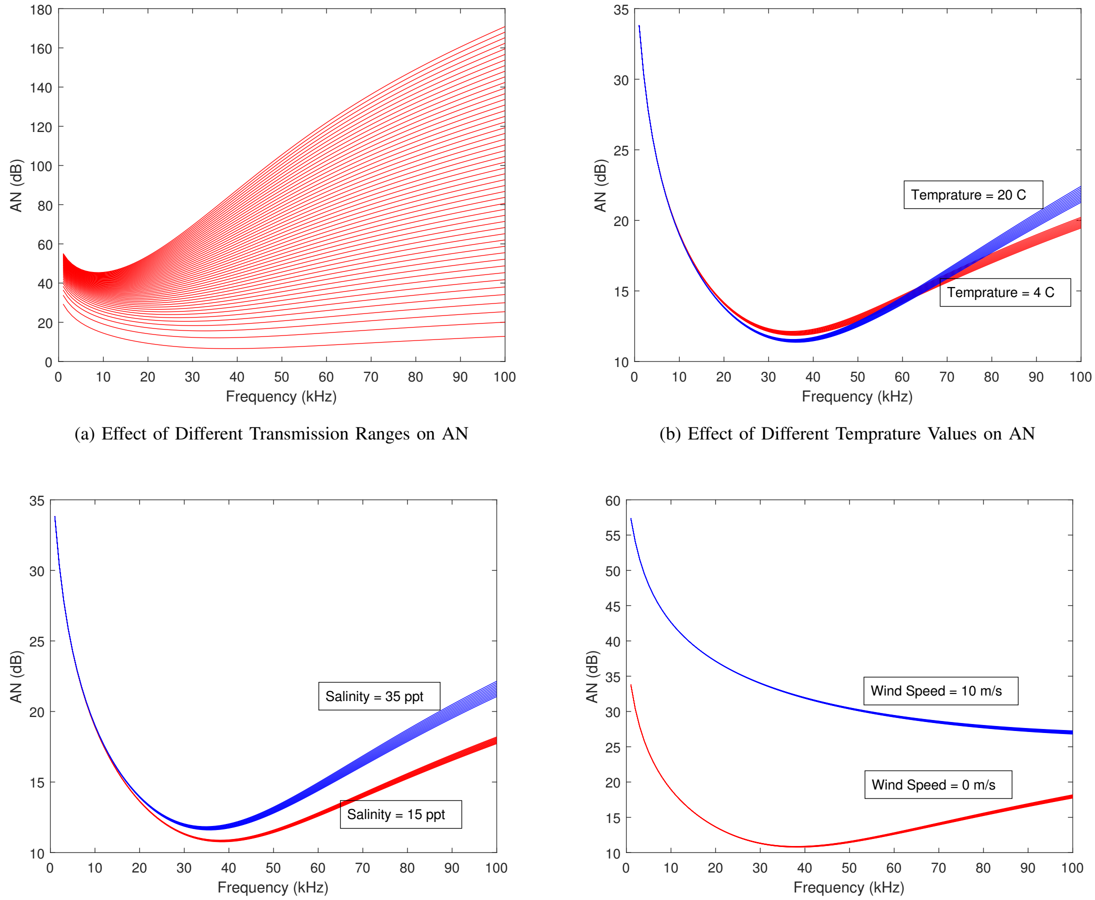
Fig. 5. Noise Loss Spectrum Level (dB re 1 μ Pascal)

where rx_{power} is the reception power level measured at the destination node. Using dB to simplify calculations, rx_{power} can be calculated as

$$rx_{power} = tx_{power} - A(f) - N(f) \tag{2}$$

Where $A(f)$ is the signal loss due to attenuation, and $N(f)$ is the loss due to Noise. The signal attenuation loss $A(f)$ in dB given in Equation 3[1] composed of two losses namely, spreading and absorption. The spreading loss is due to the geometric spreading of signal propagation it is a function of transmission range r and the spreading factor κ.For our

[1] One should note the scale and units of parameters plugged into such equations, please refer to table I



t

(a) Effect of Different Transmission Ranges on AN

(b) Effect of Different Temprature Values on AN

(c) Effect of Different Salinity Levels on AN

(d) Effect of Different Wind Speed Values on AN

Fig. 6. Effect of Different Single Parameter Values on AN

calculations and graphs in this paper, we will always use $\kappa = 1.5$.

$$10log\ A(f) = \kappa.10\ log\ r + r.10log\ \alpha(f) \tag{3}$$

The absorption loss is a function of transmission range r and absorption coefficient α, which describes water body capability to absorb the energy from the acoustic signal and convert it into heat. A higher absorption coefficient means a higher dB loss from the acoustic signal. The absorption coefficient α value is dominated by frequency but also temperature, pH level, depth level, water salinity can affect its value. There are many models that approximate the absorption coefficient empirically, such as Thorp's model and Fisher models [12], [14]. However, in this paper, we will use an approximation suggested by Ainslie and McColm [13] presented in the following formula

$$\alpha = \gamma_1 \frac{f_1 f^2}{f_1^2 + f^2} + \gamma_2 \frac{f_2 f^2}{f_2^2 + f^2} + \gamma_3 f^2 \tag{4}$$

where,

$$\begin{aligned}
f_1 &= 0.78\sqrt{\frac{s}{35}}\exp^{\frac{t}{26}}, \\
f_2 &= 42\exp^{\frac{t}{17}}, \\
\gamma_1 &= 0.106\exp^{\frac{pH-8}{0.56}}, \\
\gamma_2 &= 0.52(1+\frac{t}{43})(\frac{s}{35})\exp^{\frac{-d}{6}}, \\
\gamma_3 &= 0.00049\exp^{-(\frac{t}{27}+\frac{d}{17})}
\end{aligned}$$

For correct implementation of all equations, it is very important to understand and know units for all parameters which are summarized in Table I.

We have implemented the attenuation loss approximation above and calculated the resulted absorption coefficient for many possible combinations of input variables shown in Fig. 4. Note the increasing trend of the absorption coefficient with increasing frequency.

The Noise Loss in 2 is mainly due to ambient noise. There are four major sources for ambient noise in underwater acoustic channel namely; turbulence, shipping, wind driven waves and thermal noise. Noise is measured as power spectral

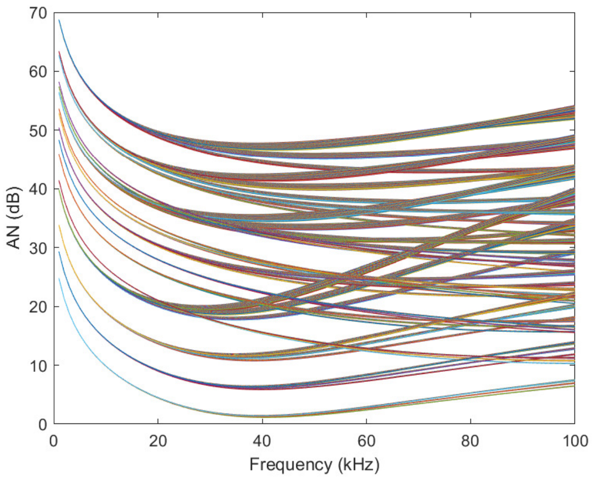


Fig. 7. Absorption Coefficient,α [dB/km] for Different Combination of Input Variables as shown in Table II

TABLE I. Acoustic Channel Model Parameters and Units

Parameter	Description	Unit
rx_{power}	Received Signal Power	dB
rx_{level}	Received Signal Threshold	dB
tx_{power}	Transmitting Signal Power	dB
A	Attenuation Loss	dB
N	Noise Loss	dB
r	Communication Range	Km
f	Frequency	kHz
t	Water Body Temperature	° Celsius
d	Water Depth	Km
s	Water Salinity	ppt
pH	Water Acidity Level	
w	Sea Surface Wind Speed	m/s
sh	Shipping Activity Factor	
κ	Spreading Coefficient	

density and its unit is dB relative to μ Pascal. Noise can be approximated as given by the following formula:

$$N = N_t + N_{sh} + N_{th} + N_w \tag{5}$$

where N_t, N_{sh}, N_{th} and N_w are given by the following formulas:

TABLE II. Selected Parameters Values

Parameter	Values	Unit
r	100 to 1000	m
f	[1 to 200]	kHz
t	[4 to 20]	° Celsius
d	[0 to 10]	Km
s	15, 25, 35	ppt
pH	8.0	
w	0, 5, 10, 15	m/s
κ	1.5	

$$\begin{aligned}
&10logN_t = 17 - 30log(f)\\
&10logN_{sh} = 40 + 20(sh - 0.5) + 26log(f) - 60log(f + 0.03)\\
&10logN_w = 50 + 7.5w^{1/2} + 20log(f) - 40log(f + 0.4)\\
&10logN_{th} = -15 + 20log(f)
\end{aligned} \tag{6}$$

where sh and w are Shipping Activity Factor and Wind Speed, respectively.

Each noise source affects a particular range of frequencies. Low-frequency region, $f < 10$ Hz is influenced by turbulence noise. The frequency range of 10 Hz -100 Hz is majorly influenced by shipping activity factor sh, whose value ranges between 0 and 1 for low and high activity. Wind-driven waves cause surface motion, which is the dominant factor of noise in the frequency region 100 Hz to 100 kHz. It is measured in m/s, and this frequency operating region is used by the majority of acoustic systems. Thermal noise contributes for $f > 100$ kHz [15]. We have implemented Equations 5 and 6 for different values of sh and w over the frequency spectrum [1 Hz to 100000 Hz]. Fig. 5 shows the Noise levels in different spectral regions with the dominant factors in each region. One can notice the decrease trending line as the frequency increases, which shows an opposite behavior compared to attenuation loss above.

Combining both losses effects, i.e., Noise and Attenuation, in the product form, AN would give us insight about communication quality for different sets of conditions. AN is the total loss incurred by the acoustic signal, which in dB can be expressed as in equation 2. Now, let us first examine the single effect of different parameters on AN, as shown in Fig. 6. We run the mathematical model extensively using the parameter combinations listed in Table II[2].

Fig. 6(a) shows the effect of different communication ranges while fixing all other parameters. You can notice the rabid increase of AN product as the transmission range increases especially with higher frequency. In (b), the increase in the temperature slightly increases the loss value as the frequency increases. Salinity level changes affect AN, as shown in (c), which is also has a limited effect. The major effect happens in Figure (d) with wind speed. As wind speed increases from $0m/s$ to $10m/s$, AN increases up to three times.

Fig. 6 shows three main observations as follows:

- Different parameters used in the acoustic channel model have different effects on the AN product.
- Noise loss has two different trending effects as frequency increases with various affecting factors in each frequency range. In general, the loss due to noise decreases as frequency increases, but at the same time, the loss due to attenuation increases as frequency increases.
- The contradicting trends of both losses create a minimal turning point where AN is the minimum. The frequency that generates that minimal AN value is the

[2] The implementation of the mathematical model is available at https://emadfelemban.org/coralsense

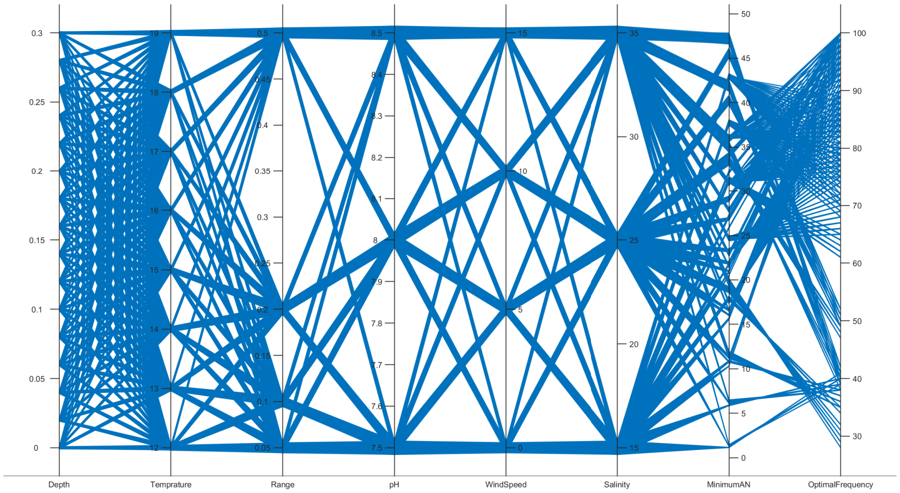


Fig. 8. A Parallel Coordinate Plot Showing All Combinations of Parameter Values Selected

optimal frequency. This optimal frequency changes as the conditions and requirements change.

- Among all parameters that are used in the acoustic channel model, wind speed has a very strong effect on AN.

Fig. 7 shows all the curves for all different combinations to get a comprehensive view. Fig. 8 is a parallel coordinate graph that shows the parameters and their values used to create Fig. 7. The same figure shows the different various optimal frequencies for each case.

IV. Optimal Selection of Communication Parameters

The quality of underwater acoustic communication depends on multiple parameters that can be categorized into three categories:

- Environmental parameters that are related to the ambient environment conditions around the communication area such as temperature, salinity level, wind speed, shipping factor.
- Deployment parameters that are related to deployment conditions of the network such as depth level and transmission range between nodes. Note that these parameters can be changed either manually in the case of fixed networks or using mobile capabilities in the case of mobile or ROV network.
- Communication Parameters that can change and affect the communication channel performance between the source and destination nodes such as transmission power and frequency. Most of the commercially available underwater acoustic modems provide flexibility in setting frequency and transmission power and changing them by software.

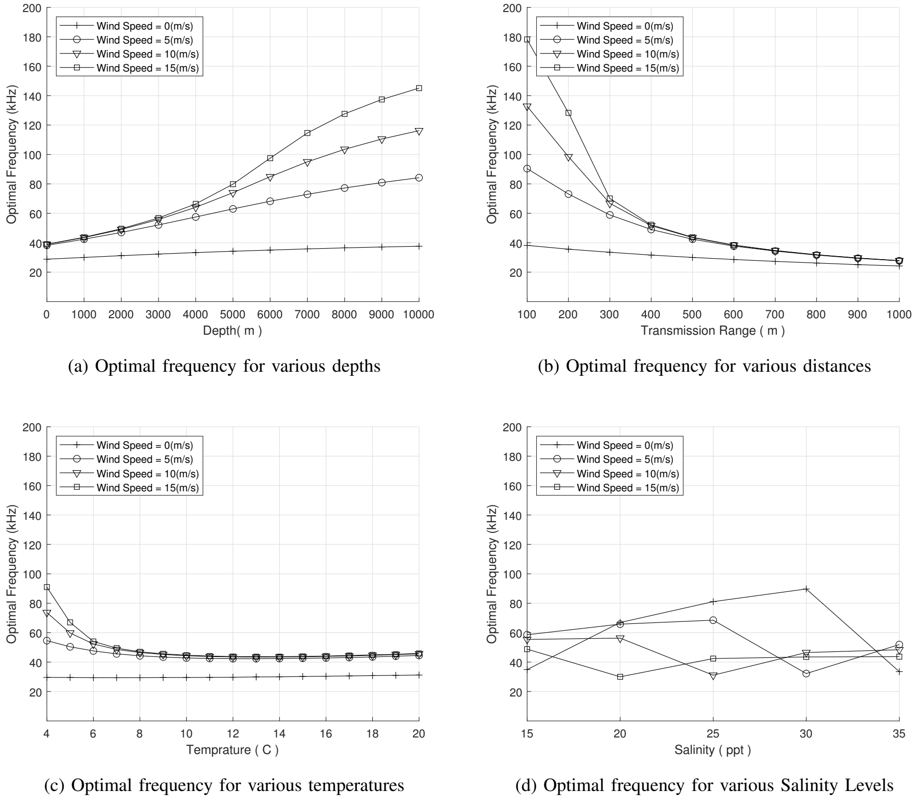
Fig. 9 shows different AN vs Frequency curves calculated with different combinations of parameters. Using the same data, we find the optimal frequency for each case and plot Fig. 6. The optimal transmission frequency provides the lowest AN value and thus most likely provides the best performance in the communication channel. For all plots, we changed the wind speed from 0 m/s to 15 m/s and plotted four curves for each parameter. Fig. 6(a) shows the optimal frequencies as depth changes from 0 Km to 10 Km vs. the increase of wind speed. In Fig. 6(b), (c) and (d) we varied the transmission range, temperature and salinity levels. It is very clear that we need to have a dynamic way to select the appropriate transmission frequency to establish a good communication channel.

V. Conclusion

This paper provides a gateway to find the optimum communication parameters for underwater communication. It offers an insight into the relationship between the different parameters that govern the underwater acoustic communication channel. We have reviewed many mathematical models available for underwater acoustic communication. We implemented our own version of the model and made it available online. We ran many input parameter combinations by changing depth, temperature, wind speed salinity to measure the effect on path loss. We found that wind speed has the most impact on path loss. Finally, we shed some light on optimizing the communication parameters, specifically the frequency. Our future work will include developing the frequency optimizer module and run experimental simulation scenarios to measure the effect of optimal frequency in the simulation environment.

Acknowledgment

This work is supported by the National Science, Technology and Innovation program NSTIP Grante number ”11-INF1688-10” by King Abdulaziz City of Science and Technology.



(a) Optimal frequency for various depths

(b) Optimal frequency for various distances

(c) Optimal frequency for various temperatures

(d) Optimal frequency for various Salinity Levels

Fig. 9. Optimal Frequency Graphs while Changing Different Parameters

REFERENCES

[1] R. Christ and R. Wemli, ”The ROV Manual”, 2nd Edition, Elsevier 2014

[2] https://popotomodem.com/

[3] M. Murad, A. Sheikh, M. Manzoor, E. Felemban, and S. Qaisar, ”A Survey on Current Underwater Acoustic Sensor Network Applications”. International Journal of Computer Theory and Engineering, 2015

[4] P. Amoli, ”An Overview on Current Researches on Underwater Sensor Networks: Applications, Challenges and Future Trends”, International Journal of Electrical and Computer Engineering, vol. 6, p. 955, 2016.

[5] E. Felemban, F. Shaikh, U. Qureshi, A. Sheikh, S. Qaisar, ”Underwater sensor network applications: A comprehensive survey”, International Journal of Distributed Sensor Networks, vol. 2015, pp. 1-14, Aug. 2015.

[6] H. Jindal, S. Saxena and S. Singh, ”Challenges and issues in underwater acoustics sensor networks: A review”, International Conference on Parallel, Distributed and Grid Computing, 2014.

[7] S. Kohli and P. Bhattacharya, ”Characterization of Acoustic Channel for Underwater Wireless Sensor Networks”, Annual IEEE India Conference (INDICON), pp. 1-4, 2005.

[8] I. Akyildiz, D. Pompili and T. Melodia, ”Underwater acoustic sensor networks: Research challenges”, Ad Hoc Networks, vol. 3, pp. 257–279, 2005.

[9] X. Jin, Y. Chen and X. Xu, The analysis of hops for multi-hop cooperation in Underwater Acoustic Sensor Networks, 2016 IEEE/OES China Ocean Acoustics (COA), Harbin, 2016, pp. 1-5.

[10] M. Felemban and E. Felemban, ”Energy-delay tradeoffs for Underwater Acoustic Sensor Networks”, ”First International Black Sea Conference on Communications and Networking (BlackSeaCom)”, Batumi, pp. 45-49. 2013.

[11] A. Stefanov, ”Distortion analysis of underwater acoustic sensor networks”, 7th International Conference on New Technologies, Mobility and Security (NTMS)”, pp. 1-4, 2015.

[12] L. Berkhovskikh and Y.Lysanov, ”Fundamentals of Ocean Acoustics” New York: Springer, 1982.

[13] M. Ainslie and J. McColm, “A simplified formula for viscous and chemical absorption in sea water,” Acoustical Society of America Journal, vol. 103, pp. 1671–1672, Mar. 1998.

[14] F. Fisher and V. Simmons, “Sound Absorbtion in sea water” Acoustical Society of America Journal, vol.62 , pp. 558-564, 1977.

[15] M. Stojanovic, “On the relationship between capacity and distance in an underwater acoustic communication channel,” ACM SIGMOBILE Mobile Computing and Communications Review, vol. 11, no. 4, pp. 34–43, 2007

[16] K. Mackenzie, ”Nineterm equation for sound speed in the oceans,” The Journal of the Acoustical Society of America, vol. 70, p. 807, 1981.

[17] A. Sehgal, I. Tumar, and J. Schonwalder, ”Variability of Available Capacity due to the Effects of Depth and Temperature in the Underwater Acoustic Communication Channel”, IEEE Ocean’s, pp. 1-6, 2009.

[18] A. Sehgal, C. David and J. Schonwalder, ”Energy consumption analysis of underwater acoustic sensor networks,” IEEE OCEANS’1, pp. 1-6, 2009.

[19] E. Sozer, M. Stojanovic, and J. Proakis, ”Underwater acoustic networks,” IEEE Journal of Oceanic Engineering, vol. 25, no. I, pp. 72-83, January 2000.

[20] M. Stojanovic, J. Preisig, ”Underwater Acoustic Communication Channels: Propagation Models and Statistical Characterization,” IEEE Communications Magazine, January 2009.

[21] C. Bassett, J. Thomson, B. Polague, ”Characteristics of Underwater Ambient Noise at a Proposed Tidal Energy Site in Puget Sound,” Oceans 2010, September 2010.

[22] https://emadfelemban.org/coralsense

Automated Recognition of Sincere Apologies from Acoustics of Speech

Zafi Sherhan Syed[1]
Mehran University,
Pakistan

Muhammad Shehram Shah[2]
RMIT University,
Australia

Abbas Shah Syed[3]
University of Louisville,
USA

***Abstract*—Sincerity is an important characteristic of communicative behavior which represents an honest, truthful, and genuine display of verbal and non-verbal expressions. Individuals who are deemed sincere often appear more charismatic and can influence a large number of people. In this paper, we propose a multi-model fusion framework to identify sincerely delivered apologies by modelling difference between acoustics of sincere and insincere utterances. The efficacy of this framework is benchmarked using the Sincere Apology Corpus (SAC). We show that our proposed methods can improve the baseline classification performance (in terms of unweighted average recall) for SAC from 66.02% to 70.97% for the validation partition and 66.61% to 75.49% for the test partition. Moreover, as part of our investigation, we found that gender dependency can influence the classification performance of machine learning models, with models trained for male subjects performing better than those trained for female subjects.**

***Keywords*—Sincerity; affective computing; social signal processing**

I. Introduction

Sincerity is an act of being sincere. It is a quality of human beings that makes them free from pretense, deceit, and hypocrisy. Generally, it is believed that if a person is perceived to be sincere, more people will trust them. Sincerity and trust are at the heart of social interactions, be it in the form of relationships about business or personal life.

Sincerity is an important aspect of human behavior which is useful for many different day-to-day activities. For example, sincerity is a vital part of business dealing and along with honesty is considered to be one of its core values. Similarly, the perceived sincerity of public representatives, such as politicians, can significantly improve their chances of winning elections. Sincerity is also an important factor in healthcare where the trust needs to be established between clinicians and their patients. Finally, sincerity and truthfulness are important aspects of law prosecution where it needs to be ensured that the witness is being truthful in the court of law. To summarize, sincerity is an important aspect of human behavior that affects almost every part of society.

Recent advances in social signal processing have encouraged researchers to investigate aspects of human behavior that influence major sectors such as health and commerce. While sincerity recognition has not been investigated by the social signal processing research community in great detail, one can note that research into deception recognition has been a popular field of research [1], [2], [3], [4], [5], [6], [7], [8].

At the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH) held in late 2019, Baird et al. [9] published a relatively large corpus of speech recordings called the *Sincerity Apology Corpus (SAC)* which have been labeled explicitly for the task of sincerity recognition. To the best of our knowledge, this is the largest publicly available dataset in this field and therefore it provides researchers an opportunity to develop frameworks to recognize whether speech is perceived sincere or insincere.

In addition to releasing the dataset, Baird et al. [9] also investigated the efficacy of three types of audio features for the task of sincerity detection from speech. These features include, a) Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [10], b) Computational Paralinguistics Challenge (ComParE) feature set [11], and c) DeepSpectrum features [12]. While eGeMAPS and ComParE features are traditional non-deep learning based features, the DeepSpectrum features are generated by feeding audio signals into the AlexNet network for large scale image classification [13]. They reported classification performance in terms of unweighted average recall (UAR) for the three types of features which showed that the Deep Spectrum achieved 79.2% on the test partition of the cross-validation folds, whereas eGeMAPS and ComParE features achieved a UAR of 72.0% and 76.2%, respectively. As per Baird et al., these results are meant to serve as the baseline classification performance for further research in the field of sincerity detection, in particular research based on the SAC corpus. It is important to mention here that a part of the Sincerity Apology Corpus was also used for Computational Paralinguistics Challenge of 2016 [11]. There, the objective was to train machine learning models for a regression task with the objective to predict the sincerity score allocated to each audio recording by a group 16 annotators, whereas Baird et al. focus on a classification task to differentiate between sincere and insincere apologies.

In this paper, we propose a multi-model fusion framework for automated recognition of sincere apologies from acoustics of speech and test it using the Sincere Apology Corpus, a dataset publicly available for academic research. In addition to this, we investigate the influence of gender on the classification performance of machine learning models for the task at hand. The rest of this paper is organized as follows: In section II we provide a summary of the Sincere Apology Corpus, whereas in section III we detail the methodology for feature aggregation and model fusion methods. Experimental results and discussion is provided in section IV, and conclusion is provided in section V.

II. Dataset

Audio recordings in the Sincere Apology Corpus are provided as dual-channel stereo audio files that are sampled at 44100 Hz. As per the convention of the field, we first converted the audio recordings into a mono-channel by taking the average signal value per sample for the two channels of the stereo signal. Next, the signal is downsampled to a sampling frequency of 16000 Hz using the Librosa library [14]. Finally, each audio recording is normalized such that the dynamic range of the signal lies between -1 and $+1$. Whereas details of the Sincere Apology Corpus are available in [15], a summary of dataset partitions is provided in Table I.

TABLE I. Summary of dataset partitions for gender independent and gender dependent settings

Gender Independent			
	NS	**S**	**Total**
Train	143	142	285
Val	186	184	370
Test	105	151	256
Male			
	NS	**S**	**Total**
Train	98	45	143
Val	59	61	120
Test	77	80	157
Female			
	NS	**S**	**Total**
Train	45	97	142
Val	127	123	250
Test	28	71	99

III. Methodology

In Fig. 1, we illustrate the process flow pipeline of our proposed multi-model fusion framework for automated recognition of sincere apologies. Here, one starts with speech based audio recordings of subjects which are preprocessed into a standard format as discussed in the previous section. The next step is to compute acoustic low-level descriptors (LLDs) which quantify characteristics of speech paralinguistics. In the current work, we use the IS10Paralinguistics feature set, the ComParE feature set and the eGeMAPS feature set. These LLDs need to undergo a process of feature aggregation which yields a higher level representation of speech acoustics. To this end, we use functionals, bag of audio words, and Fisher Vector encoding based feature aggregation. Next, machine learning models are trained using the training partition, their hyper-parameters are optimized using the validation, and the performance of machine learning models is compared against one-another in an unbiased manner using the test partition. Finally, model fusion approaches are used with the aim to improve the classification performance of the sincerity recognition framework.

A. Acoustic Features for Speech Paralinguistics

The speech signal is inherently non-stationary in nature and therefore acoustic features need to be computed over short intervals of time over which the speech signal demonstrates some form of stationarity. In speech signal processing, it is common to compute acoustic features over time intervals in the range of $15 - 30$ ms [16], and as a result, such features are called low-level descriptors (LLDs).

These LLDs quantify various acoustic characteristics of a speech signal such as its fundamental frequency (also known as pitch), the quality of voice, spectral characteristics, and more. Given that elementary discussion on the characteristics of acoustic features is beyond the scope of this paper, we refer the reader to [10] for further discussion. Typically, these acoustic LLDs are packaged together and used in the form of feature sets. In our work, we shall use feature sets that have been shown to be useful for a variety of tasks related to speech paralinguistics.

In the baseline paper for the SAC corpus, Baird et al. [15] had used four fundamental types of feature sets. These include two domain-knowledge based feature sets: the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) and the Computational Paralinguistics Challenge (ComParE), as well as acoustic representations derived from the AlexNet deep neural network for image classification [13], [12]. In our work, we shall make use of eGeMAPS and ComParE feature sets (similar to the baseline paper) but also include the IS10-Paralinguistics feature set which we have previously found to be useful for tasks related to speech paralinguistics [17], [18]. We shall provide a brief description of these features in the following paragraphs.

The IS10-Paralinguistics feature set consists of 38 acoustic LLDs which include 31 LLDs that describe spectral characteristics of speech, 6 LLDs which describe voicing related characteristics, and an LLD to describe the energy of voice (in terms of loudness). As the name suggests, the IS10-Paralinguistics feature set was especially designed to characterize paralinguistics characteristics of speech. For further details of this feature set, we refer the reader to [19]. Meanwhile, the ComParE feature set consists of 65 acoustic LLDs of which 55 LLDs describe the spectral characteristics of the speech signal, 6 LLDs quantify voicing related characteristics, and 4 LLDs describe energy-related LLDs. The ComParE feature set is often called a brute force feature set since it provides a more holistic approach to modeling speech characteristics. Finally, the eGeMAPS feature set was proposed as an optimized version of the ComParE feature set in terms of feature set dimensionality. The eGeMAPS feature set consists of 23 acoustic LLDs which include 9 spectral LLDs, 13 voicing related LLDs, and 1 energy-related LLDs. For further details of this feature set, we refer the reader to [10].

B. Feature Aggregation: Functionals

As mentioned earlier, due to the non-stationary nature of speech, acoustic features are computed for short duration frames of the audio signal (typically in the range of $20 - 30$ ms). These acoustic features, called low-level descriptors (LLDs) only provide low-level information. Therefore, in order to generate a global representation for an audio recording, the information provided by acoustic LLDs needs to be aggregated by appropriate methods. The simplest and commonly used feature aggregation method uses functionals of descriptive statistics such as mean, variance, range et cetera. In this work, we use a set of standard functionals as defined in the openSmile toolkit [20]. The toolkit is the defacto standard in the field of social signal processing due to its open-source nature and free availability for academic research.

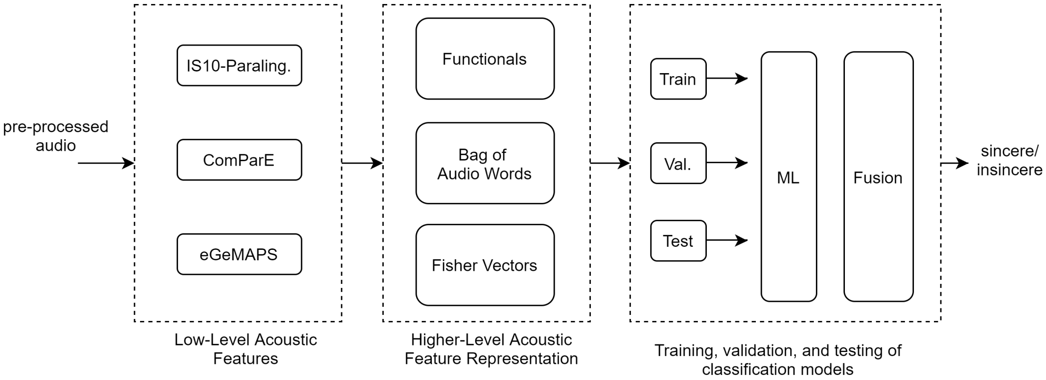

pre-processed audio
IS10-Paraling.
ComParE
eGeMAPS
Low-Level Acoustic Features
Functionals
Bag of Audio Words
Fisher Vectors
Higher-Level Acoustic Feature Representation
Train
Val.
Test
ML
Fusion
Training, validation, and testing of classification models
sincere/ insincere

Fig. 1. Illustration of the Pipeline for Baseline Classification

C. Feature Aggregation: Bag of Audio Words

An alternate to functionals based feature aggregation is the Bag-of-Audio-Words (BoAW) method which is an extension of the bag-of-words (BoW) method from the field of natural language processing. BoW has been a popular approach to generate word-frequency histogram-based representation of text documents for applications related to text processing. The same concept has also been extended in audio signal processing to yield a global histogram-based representation for audio recordings [21], [22]. Unlike the text domain where textual words naturally exist, one needs to compute *audio words* through a process called vector quantization. Here, acoustic LLDs from all audio files are concatenated into a matrix and a clustering algorithm is used to learn representative clusters for the LLDs. Each cluster is called an audio word and the set of clusters for acoustic LLDs is called the codebook. It was common to use the k-means clustering algorithm, however Rawat et al. discovered that clustering based on random selection performs just as well with the advantage of a considerably smaller computational complexity [23]. Henceforth, it has become common practice to use random sampling for learning the codebook. As a result, we shall also use a random sampling approach for clustering in this work. The BoAW approach requires tuning of hyperparameters such as the codebook size, the number of simultaneous assignments to multiple audio words (in case an acoustic LLD is close in terms of Euclidean distance to multiple audio words), and normalization. We shall perform hyperparameter optimization for BoAW using the validation partition and from there select the best performing model for predicting test partition labels. In order to compute BoAW, we use the openXBOW toolkit [24].

D. Feature Aggregation: Fisher Vectors

Fisher vector encoding is a feature aggregation method which was initially proposed by Perronin et al. [25], [26] for applications in the field of computer vision, achieving state-of-the-art performance [27] for object recognition before the advent of deep learning for computer vision era [13]. This method has also been found useful for applications related to speech paralinguisitics [28], [29], [30]. In our previous works, we found the Fisher Vector based feature aggregation to be useful for speech screening of depression [31] and bipolar disorder [17]. Fisher Vector representation combines the advantages of generative models i.e. the ability to work with variable-length data and discriminative models i.e. ability to learn class-specific boundaries.

In order to compute Fisher Vector features for the task of sincerity recognition, we follow the approach detailed by Perronin et al. [25], [26] but adapt it for audio signals. To this end, we first concatenate acoustic LLDs from all audio recordings and train a Gaussian Mixture Model (GMM) [32], which serves as the generative model of the Fisher Vector framework. Next, the first and second-order statistics for gradients between acoustic LLDs from an audio recording and the generative model are computed. These statistics are then concatenated to yield a single feature vector and called Fisher Vector (FV) features. In the current work, we use the VLfeat toolkit [33] in order to train GMMs as well as compute FV features.

E. Fusion

Fusion is a method through which it is possible to combine information from multiple machine learning models with the aim to improve overall classification performance[34]. There are two fundamental ways to fuse such information i.e. label fusion and confidence fusion. In label fusion, class-label predictions from machine learning models are stacked and the class which is predicted by the majority of models is deemed to be the correctly predicted class. For example, if three models in five-model fusion predict the label for a speech recording to be *Sincere* whereas two models predict the label as *Not-Sincere* then the final label will be decided as *Sincere* based on a majority vote. Meanwhile, confidence fusion takes place on probabilistic or confidence outputs of the classifier. It is reminded that each classification model returns a confidence metric for the predicted label which essentially quantifies how sure it is about its predictions. Naturally, it predicts the label for which it has the most confidence. Given that different models are trained with different features and parameters, the confidence may be different and such information may help to improve the overall accuracy of prediction. In confidence fusion, the idea is to decide on the class-label based on the confidence of multiple machine learning models. The simplest way to implement confidence fusion is to take the arithmetic average of the confidence metric from multiple machine learning models and predict the label which has more confidence.

IV. Experimentation, Results and Discussion

We use the implementations of logistic regression (LR), support vector machine (SVM), and random forest (RF) classifier which are available in the scikit-learn toolkit [1]. The complexity value of the LR and SVM is optimized over a logarithmically spaced grid between 10^{-7} to 10^{7}. The RF classifier has a number of hyperparameters which need to be optimized, such as a) the number of trees in the forest (num_est), b) the maximum depth of each tree (*max_depth*), c) the minimum number of samples before splitting at a node (*min_samps_split*), and d) the minimum number of samples required to be at a leaf node (*min_samps_leaf*). To this end, we conduct gridsearch based optimization with the following parameter values: *num_est* = $\{25, 50, 100, 200, 400\}$, *max_depth* = $\{2, 5, 10, 15, 20\}$, *min_samps_split* = $\{2, 5, 10, 15, 20\}$, and *min_samps_leaf* = $\{2, 5, 10, 15, 20\}$. These classifiers are trained using the training partition, their hyperparameters are optimized using the validation partition, and the classification results being compared against the test partition. For the sake of completeness, we report the results for both validation and test partitions.

A. Baseline Classification Performance

In Table II, we provide a summary of the baseline classification performance for the Sincere Apology Corpus which was reported by Baird et al. [15]. Furthermore, we also report the results we achieved using the same feature set (note that along with audio recordings, Baird et al. also provided their features) albeit with three different classifiers, that are LR, SVM, and RF whereas Baird et al. only used SVM.

Ideally, one expects that the results provided in the baseline paper and those computed by us with the SVM classifier would be the same given that features and the SVM classifier are similar; but these are not. In fact, we find that all results of machine learning models reported by Baird et al. achieve a greater classification accuracy on the test partition as compared to the results we computed. One can think of two possible reasons: 1) the random seed and the number of iterations for training the SVM classifier could be different between our implementation and that of [15] which can lead to a difference in results, and 2) Baird et al. optimized the SVM for results on the test partition directly whereas we optimized the SVM classifier for the validation partition and only used the test partition for comparing classification performance across different models. Furthermore, one should also note that Baird et al. did not report results for the validation partition which would have otherwise made their results easier to interpret.

Given this, we shall use classification performance achieved through our experiments as the baseline and shall carry out our investigations on feature aggregation, gender dependency, and model fusion for binary classification between sincere and insincere apologies. To this end, we report that eGeMAPS functionals when used with the SVM classifier, provide the best classification performance of UAR = 66.20% for the development partition, and the corresponding model achieves a UAR = 66.61% for the test partition. This shall be the baseline classification performance.

B. Experiments with Gender Independent Partitions

We first compare the classification performance of feature aggregation methods in a gender-independent setting. In Table III, we provide a summary of results for classification between sincere and insincere apologies for functionals, BoAW, and FV based feature aggregation of IS10-Paralinguistics, eGeMAPS, and ComParE features for a gender independent setting. It is clear to note that the baseline UAR = 66.02% for the development partition can be improved by all three feature aggregation methods. The top performing models from each method are Funcs-IS10Paraling-RF, BoAW-ComParE-RF, and FV-ComParE-RF. Overall, the best performing model for the validation partition is BoAW-ComParE-RF with a UAR = 67.08% which goes on to achieve a UAR = 68.78% on the test partition.

C. Experiments with Gender Dependent Partitions

It is known that the subject's gender can influence the paralinguistic characteristics of speech for applications such as emotion valance recognition [35] and depression recognition [36], [37]. We, therefore, investigate the effect of gender on the accuracy with which machine learning models can differentiate between sincere and insincere apologies. To this end, we conducted experiments as discussed previously with gender-dependent partitions and provide a summary of results for male gender in Table IV and female gender in Table V. It is important to mention here that since Baird et al. did not report results of classification performance under a gender-dependent setting, therefore, a baseline does not exist, and we shall introduce a baseline for gender-dependent settings as part of our work.

From Table IV, we note that the best performance for the validation partition is achieved by Funcs-eGeMAPS-RF with a UAR = 80.79% although the performance drops significantly to 55.52% on the test partition. It is interesting to note that a number of models achieve similar performance as Funcs-eGeMAPS-RF on the validation partition, such as BoAW-IS10Paraling-RF with UAR = 80.02%, FV-ComParE-SVM with UAR = 79.91%, FV-ComParE-LR with UAR = 79.04%, and FV-IS10Paraling-SVM with UAR = 79.01%. Amongst these, FV-IS10Paraling-SVM achieves the highest performance on the test partition with a UAR = 69.34% whereas FV-ComParE-SVM and FV-ComParE-LR achieve a UAR of approximately 67.50%. These results suggest overfitting on the validation partition since there exists a large difference between the UAR values achieved for validation and test partition.

As far as recognition of sincere and insincere apologies for female subjects is concerned, we find somewhat poorer classification performance of machine learning models as compared to the case of male subjects. Here, the best performing model FV-IS10Paralig-LR achieves a UAR = 67.89% on the validation partition and a UAR = 72.33% on the test partition. The best performing model amongst BoAW based feature aggregation is the BoAW-IS10Paraling-LR which achieved a UAR = 64.88% for the validation partition whereas the best performing with functionals based aggregation achieved a UAR = 64.82% for the same partition. This suggests that gender does influence the paralinguistic characteristics of sincere and insincere speech.

[1] https://scikit-learn.org

TABLE II. SUMMARY OF RESULTS PROVIDED AS BASELINE IN [15] AND FROM EXPERIMENT PERFORMED BY THESE AUTHORS USING BASELINE FEATURES

Feature Name		*UAR (%)*					
		SVM		*LR*		*RF*	
		Val.	*Test*	*Val.*	*Test*	*Val.*	*Test*
Results provided	ComParE-funcs	-	70.00	-	-	-	-
	eGeMAPS-funcs	-	70.20	-	-	-	-
	DeepSpectrum-MelSpec-fc6	-	69.80	-	-	-	-
	DeepSpectrum-LinSpec-fc6	-	69.90	-	-	-	-
	DeepSpectrum-MelSpec-fc7	-	65.90	-	-	-	-
	DeepSpectrum-LinSpec-fc7	-	68.60	-	-	-	-
Our experiments	ComParE-funcs	62.07	66.40	64.46	64.50	**65.24**	62.80
	eGeMAPS-funcs	**66.02**	66.61	63.62	63.75	63.07	63.90
	DeepSpectrum-MelSpec-fc6	51.87	53.95	58.17	66.05	57.88	66.34
	DeepSpectrum-LinSpec-fc6	57.23	57.76	61.65	65.16	58.98	65.74
	DeepSpectrum-MelSpec-fc7	59.85	65.80	60.11	65.43	58.46	64.64
	DeepSpectrum-LinSpec-fc7	53.48	53.00	58.20	66.05	**59.81**	65.30

TABLE III. SUMMARY OF RESULTS FOR FUNCTIONALS, BAG-OF-AUDIO WORDS, AND FISHER VECTOR FEATURES FOR GENDER INDEPENDENT SETTING OF TRAINING, VALIDATION, AND TEST PARTITIONS

Feature Name		*UAR (%)*					
		SVM		*LR*		*RF*	
		Val.	*Test*	*Val.*	*Test*	*Val.*	*Test*
Functionals	IS10Paral.	64.46	70.75	64.42	68.97	**66.84**	70.01
	ComParE	62.84	66.87	65.23	65.37	64.71	64.54
	eGeMAPS	63.02	59.47	62.57	66.34	**66.05**	63.13
BoAW	IS10Paral.	**66.00**	62.68	65.73	65.00	64.67	65.29
	ComParE	61.59	56.01	62.98	62.18	**67.08**	68.78
	eGeMAPS	59.48	64.11	59.19	67.48	64.66	60.96
FV	IS10Paral.	62.72	71.87	63.78	70.01	63.33	67.50
	ComParE	65.73	69.39	65.45	69.82	**66.60**	69.40
	eGeMAPS	63.82	62.70	64.08	65.16	64.17	68.45

TABLE IV. SUMMARY OF RESULTS FOR FUNCTIONALS, BAG-OF-AUDIO WORDS, AND FISHER VECTOR FEATURES FOR TRAINING, VALIDATION, AND TEST PARTITIONS WITH MALE SUBJECTS ONLY

Feature Name		*UAR (%)*					
		SVM		*LR*		*RF*	
		Val.	*Test*	*Val.*	*Test*	*Val.*	*Test*
Functionals	IS10Paraling	72.06	69.72	71.33	71.14	74.91	52.32
	ComParE	73.81	69.05	73.02	66.50	72.05	57.23
	eGeMAPS	68.78	60.78	75.62	64.15	**80.79**	55.74
BoAW	IS10Paraling	76.55	66.82	78.19	66.70	**80.02**	60.76
	ComParE	69.03	62.32	68.30	57.32	77.42	66.19
	eGeMAPS	71.77	62.35	67.81	55.06	75.90	50.64
FV	IS10Paraling	**79.01**	69.34	79.01	67.42	73.63	54.16
	ComParE	**79.91**	67.54	**79.04**	67.47	76.09	67.18
	eGeMAPS	75.70	59.82	78.27	61.31	74.34	62.09

TABLE V. SUMMARY OF RESULTS FOR FUNCTIONALS, BAG-OF-AUDIO WORDS, AND FISHER VECTOR FEATURES FOR TRAINING, VALIDATION, AND TEST PARTITIONS WITH FEMALE SUBJECTS ONLY

Feature Name		*UAR (%)*					
		SVM		*LR*		*RF*	
		Val.	*Test*	*Val.*	*Test*	*Val.*	*Test*
Functionals	IS10Paraling	59.97	59.68	60.43	58.98	59.32	64.71
	ComParE	56.64	58.60	56.98	67.91	58.22	67.91
	eGeMAPS	**64.82**	50.48	**65.21**	56.54	62.68	64.34
BoAW	IS10Paraling	64.22	69.52	**64.88**	64.81	**64.86**	64.76
	ComParE	63.66	68.86	63.38	62.75	62.33	68.66
	eGeMAPS	61.83	50.96	62.61	54.15	64.31	68.66
FV	IS10Paraling	60.65	63.63	**67.89**	72.33	63.17	64.71
	ComParE	61.61	72.56	60.50	62.93	60.88	66.83
	eGeMAPS	63.43	58.32	63.78	62.22	61.79	58.27

D. Model Fusion

Finally, in Table VI, we provide a summary of results for label- and confidence-based fusion for predicting sincerity for the gender independent setting. Here, we chose to fuse the results from top-5 performing models. The results show that both fusion approaches can help improve the classification performance for validation as well as test partitions. Interestingly, there is little difference between the UAR achieved by label-based and confidence-based fusion approaches for the validation partitions but for the test partition, label-based fusion provides a better UAR with 75.49% compared to 73.22% as achieved by confidence-based fusion.

TABLE VI. SUMMARY OF RESULTS FOR LABEL- AND CONFIDENCE-BASED FOR TOP-5 PERFORMING MODELS

Model Name	**UAR (%)**	
	Val.	*Test*
BoAW-ComParE-RF	**67.08**	68.78
Funcs-IS10Paraling-RF	66.84	**70.01**
FV-ComParE-RF	66.60	69.40
Funcs-eGeMAPS-RF	66.05	63.13
BoAW-IS10Paralig-LinSVM	66.00	62.68
Label Fusion	70.79	**75.49**
Conf. Fusion	**70.97**	73.22

V. CONCLUSION

The purpose of the current study was to propose a multi-model fusion based framework for identifying speech recordings which carry insincere apologies amongst a corpus which also contains recordings of sincere apologies. To this end, our proposed methods were able to improve the classification performance for the Sincere Apology Corpus from 66.02% to 70.97% for the validation partition and 66.61% to 75.49% for the test partition. We also proposed new baselines for gender dependent classification between sincere and insincere apologies and report that classification models tend to perform better for male subjects as compared to female subjects.

REFERENCES

[1] J. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, B. Pellom, E. Shriberg, and A. Stolcke, "Distinguishing deceptive from non-deceptive speech," in *INTERSPEECH*, 2005, pp. 1833–1836.

[2] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining prosodic lexical and cepstral systems for deceptive speech detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2006, pp. 1033–1036.

[3] J. F. Torres, E. Moore, and E. Bryant, "A study of glottal waveform features for deceptive speech classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4489–4492.

[4] N. Raiman, H. Hung, and G. Englebienne, "Move, and I will tell you who you are: Detecting deceptive roles in low-quality data," in *ACM International Conference on Multimodal Interaction*, 2011, pp. 201–204.

[5] C. Fan, H. Zhao, X. Chen, X. Fan, and S. Chen, "Distinguishing deception from non-deception in Chinese speech," in *International Conference on Intelligent Control and Information Processing (ICICIP)*, 2015, pp. 268–273.

[6] S. I. Levitan, G. An, M. Wang, G. Mendels, J. Hirschberg, M. Levine, and A. Rosenberg, "Cross-cultural production and detection of deception from speech," in *WMDD 2015 - Proceedings of the ACM Workshop on Multimodal Deception Detection, co-located with ICMI 2015*, 2015.

[7] C. Montacie and M. J. Caraty, "Prosodic cues and answer type detection for the deception sub-challenge," in *INTERSPEECH*, 2016, pp. 2016–2020.

[8] G. Mendels, S. I. Levitan, K. Z. Lee, and J. Hirschberg, "Hybrid acoustic-lexical deep learning approach for deception detection," in *INTERSPEECH*, 2017, pp. 1472–1476.

[9] A. Baird, S. Amiriparian, N. Cummins, S. Sturmbauer, J. Janson, E.-M. Messner, H. Baumeister, N. Rohleder, and B. Schuller, "Using Speech to Predict Sequentially Measured Cortisol Levels During a Trier Social Stress Test," in *INTERSPEECH*, 2019, pp. 534–538.

[10] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[11] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native language," in *INTERSPEECH*, 2016, pp. 2001–2005.

[12] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech," in *ACM on Multimedia Conference*, 2017, pp. 478–484.

[13] A. Krizhevsky, I. Sutskever, and H. Geoffrey E., "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1–9.

[14] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "c," in *14th Python in Science Conference*, 2015, pp. 18–24.

[15] A. Baird, E. Coutinho, J. Hirschberg, and B. Schuller, "Sincerity in Acted Speech: Presenting the Sincere Apology Corpus and Results," in *INTERSPEECH*, 2019, pp. 539–543.

[16] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis*, 1st ed. Academic Press Inc. (London) Ltd., 2014.

[17] Z. S. Syed, K. Sidorov, and D. Marshall, "Automated Screening for Bipolar Disorder from Audio/Visual Modalities," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2018, pp. 39–45.

[18] Z. S. Syed, S. A. Memon, M. S. Shah, and A. S. Syed, "Introducing the Urdu-Sindhi Speech Emotion Corpus: A novel dataset of speech recordings for emotion recognition for two low-resource languages," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 1–6, 2020.

[19] S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, M. Christian, S. Language, P. Group, D. Telekom, and A. G. Laboratories, "The INTERSPEECH 2010 Paralinguistic Challenge," in *INTERSPEECH*, 2010, pp. 2794–2797.

[20] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *ACM international conference on Multimedia*, 2013, pp. 835–838.

[21] Y. Liu, W. L. Zhao, C. W. Ngo, C. S. Xu, and H. Q. Lu, "Coherent bag-of audio words model for efficient large-scale video copy detection," in *ACM International Conference on Image and Video Retrieval*, Xi'an, China, 2010, pp. 89–96.

[22] S. Pancoast and M. Akbacak, "Bag-of-Audio-Words Approach for Multimedia Event Classification." in *INTERSPEECH*, Portland, Oregon, USA, 2012, pp. 2105–2108.

[23] S. Rawat, P. F. Schulam, S. Burger, D. Ding, Y. Wang, and F. Metze, "Robust Audio-Codebooks for Large-Scale Event Detection in Consumer Videos," in *INTERSPEECH*, 2013, pp. 2929–2933.

[24] M. Schmitt and B. Schuller, "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[25] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[26] F. Perronnin, J. Sanchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Lecture Notes in Computer Science*, vol. 6314, 2010, pp. 143–156.

[27] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, 2013.

[28] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression Estimation Using Audiovisual Features and Fisher Vector Encoding," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014, pp. 87–91.

[29] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 255–259.

[30] H. Kaya, F. Gurpinar, S. Afshar, and A. A. Salah, "Contrasting and Combining Least Squares Based Learners for Emotion Recognition in the Wild," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2015, pp. 459–466.

[31] Z. S. Syed, K. Sidorov, and D. Marshall, "Depression Severity Prediction Based on Biomarkers of Psychomotor Retardation," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2017, pp. 37–43.

[32] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[33] A. Vedaldi and B. Fulkerson, "VLFeat: An Open and Portable Library of Computer Vision Algorithms," in *ACM International Conference on Multimedia*, 2010, pp. 1469–1472.

[34] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Information Fusion*, vol. 57, no. 1, pp. 115–129, 2020.

[35] H. Sagha, J. Deng, and B. Schuller, "The effect of personality trait, age, and gender on the performance of automatic speech valence recognition," in *ACM International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 1–5.

[36] Z. Huang, B. Stasak, T. Dang, K. Wataraka Gamage, P. Le, V. Sethu, and J. Epps, "Staircase Regression in OA RVM, Data Selection and Gender Dependency in AVEC 2016," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 19–26.

[37] G. Stratou, S. Scherer, J. Gratch, and L. P. Morency, "Automatic Nonverbal Behavior Indicators of Depression and PTSD: Exploring Gender Differences," in *IEEE Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 147–152.

Document Classification Method based on Graphs and Concepts of Non-rigid 3D Models Approach

Lorena Castillo Galdos[1], Cristian Lopez Del Alamo[2], Grimaldo Dávila Guillén[3]
Escuela Profesional de Ciencia de la Computación
Universidad Nacional de
San Agustín de Arequipa
Arequipa, Peru

Abstract—**Text document classification is an important research topic in the field of information retrieval, and so it is how we represent the information extracted from the documents to be classified. There exists document classification methods and techniques based on the vector space model, which doesn't capture the relation between words, which is considered of importance to make a better comparison and therefore classification. For this reason, two significant contributions were made, the first one is the way to create the feature vector for document comparison, which uses adapted concepts of non-rigid 3D models comparison and graphs as a data structure to represent such documents. The second contribution is the classification method itself, which uses the representative feature vectors of each category to classify new documents.**

Keywords*—*Document classification; graphs; non-rigid 3D models; Universidad Nacional de San Agustín de Arequipa (UNSA)

I. Introduction

Nowadays with the increase of the use of technology, a great amount of textual information is generated as well as the need of innovative methods and techniques for its analysis, comparison, and classification, being the latter defined as the assignment of a category to an unclassified document finding similarities between this and the documents of the different known categories.

There is a wide variety of document classification algorithms, plenty of them are based on similarity comparison techniques [1], whether they are based on the vector space model [2] which treats words independently and does not capture the semantic relations between documents; or methods that do consider them important, which create graphs from the relation between words inside a document [3], [4], [5], [6].

The efficiency of these methods depends mainly of the representation of the documents to be classified, so in this paper it was decided to follow the path of [5] in the utilization of graphs as structure to represent said documents.

Graphs are data structures that are used to represent complex non-structured information about entities and the interaction between them. On the other hand, documents can also be represented as graphs using account concepts of frequencies and relationship between words. Finally, this graph can be used to apply techniques similar to that of three-dimensional meshes for classification.

Also, other areas in computer science can provide some applicable ideas and concepts to the information retrieval field, for example the approach in which this method is basing, uses adapted notions of the area of computer graphics to do a better document similarity comparison, resembling the definition of isomorphism with document semantic similarity. This method takes into account both the individuality and the relations between words, which is used for the document classification since the documents belonging to one category have a very high similitude between one another because when talking about the same topic, exists a very high quantity of words that appear in many documents inside this category, just like consecutive words, which will be detailed in Section IV.

In this paper we propose two significant contributions, the first one is the modification of the work of [5] to obtain feature vectors and the second is the classification method itself, which is based on the obtaining of representative feature vectors per category.

The general objective of this work is to develop a new method of document classification, based on rigid models analysis concepts in geometry processing. The steps to follow are these:

- Select documents to create the training and testing sets.
- Adapt the document comparison approach proposed by [5] to obtain a feature vector representing each category.
- Analyze the new document to obtain its feature vector.
- Apply the proposed classification method to the feature vector of the new document using the feature vectors of all the categories.
- Identify the category the new document belongs to.
- Experiment with the testing set.

This rest of the paper is organized as follows. Section II presents previous concepts. Section III provides an overview of the state of the art. Section IV describes the methodology. Section V evaluates experimental results and we present conclusions on Section VI.

II. Previous Concepts

For a better understanding of the problem and the proposed solution, we define the following concepts.

- **Keypoint**: In 3D models, a *keypoint* is a point which is distinctive in its locality and it is present at all different instances of the object [7].
- **Keyword**: The *keywords* of a document are defined as the words which bring most information about a set of words

inside a neighborhood. Such that, its frequency and the grade in which it is related to its neighbor words are high [5].

- **K-rings and neighborhood**: In 3D models a *k-ring* $R_k(v)$ of a profundity level of k with center on the vertex v is defined by:

$$R_k(v) = \{v' \in V',\ \mid C(v', v) \mid = k\} \tag{1}$$

Where $C(v', v)$ is the shortest path from vertex v' to v and $\mid C(v', v) \mid$ is the size of the path $C(v', v)$. It is important to mention that the size of an edge is always 1 [8].

Then we adapted the concept of *k-ring* so that in documents it is called *neighborhood*.

- **Document graph**:

According to the work of [5], a document graph $G(N, A, W)$ is a representation in which the vertexes N are the terms of a document, the outgoing edges A of each node represent the existing relations between them, while W are the weights of the edges which indicate the importance of a relation. Fig. 1 shows an example of a document graph.

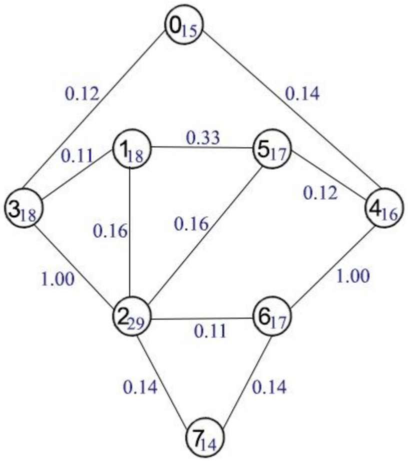


Fig. 1. Example of a Document Graph. [5]

III. State-of-the-Art

In the past years, the document classification task has been widely studied, including approaches of *machine learning* like Bayesian classifiers, Decision trees, K-nearest neighbor (KNN), Support Vector Machines (SVMs), Neural Networks, [9], [10], [11], [12], [13], [14], among others.

This paper focuses on supervised classification since it requires a learning or training process by the classifier. The main idea of supervised classification techniques or algorithms is to build a pattern from each class or category to then find the similitude between this and the new document to be classified.

To perform a better text document classification these can be represented in plenty of ways, this is done to reduce their complexity and to make them easier to handle. The more commonly used representation is the *vector space model* [2], in which the documents are represented as vectors of words. This model does not capture the relationships among words, or the semantic relations between them, for this reason, there exist methods of term weighting a matrix as it is shown in Fig. 2 [15]. A big problem of this representation is that because each entry represents a word of the document, and not all the words appear in every document to be classified, this becomes highly dimensional resulting in a very large disperse matrix [15].

$$\begin{pmatrix} T_1 & T_2 & \dots & T_{at}c_i \\ D_1 & w_{11}w21 & \dots & w_{t1}c_1 \\ D_2 & w_{12}w22 & \dots & w_{t2}c_2 \\ \vdots & \vdots & & \vdots \\ D_n & w_{1n}w2n & \dots & w_{tn}c_n \end{pmatrix} \tag{2}$$

Likewise, documents can be represented using structures like graphs, which demonstrate to better capture the relations among words or terms according to the edges between its vertexes. There are several related works that use this representation [5], [1], [3], [16], [17], [18], [4].

A. Subgraphs and Term Graphs

In the work of [17], they state that a document D_i is represented as a vector of terms $D_i =< w_{1i}, \dots, w_{|T|i} >$ where T is the ordered set of terms that appear at least once in a document inside a collection of documents. Each weight w_{ij} represents how much a term t_j contributes to the semantic of the document. The weight of each term inside a collection of documents is found by building a term graph. The relations between terms are captured using the frequent itemset mining method[1].

In the work of [3], they also use a graph-based approach to classify documents. Their algorithm W-gSpan (weighted subgraph mining algorithm) is applied to identify the subgraphs with frequent weights of the documents, these subgraphs are then used to generate a set of binary feature vectors (one per document), which then serve as entry to the TFPC classifiers (a mining classification association rule), Naive Bayes and decision tree classifier C4.5 showing as a result, a percentage greater than 84% of classification precision using two methods described as follows.

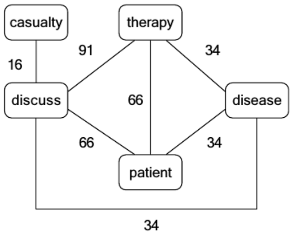
The first classification method consists in treating each term of a graph as a web page to find a PageRank score, which is a method that consists in the idea that if a web page is pointed by several other web pages, then its ranking will be high, or if pages with a high score point to it. Then a rankings vector representing the document is created, and the category whose ranking co-relation coefficients (found with the Spearman algorithm) are higher with this test document is assigned. This vector is used with SVM, obtaining an average of 92% of precision.

The second method is based on the term distance matrix and the distance-weight similarity function. Given a distance matrix set $\{T_1, T_2, \dots, T_n\}$ representing the categories $\{C_1, C_2, \dots, C_n\}$ and a test document D, the document will be classified into the category C_i if and only if the distance-weight similarity of C_i and D is the longest among all the

[1] These algorithms can be used to find subsets of items that surpass a threshold inside a collection.

Itemset	Support
$\{therapy, discuss\}$	91
$\{therapy, discuss, patient\}$	66
$\{therapy, discuss, patient, disease\}$	34
$\{casualty, discuss\}$	16

(a) Frequent Itemsets



(b) The Corresponding Graph

Fig. 2. Example of a Term Graph, in a) the Frequent Itemsets are shown, and in b) its Corresponding Graph [17].

categories. This method obtained an average precision of more than 60%.

Also, there exists other methods that combine this subgraph and term graph approaches to perform the classification task [19], [20], [21].

B. Graphs and Graph-Kernels

In the work of [18], they consider the text classification task as a graph classification problem, model text documents as a graph-of-words, which correspond to a graph in which vertexes represents unique terms of the document and the edges represent co-occurrences between the terms inside a fixed size window. An example of this graph is shown in Fig. 3.

Then, they used linear SVMs to perform the classification because the objective was discovering and exploring new characteristics. To perform the characteristics extraction the used gSpan (graph-based Substructure pattern) to get frequent subgraphs, the minimum quantity of these depends of a parameter known as support, the optimal value of this parameter can be learned trough cross-validation to maximize the prediction precision of the classifier, turning this whole process in a supervised process. When reducing the graphs, it is necessary to keep the more dense parts for which they extract its main cores. This method obtained results of up to more of 90% of precision.

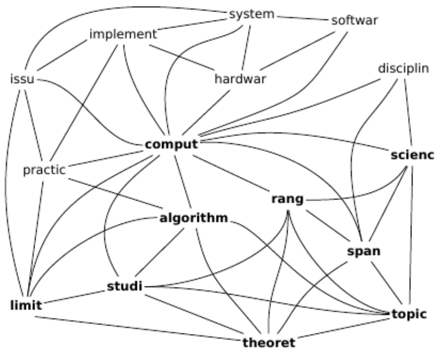


Fig. 3. Graph of Words, Bold Words Represent the Words of the *Main Core* [18].

Author in [4] present a similarity measure based in the definition of a graph-kernel[2] between pairs of documents, using the terms contained in the documents and the relations among them, representing them as a graph-of-words. Specifically they capitalize on the kernel and modify it to compare the graph representations of a pair of documents.

The method takes as entry a pair of documents and automatically computes how similar are one of another based only on their content. This method was tested by doing text categorization for which a SVM classifier was used taking as entry the kernel matrix of the training set, showing results of up to 77% of precision in one database and more than 91% in the other three.

IV. Methodology

Due to the wide variety of techniques that have been developed to solve the document classification problem, is that this paper adopts the innovative approach of Lorena *et al.* [5] of document similarity comparison using graphs and concepts of 3D models and applies it towards document classification.

At the time of obtaining feature vectors from the document graphs, what we look for is to capture better the relation among words inside a document, extracting this way a semantic representation of it, to then be able to use them with the classification method.

A general diagram of the document classification process for a new unclassified document is shown in Fig. 4.

As it was mentioned previously, this paper modifies a previous work approach. Then, its general functioning is explained as well as the modification to obtain the feature vectors and the classification method. The steps performed are enumerated according to Fig. 4.

a) **Preprocessing and graph construction**:
For the preprocessing phase, first we do the cleaning step, which consists in the elimination of stop words[3]. Then the Porter algorithm is applied for the stemming step, which preserves only the roots of the words to avoid the different time, gender and number variations; and because there will be repeated roots we proceed to the ID Assignment step, which assigns numeric IDs to each root, to later be inserted on the list L.

[2] Graph-kernels can be intuitively understood as functions that measure the similarity of pairs of graphs.

[3] They can be pronouns, articles, etc.

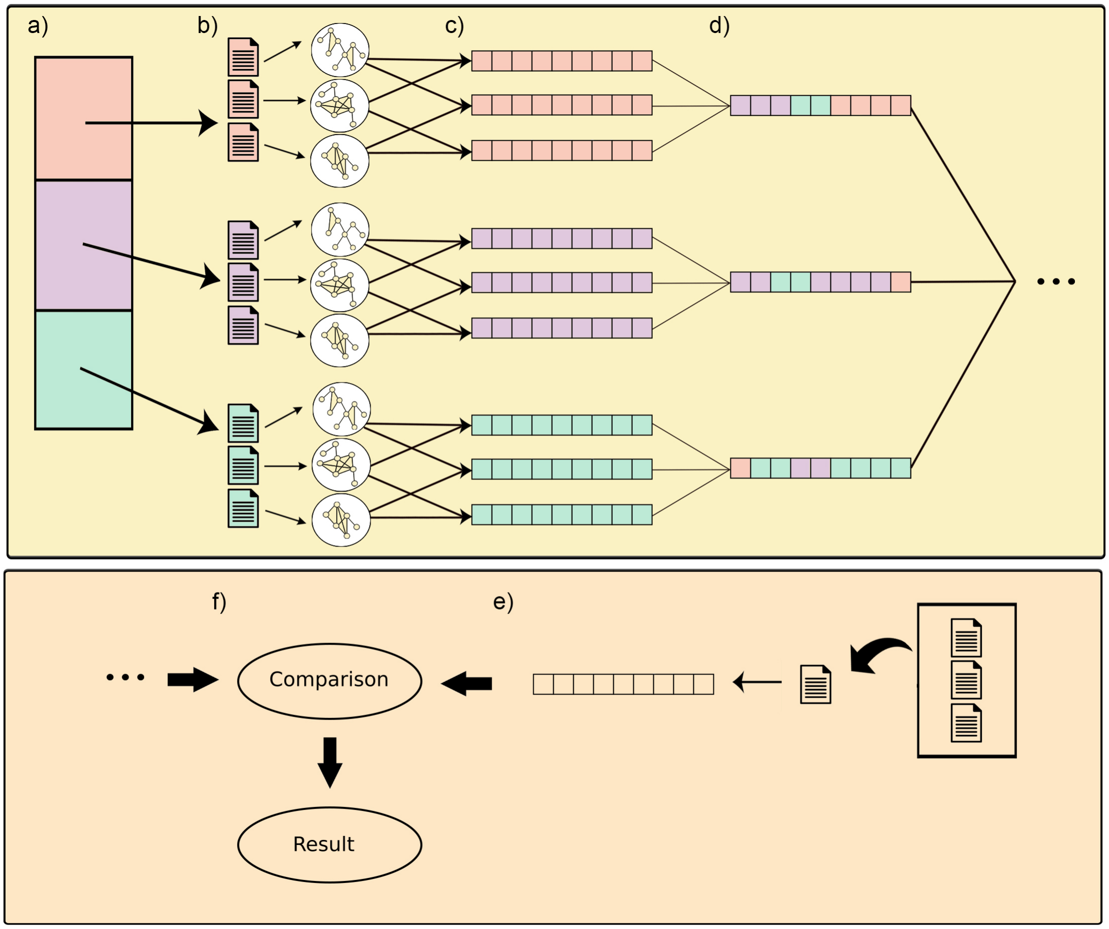


Fig. 4. Pipeline of the Proposed Model.

b) **Graph construction** After the preprocessing step, we proceed to build the graph $G(N, A, W)$ where N are the nodes of the graph, which represent the elements of the list L, A indicates the edges which are the existing relations between the elements of the list L, and W are the weights of the edges. The protruding edges of the nodes represent the grade in which these are related with their neighbors, as it is shown in Fig. 1.

c) **Comparison**:
Following the approach of [5], to perform the comparison between two document graphs G_1 and G_2, first we obtain a list of keywords (L_{kw}) of each graph, which are the μ nodes with greater weights. Then we found a list with the intersection of both lists, which will contain the common keywords between both graphs as it is shown in Equation 3.

$$KW(G_1, G_2) = \max_{\mu}(G_1) \cap \max_{\mu}(G_2) \tag{3}$$

Where $\max_{\mu}$ represents the μ higher values, G_1 and G_2 are the graphs that represent two different documents and finally $KW(G_1, G_2)$ is the set of common keywords between G_1 and G_2. Given that w is the number of times that a relation between two words (a, b) appears on the text, to find the distance between the nodes that represent these words the Equation 4 is applied.

$$D_{a,b} = \{\frac{1}{w_{a,b}}\} \tag{4}$$

Then we use the Equation 5 to find the neighborhood.

$$R = \{F_{\rho}(L_{kw_1}) \amalg \cdots \amalg F_{\rho}(L_{kw_{|L_{kw}|}})\} \tag{5}$$

Where $F_{\rho}(L_{kw_j}) = \{n \in G_1, G_2 : D(n, L_{kw_j}) \leq \rho\}$, D denotes the shortest distance between the node n and L_{kw_j} applying the Dijkstra algorithm, n are all the nodes which distance D is shorter than a radio ρ.

Subsequently, instead of obtaining a comparison coefficient per each pair of documents as the authors do in [5], we perform the comparison between them following the Equation 6, obtaining the comparison vectors B which are the union of the keywords in common plus the neighborhood of these, keeping like this more information than just a coefficient. This is performed for every document inside each category.

$$B = R \cup L_{kw} \tag{6}$$

TABLE I. TABLE OF CLASSIFICATION PERCENTAGES WITH 4 CATEGORIES.

	Method 1				Method 2			
threshold ϕ>3000								
$kw = 15$, $\rho = 2$, grade $k = 2$								
baby	73	0.75	5.75	20.5	89.25	0.5	1.25	9
dvd	2.75	80	4	13.25	7.25	81.25	4.5	7
software	3.75	4	78.25	14	6.5	2.5	75.25	15.75
toys_&_games	10.25	5	4	80.75	34	2.25	2.25	61.5
$kw = 10$, $\rho = 3$, grade $k = 2$								
baby	73.25	2	2.25	22.25	85	2.25	1.25	11.5
dvd	2	86.5	1.25	10.25	3.75	89.5	1.25	5.5
software	4.75	7	72	16.25	6	6	71.5	16.5
toys_&_games	9.25	6	1.5	83	32.25	5	2	60.75
threshold ϕ>10000								
$kw = 15$, $\rho = 2$, grade $k = 2$								
baby	66	1.5	8.25	24.25	67.75	1.25	8.75	22.25
dvd	2.25	80	3.75	14	2.25	70.25	9.5	18
software	5.25	6	73.75	15	6	1.25	76.5	16.25
toys_&_games	8	4.75	4.75	82.5	6.75	2	5.75	85.5
$kw = 10$, $\rho = 3$, grade $k = 2$								
baby	64.25	3	3.25	29.5	67	3.5	3.25	26.25
dvd	2	84.75	1.25	12	2.25	78.5	4	15.25
software	7	10	65.25	17.75	14.25	6.25	57	22.5
toys_&_games	7	6.5	2.5	84	7.25	3	2.5	87.25

d) **Feature vectors per category**

To obtain the representative vectors $\Gamma_1, \Gamma_2, \dots \Gamma_n$ where n is the number of categories, we considered to apply the intersection of the vectors B; this concept was initially considered to obtain the common IDs of all vectors, but because of the low probability of a word being considered a keyword and also appear in every document inside a category, this idea was dismissed, also because in the experimentation step the results of the intersection came to be 0 or the size of the resulting vector was too small.

Instead of this, we obtain the occurrence frequencies δ of each word of the dictionary of all vectors B. This frequencies vector is then ordered in a decreasing way to obtain the words with higher frequencies according to a threshold ϕ which is passed by parameter.
Finally, the resulting vectors Γ are obtained using the Equation 7, each vector will represent a category and will contain the IDs of their more representative words.

$$\Gamma = B_1\{id_{i=1}, \dots, id_\phi\} \cup \dots \cup B_n\{id_{i=1}, \dots, id_\phi\} \quad (7)$$

Where n is the number of the obtained feature vectors.

e) **Feature vector of a new document**

In order to obtain the feature vector Z of a new document, first we do the preprocessing and graph obtaining steps. Then, each ID will be placed as a position of the vector as it is shown in Equation 8 to then perform the classification method.

$$Z = \{id_1, id_2, \dots, id_t\} \quad (8)$$

Where t is the total number of obtained IDs of the document.

f) **Classification Method**

Once obtained the vector Z of the new document and the representative vectors Γ of each category, we find the intersection of this vector with all the vectors Γ, to get this way the belonging grade X with each category. Then, to obtain X two methods are proposed.

a) **Method 1**: X is the number of elements of the intersection between Z and Γ.

$$X = \sum_{i=0}^{n(Z \cap \Gamma)} 1 \quad (9)$$

b) **Method 2**: X is the sum of the frequencies of the words in Γ that are in the intersection with Z.

$$X = \sum_{i=0}^{n(Z \cap \Gamma)} \delta(Z \cap \Gamma)_i \quad (10)$$

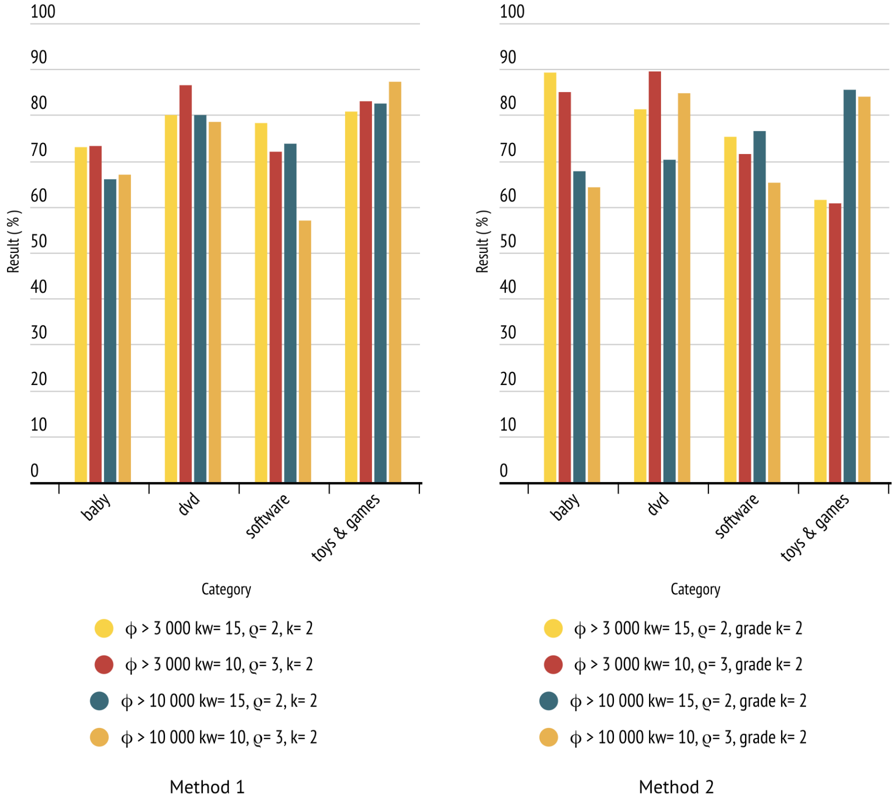
By last, the category to which the new document will belong to, will be the one with which it obtained the higher belonging grade X.

V. EXPERIMENTS AND RESULTS

For the experimentation phase, we used the amazon database [22], from which we randomly chose 4 categories and 8000 documents, being 2000 per category. The set of documents was then divided in 1600 training documents and 400 testing documents. After this the next steps were performed:

First we get the vectors B from the training set. Then, by analyzing the obtained results we can assign the value of the threshold ϕ, which controls how many IDs will be extracted for the classification method. The results of the category vectors B showed values of δ superior to 3000 and 10000 becoming these the assigned values to the threshold ϕ to then get the representative vectors per category Γ.

Next, in the Tables I and II, the values of the diagonals represent the percentage of correct classified documents as well as the error percentage, this is to say, documents assigned



Method 1

Method 2

Fig. 5. Bar Chart of the Percentages of the Correctly Classified Documents in Table I, using Method 1 and 2, where each Bar represents a Different Experiment.

TABLE II. Table of classification percentages with 3 categories.

	Method 1			Method 2		
threshold ϕ>3000						
$kw = 15, \rho = 2$, grade $k = 2$						
baby	92	1.5	6.5	97.25	0.75	2
dvd	5.5	89.25	5.25	8.75	86	5.25
software	10	6.25	83.75	10.75	6	83.25
$kw = 10, \rho = 3$, grade $k = 2$						
baby	94.5	2.75	2.75	94.25	3	2.75
dvd	3.5	95	1.5	4	94	2
software	10.5	11.75	77.75	9	11.25	79.75
threshold ϕ>10000						
$kw = 15, \rho = 2$, grade $k = 2$						
baby	87.25	2	10.75	80	3.25	16.75
dvd	4.5	89.5	6	2.75	80.5	16.75
software	11	7.5	81.5	8.5	5.5	86
$kw = 10, \rho = 3$, grade $k = 2$						
baby	90	5.75	4.25	86.25	7	6.75
dvd	3.5	94.25	2.25	2.75	89.25	8
software	12.5	14	73.5	17.5	13.25	69.25

to an incorrect category; for this we assigned different input parameters like $\rho = 2$ and $\rho = 3$, keywords number of $kw = 15$ and $kw = 10$, and grade $k = 2$.

Fig. 5 and 6 show the bar charts of the percentages of correctly classified documents, which are shown in the diagonals of the Tables I and II. We can observe that in Fig. 5, the results achieved using Method 1 with different input parameters ϕ, ρ, and k, tend to have less variation between them in most categories in comparison with the results obtained with Method 2. We can note that this behavior persists if we vary the number of categories, as showed in Fig. 6.

Also, in Fig. 5 we can see that the results of Method 2 were higher in some experiments in comparison to Method 1, these results vary if we change the input parameters, for example, the results of the category *baby* differ from 67% up to 89.25% as shown in Table I.

VI. Conclusions

In this paper, we presented a text document classification method based on a similarity comparison approach, which adapts concepts taken from the analysis of non-rigid tridimensional models and uses graphs as the structure to represent

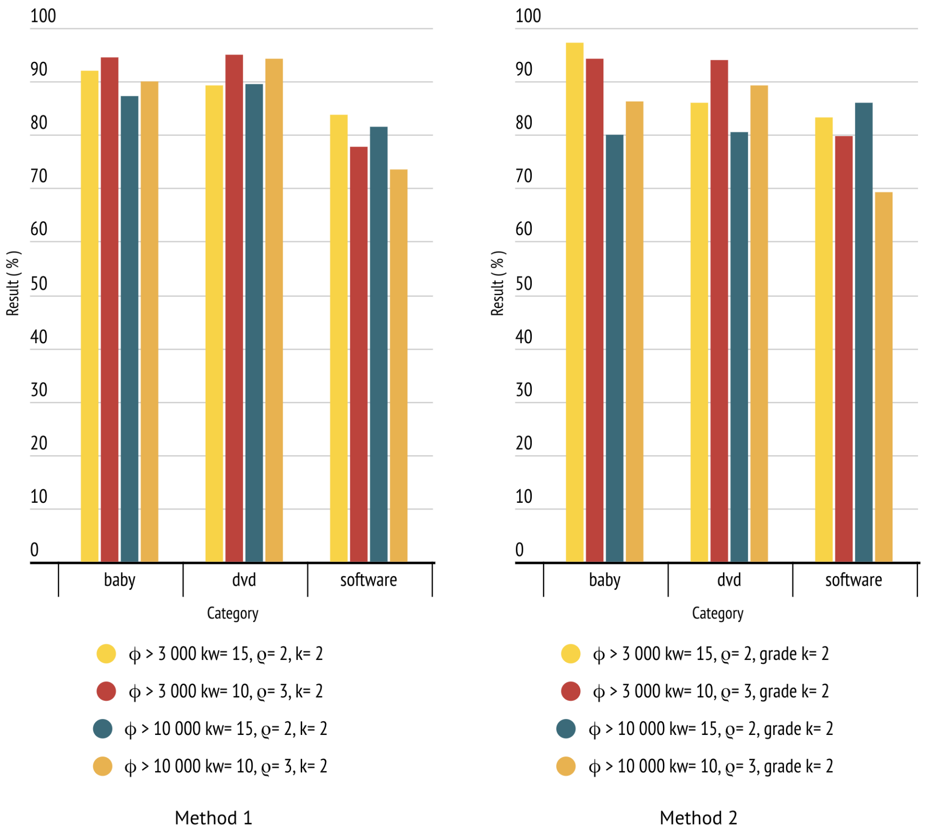


Fig. 6. Bar Chart of the Percentages of the Correctly Classified Documents in Table II, using Method 1 and 2, where each Bar represents a Different Experiment.

such documents. The method proved to have average results from 75.5% up to 78.6% of correctly classified documents with 4 categories and 82.1% up to 89.3% with 3 categories. We can observe that when performing the experiments without the category of toys toys_&_games, which generated a higher error percentage, the percentage of correct classified documents increases.

Furthermore, by the time of getting the comparison vectors B per category, their sizes can be different as well as the size of the representative vector Γ, because unlike the vector space model, in this method it would not be necessary to complete elements inside these vectors to unify their sizes according to the dictionary of words.

Worth noting that the obtained words in this representative vectors Γ are those who keep more information about the category.

Acknowledgment

The authors would like to thank the support and subvention of UNIVERSIDAD NACIONAL DE SAN AGUSTÍN DE AREQUIPA with contract No TP-7-2018-UNSA.

References

[1] Z. Wu, H. Zhu, G. Li, Z. Cui, H. Huang, J. Li, E. Chen, and G. Xu, "An efficient wikipedia semantic matching approach to text document classification," *Information Sciences*, vol. 393, pp. 15–28, 2017.

[2] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[3] C. Jiang, F. Coenen, R. Sanderson, and M. Zito, "Text classification using graph mining-based feature extraction," *Knowledge-Based Systems*, vol. 23, no. 4, pp. 302–308, 2010.

[4] G. Nikolentzos, P. Meladianos, F. Rousseau, Y. Stavrakas, and M. Vazirgiannis, "Shortest-path graph kernels for document similarity," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1890–1900.

[5] L. Castillo G., G. Dávila G., and C. López Del Alamo, "A new graph-based approach for document similarity using concepts of non-rigid shapes," pp. 41–46, june 2017.

[6] C. Paul, A. Rettinger, A. Mogadala, C. A. Knoblock, and P. Szekely, "Efficient graph-based document similarity," in *International Semantic Web Conference*. Springer, 2016, pp. 334–349.

[7] H. Dutagaci, C. P. Cheung, and A. Godil, "Evaluation of 3d interest point detection techniques via human-generated ground truth," *The Visual Computer*, vol. 28, no. 9, pp. 901–917, 2012.

[8] C. J. L. Del Alamo, L. A. R. Calla, and L. J. F. Pérez, "Efficient approach for interest points detection in non-rigid shapes," in *Computing Conference (CLEI), 2015 Latin American*. IEEE, 2015, pp. 1–8.

[9] G. N. Chandrika and E. S. Reddy, "An efficient filtered classifier for classification of unseen test data in text documents," in *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE, 2017, pp. 1–4.

[10] L. Ge and T.-S. Moh, "Improving text classification with word embedding," in *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1796–1805.

[11] A. Jain and J. Mandowara, “Text classification by combining text classifiers to improve the efficiency of classification,” *International Journal of Computer Application (2250-1797)*, vol. 6, no. 2, 2016.

[12] R. Jindal and S. Taneja, “Ranking in multi label classification of text documents using quantifiers,” in *Control System, Computing and Engineering (ICCSCE), 2015 IEEE International Conference on*. IEEE, 2015, pp. 162–166.

[13] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” *arXiv preprint arXiv:1607.01759*, 2016.

[14] Y.-S. Lin, J.-Y. Jiang, and S.-J. Lee, “A similarity measure for text classification and clustering,” *IEEE transactions on knowledge and data engineering*, vol. 26, no. 7, pp. 1575–1590, 2014.

[15] V. Korde and C. N. Mahender, “Text classification and classifiers: A survey,” *International Journal of Artificial Intelligence & Applications*, vol. 3, no. 2, p. 85, 2012.

[16] K. R. Gee and D. J. Cook, “Text classification using graph-encoded linguistic elements.” in *FLAIRS Conference*, 2005, pp. 487–492.

[17] W. Wang, D. B. Do, and X. Lin, “Term graph model for text classification,” in *International Conference on Advanced Data Mining and Applications*. Springer, 2005, pp. 19–30.

[18] F. Rousseau, E. Kiagias, and M. Vazirgiannis, “Text categorization as a graph classification problem,” 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)*, vol. 1, 2015, pp. 1702–1712.

[19] K.-A. Sohn, T.-S. Chung *et al.*, “A graph model based author attribution technique for single-class e-mail classification,” in *Computer and Information Science (ICIS), 2015 IEEE/ACIS 14th International Conference on*. IEEE, 2015, pp. 191–196.

[20] F. D. Malliaros and K. Skianis, “Graph-based term weighting for text categorization,” in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 2015, pp. 1473–1479.

[21] B. Li, Q. Yan, Z. Xu, and G. Wang, “Weighted document frequency for feature selection in text classification,” in *Asian Language Processing (IALP), 2015 International Conference on*. IEEE, 2015, pp. 132–135.

[22] J. Blitzer, M. Dredze, and F. Pereira, “Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification,” in *Proceedings of the 45th annual meeting of the association of computational linguistics*, 2007, pp. 440–447.

Predicting Number of Hospital Appointments When No Data Is Available

Harold Cáceres[1], Nelson Fuentes[2], Julio Aguilar[3]
Cesar Baluarte[4], Karim Guevara[5], Eveling Castro-Gutierrez[6],
Omar U. Florez[7]
Universidad Nacional de San Agustín de Arequipa
Arequipa, Perú

***Abstract*—Usually, in a hospital, the data generated by each department or section is treated in isolation, believing that there is no relationship between them. It is thought that while one department is in high demand, it can not influence that another may have the same demand or not have any demand. In this paper, we question this approach by considering information from departments as components of a large system in the hospital. Thus, we present an algorithm to predict the appointments of departments when data is not available using data from other departments. This algorithm uses a model based on multiple linear regression using a correlation matrix to measure the relationship between the departments with different time windows. After running our algorithm for different time windows and departments, we experimentally find that while we increase the extension of a time window and learn dependencies in the data, its corresponding precision decreases. Indeed, a month of data is the minimum sweet spot to leverage information from other departments and still provide accurate predictions. These results are important to develop per-department health policies under limited data, an interesting problem that we plan to investigate in future works.**

***Keywords*—Multi linear Regression; hospital appointments; machine learning; correlation matrix**

I. Introduction

Usually, hospital data is treated as a single entity in which appointment information, resource management, and activities are treated equally. Therefore, when the hospital offers the number of appointments available for the next month just analyze the historical data for the same department without considering the data of other departments since it is known they don't have any type of relationship between them.

There have been several studies documenting the various aspects of non-attendance in hospitals to improve the scheduling of available appointments every month just using the historical data in each department.

In this paper, we question this approach by considering information from departments as components of a large system. By doing this, we take advantage of the particular dynamics between departments that explain the system behaviour. Particularly, we observe that not all department data has the same data availability but they complement each other over time. For example, before a mother gives labor, the Emergency department shows busy schedules, but when the labor is done, usage of the Pediatrics department shows more intensity. Knowing how the different departments of a hospital interact allow us to predict the number of appointments for a target department because the other departments can explain their behavior over time. In other words, when one department does not have information, we still want to be able to make predictions using the data from other departments.

[1] found that is possible to use high-dimensional models varying in complexity based in logistic regression. The models were trained and evaluated achieving a good performance in the prediction of schedule hospital attendance.

In this paper, we experimentally confirm that it is possible to predict appointments for the next month under no available data for a target department, a critical problem in most hospitals. When the time window goes beyond a month, the predictions are not reliable and some appointment information for the target department is needed.

The rest of the article is organized as follows: Section 2 describes the related works proposed in the literature. Section 3 summarizes some previous concepts needed to understand the proposed prediction model. Section 4 explains the general scheme of the proposed model. Section 5 describes the results of applying the proposed prediction model to a health appointments dataset collected at a hospital. Finally, Section 6 concludes our work and presents some future extensions of the proposed model.

II. Related Works

Nelson A. et al. [1] found that in the United Kingdom the cost "no-show" for appointments, where the patients did not present, is around 1 000 000 000 annuals. Their purpose was to find the relation between predictive models and predictive features for "no-show" appointments. They got data from University College Hospital and Neurology and Neurosurgery National Hospital, these data had a cleaning process, after was divided into three groups: for training, validation and test for a neural network to predict "no-show" appointments. Their investigation showed that an optimal schedule appointment requires high-dimensional models based on machine learning.

Dashtban and Li [2] explain that "no-show" appointments drive to worst attention for patients, inefficient use of human resources, and an increase in waiting time. They wanted to predict the behavior's patients finding common factors. They used SSDAE (Sparse Stacked Denoising Autoencoders) for rebuilding missing data, and added a layer for making predictions, the oldest data was used for training and newest for validation. Their model surpassed other models that were compared.

Kyambille and Kalegele [3] said that Tanzania's patients complain for the time to go to the hospital and be attended, they developed a mobile application to manage appointments, they hope this application reduces the waiting time in patients.

Mieloszyk [4] faces the problem that "no-show" appointments do not allow rescheduling to other patients in this space. Their objective is to develop a system to collect data of appointments, these data were classified into three groups: relational to a patient, exam, and appointment schedule; over it, they used linear regression.

Tenagyei and Kwadwo [5] focus on the manual schedule appointments and the problems it carries on. They developed a system where patients and doctors can schedule appointments, balancing the patient's charge in doctors.

Mazurowski and Maciej [6] increasing class imbalance in the training dataset generally has a progressively detrimental effect on the classifier test performance measured by AUC and 0.9 AUC. This is true for small and moderate size training datasets that contain either uncorrelated or correlated features. In the majority of the analyzed scenarios backpropagation provided better results. The training was more susceptible to factors such as class imbalance, small training sample size, and a large number of features. Again, this finding was true for both correlated and uncorrelated features.

Class imbalance is a common problem with most medical datasets [7]. Most existing classification methods tend not to perform well on minority class examples when the dataset is extremely imbalanced. Sampling strategies have been used to overcome the class imbalance problem by either oversampling or under-sampling. Many researchers proposed different methods of over-sampling and under-sampling the majority class sample to balance the data.

Zia, Uswa Ali, and Naeem Khan. [8] they used classification algorithms Naïve Bayes, Decision Trees, and kNN for prediction diabetes. The result obtained from this study is compared with the similar study of other authors. From the comparison table, we have noticed the decision trees work better than others. The decision tree algorithms i.e. J48 and Jgraft outperform other classifiers and previous studies. It achieves the highest accuracy rate of 94.44. The decision tree is simple and a good classifier for predicting diabetes.

Data Mining is gaining its popularity in almost all applications of the real world. One of the data mining techniques i.e., classification is an interesting topic to the researchers as it accurately and efficiently classifies the data for knowledge discovery. Decision trees are so popular because they produce human-readable classification rules and are easier to interpret than other classification methods. Frequently used decision tree classifiers are studied and the experiments are conducted to find the best classifier for Medical Diagnosis. The experimental results show that CART is the best algorithm for the classification of medical data. It is also observed that CART performs well for classification on medical data sets of increased size [9].

In [10] used two large datasets, they found that DNA rates for medical appointments declined monotonically over the week. This pattern was present for both male and female patients and in all age groups but was stronger in younger age groups. Importantly, it also generalized across national hospital and single practice settings. In line with their predictions, attendance was systematically higher on days that elicit emotionally positive associations (e.g. Friday), and lower on days that elicit emotionally negative associations (e.g. Monday). These findings raise the possibility that medical appointments may be harder to face on some weekdays than on others.

Green, Linda V., and Sergei Savin. [11] Health care practices are increasingly competing not only on cost but also on quality and patient satisfaction. In this environment, timely access to care has become a more important issue. As a result, physician practices are eager to embrace new approaches to patient appointment scheduling to reduce backlogs, increase productivity, and improve patient satisfaction. They have demonstrated that the cancellation factor and its associated rescheduling probability have a significant impact on system performance and on the maximum patient panel size that can be reasonably handled by a practice. While no model is a perfect representation of reality, they believe that these are useful for guiding patient panel decisions because they capture the essential dynamics of a patient appointment system.

Almuhaideb [12] talks about the "no-show" appointments as a global problem. The data collected was from the year 2014 and used the trees JRip17 and Hoeffding for model and classified the appointments. The model uses the "no-show" appointments historical to predict future "no-show" appointments for a particular patient, the model generated by JRip 17 has 13 rules and resembles a decision tree.

III. Previous Concepts

A. Multiple Linear Regression (MLR)

Regression models are used to describe relationships between variables by fitting a line into the observed data. Regression allows the estimation of how a dependent variable changes in accordance to the changes of an independent variable(s).

Multiple linear Regression, also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of MLR is to model the linear relationship between the explanatory variables and the response variable.

In this study, MLR is used to formulate the problem of predicting the number of appointments for a department using information from the rest of hospital departments when no information is available for that specific department as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

Where Y represents the number of appointments for the next month and X represents the appointments for other departments.

B. Root Mean Square Error (RSME)

The root mean square deviation or root mean square error is a frequently used measure of the differences between the values predicted by an estimator and the observed values. RMSE is always non-negative, and a value of 0 would indicate a perfect fit to the data. In general, a lower RMSE is better than a higher one.

We used RMSE to measure the precision of the prediction as follows:

$$rmse = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2} \qquad (2)$$

In every time window, an average of RMSE was calculated by each department. Those values were used to make the comparison when the algorithm is using or not the own department data to make the prediction.

IV. Predicting Number of Appointments

A. Algorithm

The algorithm uses different time windows to generate the RMSE matrix for each department in the defined time windows. In each time window, it extracts data from the matrix to train the model and predict the next value. Thus, our input is a matrix with 41 rows representing the number of departments and 72 columns representing the number of months in 6 years.

Suppose that the time window value is 3. Then, the algorithm begins in the first position which represents January 2014 until March 2014. In each iteration of the time window, the algorithm extracts data from the matrix generating a new sub-matrix. The size of this sub-matrix is defined by the value of the time window. This time window will be advanced by until we reach row number 72, which is the number of rows in the matrix.

This matrix will be used to train the model. Once the model is trained a new input needs to be sent that will be used to predict the number of appointments for the next month. The input will be the appointments for April 2014 and the model will predict the appointments for May 2014. For the next iteration, we need to create a different subset of data moving one the time window. Meaning that the input in the next iteration to train the model will be from February 2014 until April 2014 which represents the index 1 - 4 respectively.

The next input will be the appointments for May 2014 and the algorithm will predict the appointments for June 2014. This is the main idea of the algorithm that will run for each time window defined. The algorithm flow is shown in figure 1.

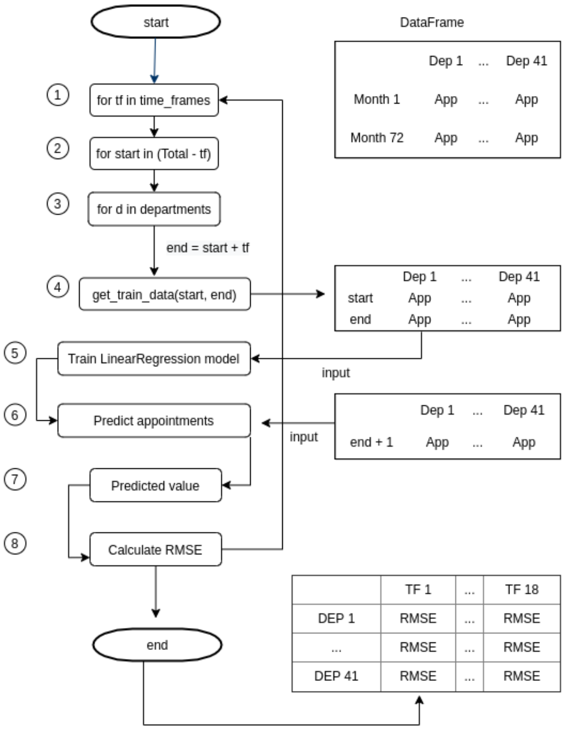


Fig. 1. Process which Represents the Algorithm Appointments Predictions

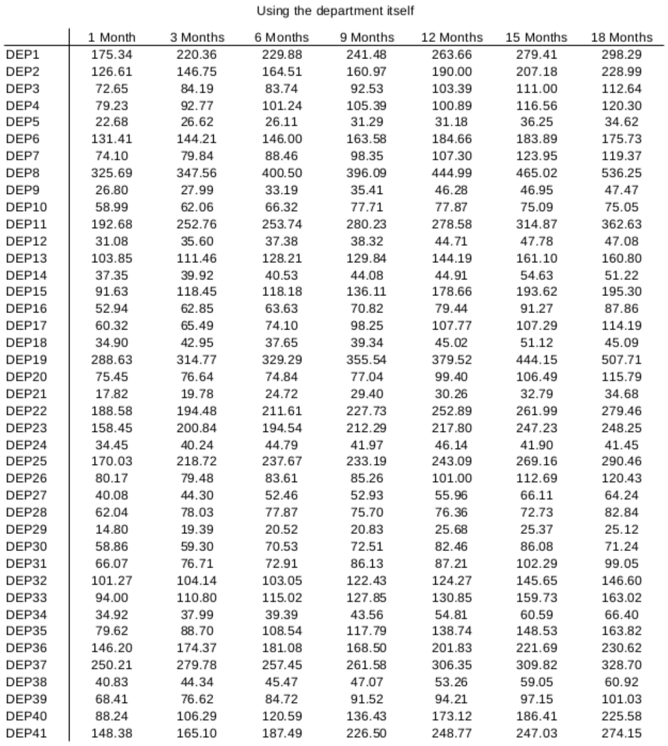
In step 1, seven time frames were defined to measure the precision in each of them. The algorithm goes over each time window(1, 3, 6, 12, 15, 18). Then, in step 2 the range for the next iteration was defined to prevent iterating more than 72 times since the data has just 72 rows. That loop goes from zero until total departments less the value of the time window. The third loop in step 3 goes over the 41 departments.

In step 4 generates a matrix using the start variable defined in step 2 and the end variable that was calculated using the start value plus the value of the time frame. The new data frame obtained in step 4 will be used as the input for the Multi linear Regression model in step 5. The algorithm was used in two different scenarios. Thus, the data obtained in step 4 were different for each of them. To predict the number of appointments for a given department using data from all departments including the data from the department to be predicted the input was 41 departments to train the model. For the second scenario where we want to predict the number of appointments for a certain department without using the data of the department that we want to predict the appointments, the entry was 40 departments.

Once the model is trained, in step 6 calculates the prediction for the next month. So, the input for the model was the data of the previous month that we want to predict. The prediction in step 7 will be used to calculate the RMSE between the real value and the value predicted. Those RMSE were saved for each iteration that will be used to get an average and save them in a table where each column represents the time window used and every row represents the departments.

The output of this algorithm is a matrix that shows the average RMSE for all the iterations for each time window. Figure 2 shows the results using all the departments and the figure 3 shows the results without using the data of the department itself.

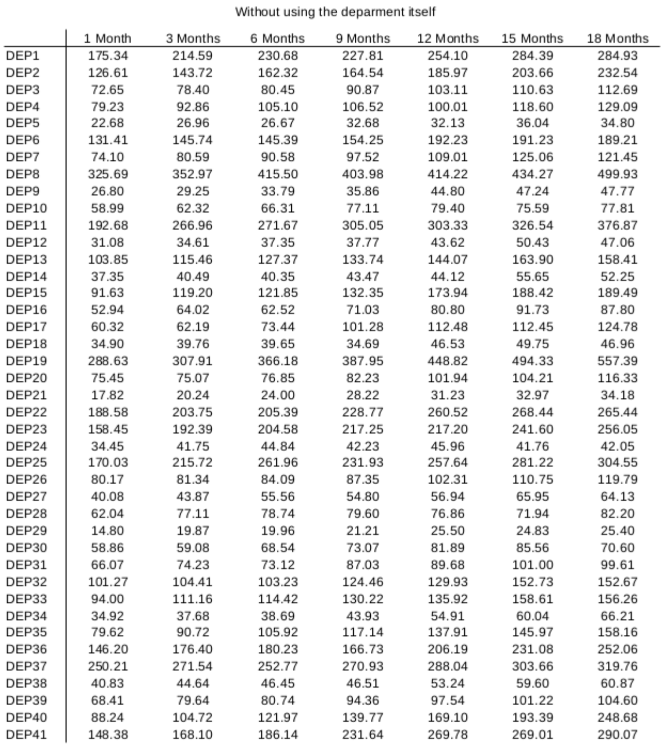
We observed while the value of the time window is lower the RMSE is pretty much the same in both scenarios. Thus, when the time window is equal to 1, the values in both scenarios are the same. It means that it is possible to predict the appointments for one department when the data is not available for the target department using the data from other



Using the department itself

	1 Month	3 Months	6 Months	9 Months	12 Months	15 Months	18 Months
DEP1	175.34	220.36	229.88	241.48	263.66	279.41	298.29
DEP2	126.61	146.75	164.51	160.97	190.00	207.18	228.99
DEP3	72.65	84.19	83.74	92.53	103.39	111.00	112.64
DEP4	79.23	92.77	101.24	105.39	100.89	116.56	120.30
DEP5	22.68	26.62	26.11	31.29	31.18	36.25	34.62
DEP6	131.41	144.21	146.00	163.58	184.66	183.89	175.73
DEP7	74.10	79.84	88.46	98.35	107.30	123.95	119.37
DEP8	325.69	347.56	400.50	396.09	444.99	465.02	536.25
DEP9	26.80	27.99	33.19	35.41	46.28	46.95	47.47
DEP10	58.99	62.06	66.32	77.71	77.87	75.09	75.05
DEP11	192.68	252.76	253.74	280.23	278.58	314.87	362.63
DEP12	31.08	35.60	37.38	38.32	44.71	47.78	47.08
DEP13	103.85	111.46	128.21	129.84	144.19	161.10	160.80
DEP14	37.35	39.92	40.53	44.08	44.91	54.63	51.22
DEP15	91.63	118.45	118.18	136.11	178.66	193.62	195.30
DEP16	52.94	62.85	63.63	70.82	79.44	91.27	87.86
DEP17	60.32	65.49	74.10	98.25	107.77	107.29	114.19
DEP18	34.90	42.95	37.65	39.34	45.02	51.12	45.09
DEP19	288.63	314.77	329.29	355.54	379.52	444.15	507.71
DEP20	75.45	76.64	74.84	77.04	99.40	106.49	115.79
DEP21	17.82	19.78	24.72	29.40	30.26	32.79	34.68
DEP22	188.58	194.48	211.61	227.73	252.89	261.99	279.46
DEP23	158.45	200.84	194.54	212.29	217.80	247.23	248.25
DEP24	34.45	40.24	44.79	41.97	46.14	41.90	41.45
DEP25	170.03	218.72	237.67	233.19	243.09	269.16	290.46
DEP26	80.17	79.48	83.61	85.26	101.00	112.69	120.43
DEP27	40.08	44.30	52.46	52.93	55.96	66.11	64.24
DEP28	62.04	78.03	77.87	75.70	76.36	72.73	82.84
DEP29	14.80	19.39	20.52	20.83	25.68	25.37	25.12
DEP30	58.86	59.30	70.53	72.51	82.46	86.08	71.24
DEP31	66.07	76.71	72.91	86.13	87.21	102.29	99.05
DEP32	101.27	104.14	103.05	122.43	124.27	145.65	146.60
DEP33	94.00	110.80	115.02	127.85	130.85	159.73	163.02
DEP34	34.92	37.99	39.39	43.56	54.81	60.59	66.40
DEP35	79.62	88.70	108.54	117.79	138.74	148.53	163.82
DEP36	146.20	174.37	181.08	168.50	201.83	221.69	230.62
DEP37	250.21	279.78	257.45	261.58	306.35	309.82	328.70
DEP38	40.83	44.34	45.47	47.07	53.26	59.05	60.92
DEP39	68.41	76.62	84.72	91.52	94.21	97.15	101.03
DEP40	88.24	106.29	120.59	136.43	173.12	186.41	225.58
DEP41	148.38	165.10	187.49	226.50	248.77	247.03	274.15

Fig. 2. RMSE Average for Each Time Window using the Department Itself



Without using the deparment itself

	1 Month	3 Months	6 Months	9 Months	12 Months	15 Months	18 Months
DEP1	175.34	214.59	230.68	227.81	254.10	284.39	284.93
DEP2	126.61	143.72	162.32	164.54	185.97	203.66	232.54
DEP3	72.65	78.40	80.45	90.87	103.11	110.63	112.69
DEP4	79.23	92.86	105.10	106.52	100.01	118.60	129.09
DEP5	22.68	26.96	26.67	32.68	32.13	36.04	34.80
DEP6	131.41	145.74	145.39	154.25	192.23	191.23	189.21
DEP7	74.10	80.59	90.58	97.52	109.01	125.06	121.45
DEP8	325.69	352.97	415.50	403.98	414.22	434.27	499.93
DEP9	26.80	29.25	33.79	35.86	44.80	47.24	47.77
DEP10	58.99	62.32	66.31	77.11	79.40	75.59	77.81
DEP11	192.68	266.96	271.67	305.05	303.33	326.54	376.87
DEP12	31.08	34.61	37.35	37.77	43.62	50.43	47.06
DEP13	103.85	115.46	127.37	133.74	144.07	163.90	158.41
DEP14	37.35	40.49	40.35	43.47	44.12	55.65	52.25
DEP15	91.63	119.20	121.85	132.35	173.94	188.42	189.49
DEP16	52.94	64.02	62.52	71.03	80.80	91.73	87.80
DEP17	60.32	62.19	73.44	101.28	112.48	112.45	124.78
DEP18	34.90	39.76	39.65	34.69	46.53	49.75	46.96
DEP19	288.63	307.91	366.18	387.95	448.82	494.33	557.39
DEP20	75.45	75.07	76.85	82.23	101.94	104.21	116.33
DEP21	17.82	20.24	24.00	28.22	31.23	32.97	34.18
DEP22	188.58	203.75	205.39	228.77	260.52	268.44	265.44
DEP23	158.45	192.39	204.58	217.25	217.20	241.60	256.05
DEP24	34.45	41.75	44.84	42.23	45.96	41.76	42.05
DEP25	170.03	215.72	261.96	231.93	257.64	281.22	304.55
DEP26	80.17	81.34	84.09	87.35	102.31	110.75	119.79
DEP27	40.08	43.87	55.56	54.80	56.94	65.95	64.13
DEP28	62.04	77.11	78.74	79.60	76.86	71.94	82.20
DEP29	14.80	19.87	19.96	21.21	25.50	24.83	25.40
DEP30	58.86	59.08	68.54	73.07	81.89	85.56	70.60
DEP31	66.07	74.23	73.12	87.03	89.68	101.00	99.61
DEP32	101.27	104.41	103.23	124.46	129.93	152.73	152.67
DEP33	94.00	111.16	114.42	130.22	135.92	158.61	156.26
DEP34	34.92	37.68	38.69	43.93	54.91	60.04	66.21
DEP35	79.62	90.72	105.92	117.14	137.91	145.97	158.16
DEP36	146.20	176.40	180.23	166.73	206.19	231.08	252.06
DEP37	250.21	271.54	252.77	270.93	288.04	303.66	319.76
DEP38	40.83	44.64	46.45	46.51	53.24	59.60	60.87
DEP39	68.41	79.64	80.74	94.36	97.54	101.22	104.60
DEP40	88.24	104.72	121.97	139.77	169.10	193.39	248.68
DEP41	148.38	168.10	186.14	231.64	269.78	269.01	290.07

Fig. 3. RMSE Average for Each Time Window without using the Department Itself

departments.

V. Results

A. Experimental Details

1) Data collection:

a) Participants: Data for this project was collected from the regional hospital located in Arequipa involving 41 departments for six years. This hospital has provided for more than 90 000 appointments from 2014 to 2019.

b) Number of departments: Forty-one departments were extracted and identified. These departments available in the hospital include nursing, pediatrics, gynecology, psychology, and other 37 departments that are included in this study. We noticed that over time in those six years, some departments were opened and closed each year. So, only the common departments over the six years were considered in our analysis.

c) Dates: Seventy-two months were extracted between January 2014 and December 2019. Where the month number 1 represents a January 2014 and month number 72 represents December 2019. For each month we have the number of appointments by each department.

2) Data processing: Data was extracted from spreadsheets. The original data that the hospital provided us was organized in folders, each folder representing a year, and within a year there was one excel per month. Thus, this data represents the overall hospital information with a total of appointments by that specific month. This data has been collected every month for over six years.

The results have the following structure: every column represents the department and every row represents the month. Departments with zero appointments in more than half of the months were removed because that data is not significant for the aim of the study. Finally, the data-set was cleaned to remove some departments that are not included in every year getting 41 columns and 72 rows.

B. Correlation Matrix

In statistics, correlation or dependence is any statistical relationship, whether casual or not, between two random variables. The most familiar measure of dependence between two quantities is the "Pearson product-moment correlation coefficient" commonly called simply "the correlation coefficient". Mathematically, it is defined as the quality of least-squares fitting the original data.

In our study, two concepts were used: strong relation and weak relation. It is assumed that the weak relationship won't affect the result and the strong relation will have a big impact on the final result. The aim is to determine if only using the departments with a high correlation coefficient will improve the prediction when data from the department itself is not used.

Figure 4 shows the coefficients between all the departments. It can be noticed that they are some of them that are pretty related and others don't have any relationships between them. Some experiments were made to use only the departments who had the correlation coefficient less 0.5 and upper 0.5 to confirm that those departments have a strong influence on the results.

After those experiments, we confirmed that using all the departments as inputs will have better results rather than just use the departments with the correlation coefficient defined above.

C. Model Validation

The model was validated by calculating the matrix of square errors for every time frame for each department. Two matrices were generated to see the differences between them since one was generated without using the department itself and the second one was generated including the data of the department.

D. With and Without Department Data

Figure 5 shows the difference of the RMSE in both scenarios that the algorithm was applied. The figure shows that for time window 1 the value is the same for both cases which confirms that it is possible to use the data of all the departments when we do not have data of the own department. The figure also shows that when the time window begins to increase, the values are not the same when the department data is used or is not used for the prediction.

Figure 6 shows the actual and predicted values using the department's data in different time windows.

Figure 7 shows the actual values and the predicted values when the department's data is not used in different time windows.

Figure 6(a) and Figure 7(a) are equal, that means, the error of the predicted values are equal. Unlike the other two graphs where we observe that they are different, therefore, the errors will also be different as shown in figure 5 in the columns of 6 and 18 months as time windows. It confirms that when one department does not have information, we still want to be able to make predictions using the data from other departments.

VI. Conclusions

The information flow in a hospital is dynamic, incomplete, but often correlated. In this paper, we discuss an algorithm to predict the appointments of departments when data is not available. We defined two scenarios to show the differences in RMSE values when the department's data is used and when is department's data is not used. After running our algorithm for different time windows and departments, we experimentally find that while we increase the extension of a time window and learn dependencies in the data, its corresponding precision decreases. Thus, the RMSE values when using the lowest time window are the same in both scenarios. Indeed, a month of data is the minimum sweet spot to leverage information from other departments and still provide accurate predictions since currently a lot of hospitals don't have the data standardized and less organized. These results are important to develop per-department health policies under limited data, an interesting problem that we plan to investigate in future works.

Acknowledgment

This research project was subsidized by Universidad Nacional de San Agustín de Arequipa, IAI-023-2018-UNSA. We would like to thank the ”Research Center, Transfer of Technologies and Software Development R + D + i” – CiTeSoft-UNSA for their collaboration in the use of their equipment and facilities, for the development of this research work.

References

[1] A. Nelson, D. Herron, G. Rees, and P. Nachev, “Predicting scheduled hospital attendance with artificial intelligence,” npj Digital Medicine, vol. 2, no. 1, pp. 1–7, Apr. 2019.

[2] RM. Dashtban and W. Li, “Deep learning for predicting non-attendance in hospital outpatient appointments,” presented at the 52nd Annual Hawaii International Conference on System Sciences (HICSS), Jan. 2019, pp. 3731–3740.

[3] G. G. Kyambille and K. Kalegele, “Enhancing Patient Appointments Scheduling that Uses Mobile Technology,” Feb. 10, 2016.

[4] R. J. Mieloszyk, J. I. Rosenbaum, P. Bhargava, and C. S. Hall, “Predictive modeling to identify scheduled radiology appointments resulting in non-attendance in a hospital setting - IEEE Conference Publication,” presented at the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

[5] Tenagyei Edwin Kwadwo , Kwadwo Kusi , Patamia Agbeshi Rutherford, “Design and Implementation of Hospital Reservation System on Android,” International Journal of Computer Science and Information Security, vol. 17, no. 10, Oct. 2019.

[6] Mazurowski, Maciej A., et al. ”Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance.” Neural networks 21.2-3 (2008).

[7] Rahman, M. Mostafizur, and D. N. Davis. ”Addressing the class imbalance problem in medical datasets.” International Journal of Machine Learning and Computing 3.2 (2013).

[8] Zia, Uswa Ali, and Naeem Khan. ”Predicting diabetes in medical datasets using machine learning techniques.” International Journal of Scientific & Engineering Research Volume 8.5 (2017).

[9] Lavanya, D., and K. Usha Rani. ”Performance evaluation of decision tree classifiers on medical datasets.” International Journal of Computer Applications 26.4 (2011).

[10] Ellis, David A., and Rob Jenkins. ”Weekday affects attendance rate for medical appointments: large-scale data analysis and implications.” PloS one 7.12 (2012).

[11] Green, Linda V., and Sergei Savin. ”Reducing delays for medical appointments: A queueing approach.” Operations Research 56.6 (2008).

[12] S. AlMuhaideb, O. Alswailem, N. Alsubaie, I. Ferwana, and A. Alnajem, “Prediction of hospital no-show appointments through artificial intelligence algorithms,” Ann. Saudi Med., vol. 39, no. 6, pp. 373–381, Nov. 2019.

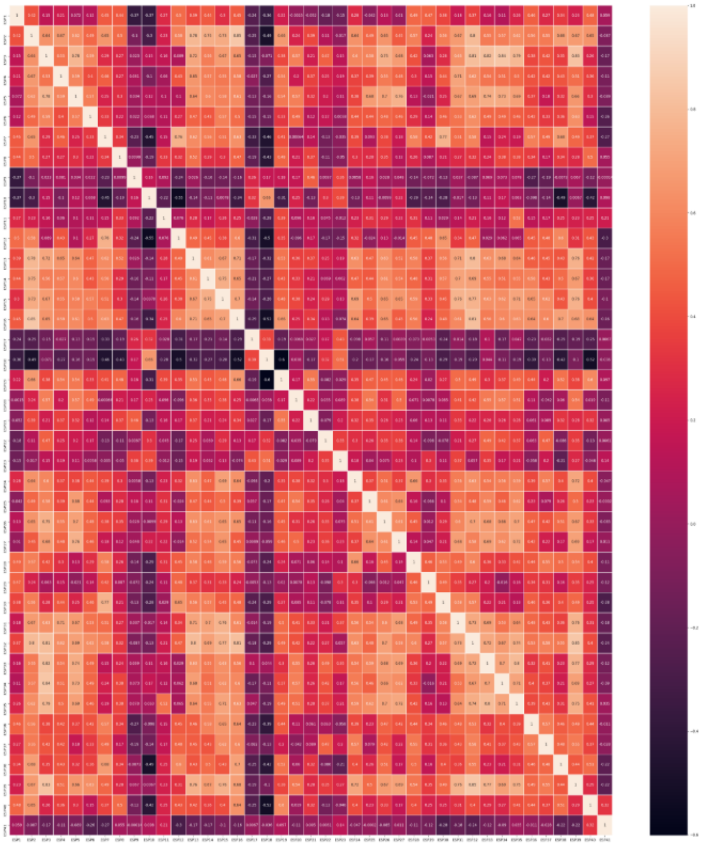
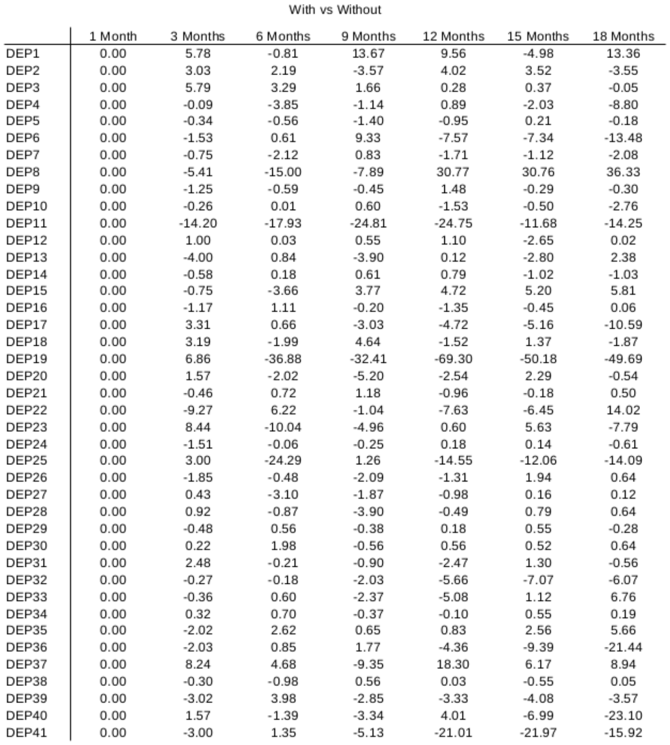


Fig. 4. Correlation Matrix between Departments. Lighter Colors Show More Positive Correlation.



With vs Without

	1 Month	3 Months	6 Months	9 Months	12 Months	15 Months	18 Months
DEP1	0.00	5.78	-0.81	13.67	9.56	-4.98	13.36
DEP2	0.00	3.03	2.19	-3.57	4.02	3.52	-3.55
DEP3	0.00	5.79	3.29	1.66	0.28	0.37	-0.05
DEP4	0.00	-0.09	-3.85	-1.14	0.89	-2.03	-8.80
DEP5	0.00	-0.34	-0.56	-1.40	-0.95	0.21	-0.18
DEP6	0.00	-1.53	0.61	9.33	-7.57	-7.34	-13.48
DEP7	0.00	-0.75	-2.12	0.83	-1.71	-1.12	-2.08
DEP8	0.00	-5.41	-15.00	-7.89	30.77	30.76	36.33
DEP9	0.00	-1.25	-0.59	-0.45	1.48	-0.29	-0.30
DEP10	0.00	-0.26	0.01	0.60	-1.53	-0.50	-2.76
DEP11	0.00	-14.20	-17.93	-24.81	-24.75	-11.68	-14.25
DEP12	0.00	1.00	0.03	0.55	1.10	-2.65	0.02
DEP13	0.00	-4.00	0.84	-3.90	0.12	-2.80	2.38
DEP14	0.00	-0.58	0.18	0.61	0.79	-1.02	-1.03
DEP15	0.00	-0.75	-3.66	3.77	4.72	5.20	5.81
DEP16	0.00	-1.17	1.11	-0.20	-1.35	-0.45	0.06
DEP17	0.00	3.31	0.66	-3.03	-4.72	-5.16	-10.59
DEP18	0.00	3.19	-1.99	4.64	-1.52	1.37	-1.87
DEP19	0.00	6.86	-36.88	-32.41	-69.30	-50.18	-49.69
DEP20	0.00	1.57	-2.02	-5.20	-2.54	2.29	-0.54
DEP21	0.00	-0.46	0.72	1.18	-0.96	-0.18	0.50
DEP22	0.00	-9.27	6.22	-1.04	-7.63	-6.45	14.02
DEP23	0.00	8.44	-10.04	-4.96	0.60	5.63	-7.79
DEP24	0.00	-1.51	-0.06	-0.25	0.18	0.14	-0.61
DEP25	0.00	3.00	-24.29	1.26	-14.55	-12.06	-14.09
DEP26	0.00	-1.85	-0.48	-2.09	-1.31	1.94	0.64
DEP27	0.00	0.43	-3.10	-1.87	-0.98	0.16	0.12
DEP28	0.00	0.92	-0.87	-3.90	-0.49	0.79	0.64
DEP29	0.00	-0.48	0.56	-0.38	0.18	0.55	-0.28
DEP30	0.00	0.22	1.98	-0.56	0.56	0.52	0.64
DEP31	0.00	2.48	-0.21	-0.90	-2.47	1.30	-0.56
DEP32	0.00	-0.27	-0.18	-2.03	-5.66	-7.07	-6.07
DEP33	0.00	-0.36	0.60	-2.37	-5.08	1.12	6.76
DEP34	0.00	0.32	0.70	-0.37	-0.10	0.55	0.19
DEP35	0.00	-2.02	2.62	0.65	0.83	2.56	5.66
DEP36	0.00	-2.03	0.85	1.77	-4.36	-9.39	-21.44
DEP37	0.00	8.24	4.68	-9.35	18.30	6.17	8.94
DEP38	0.00	-0.30	-0.98	0.56	0.03	-0.55	0.05
DEP39	0.00	-3.02	3.98	-2.85	-3.33	-4.08	-3.57
DEP40	0.00	1.57	-1.39	-3.34	4.01	-6.99	-23.10
DEP41	0.00	-3.00	1.35	-5.13	-21.01	-21.97	-15.92

Fig. 5. RMSE Comparison in Both Scenarios. Either using or not the Department's Data in Different Time Windows.

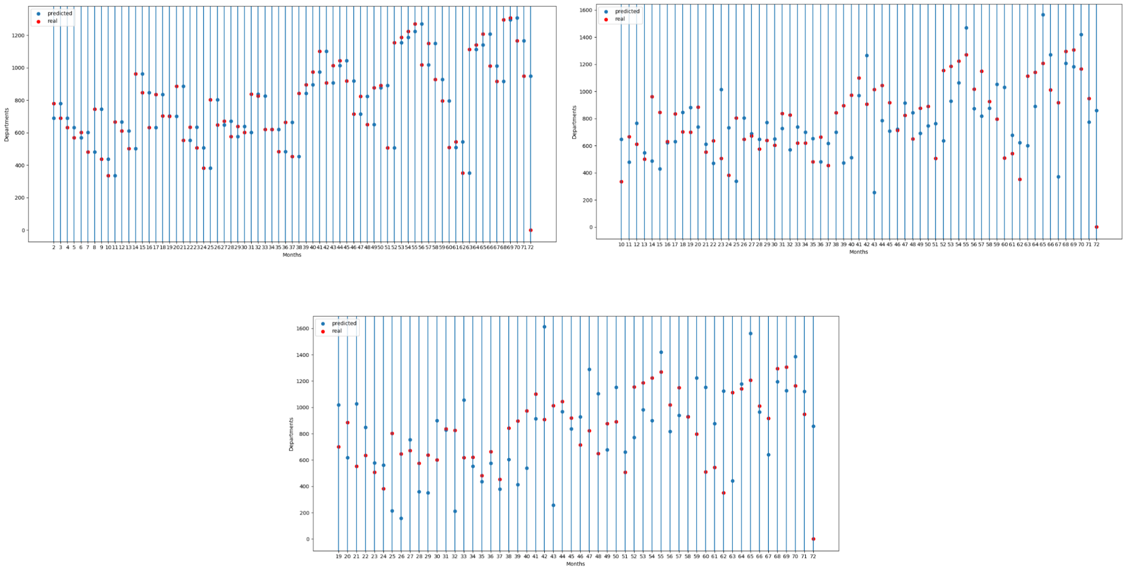


Fig. 6. Real versus the Predicted Number of Appointments using the Department's Data using Different Time Windows over the Months. (a) Time window 1 (b) Time window 9 (c) Time window 18.

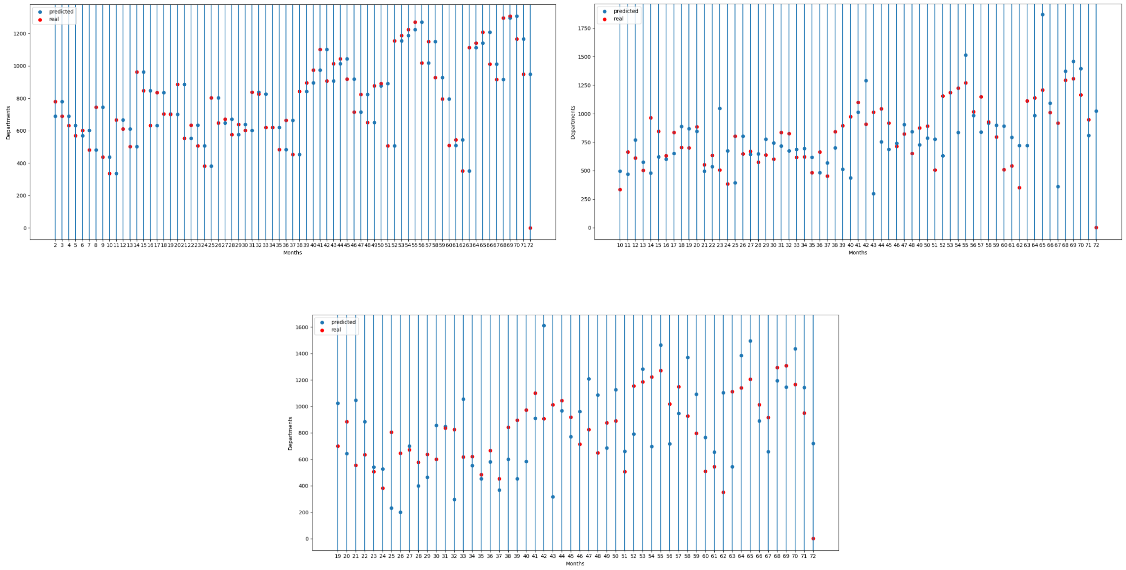


Fig. 7. Real versus the Predicted without using the Department's Data using Different Time Windows over the Months. (a) Time Window 1 (b) Time Window 9 (c) Time Window 18.

Road Damage Detection Utilizing Convolution Neural Network and Principal Component Analysis

Elizabeth Endri[1], Alaa Sheta[2], Hamza Turabieh[3]
Computer Science Department, Southern Connecticut State University, CT 06514, USA[1,2]
Department of Information Technology, Taif University, Taif, Saudi Arabia[3]

***Abstract*—Roads should always be in a reliable condition and maintained regularly. One of the problems that should be maintained well is the pavement cracks problem. This a challenging problem that faces road engineers, since maintaining roads in a stable condition is needed for both drivers and pedestrians. Many methods have been proposed to handle this problem to save time and cost. In this paper, we proposed a two-stage method to detect pavement cracks based on Principal Component Analysis (PCA) and Convolutional Neural Network (CNN) to solve this classification problem. We employed a Principal Component Analysis (PCA) method to extract the most significant features with a different number of PCA components. The proposed approach was trained using a Mendeley Asphalt Crack dataset, which contains 400 images of road cracks with a 480×480 resolution. The obtained results show how PCA helped in speeding up the learning process of CNN.**

***Keywords*—Pavement crack; Convolutional Neural Network (CNN); Principal Component Analysis (PCA)**

I. Introduction

Transportation systems depend mainly on the quality of the pavement's condition. Pavement should be able to handle traffic and environmental load for many years [1]. Subsequently, roads might be damaged over time and demonstrate distresses. To guarantee long-long-term performance and an efficient level of service, they need to be well-preserved and go through a frequent maintenance operation. Semi-automated and automated imaging-based methods are employed to provide the early detection of pavement cracks [2]. In general, roads should have good features such as shape, surface, and friction to enable users to feel safe while using them. Authorized transportation agencies are responsible for maintaining roads regularly and maintain them in good condition. In general, roads should have a prearranged schedule to keep the road safe for the public [3].

The U.S. Department of Transportation (DOT) spends billions of dollars every year for building new roads and bridges. For example, in 2018, the DOT spent more than $63 Billion on major transportation infrastructure investments throughout the USA. Meanwhile, in 2017, some unfortunate claims were reported by the American Society of Civil Engineers (ASCE) Infrastructure Report, where the roads in the USA had a "D+" grade for the road infrastructure. It was pointed out that on a scale 1 out of every five miles of highway in the USA has a bad pavement infrastructure condition [4].

One of the main difficulties of maintaining road safety is pavement crack detection, which is a challenging problem that faces road engineers all year [5]. There are several causes of pavement cracks which include poor construction, bad weather conditions, and inadequate structural support for large vehicles [6]. Traditionally, cracks have been detected through a visual process that was proven to be a tedious, time-consuming, and expensive method with an especially low rate of effectiveness. Normally, a road maintenance operator needs a great deal of related knowledge and subjectivity to deal with such problems [7]. The manual inspection is also extremely dangerous for inspectors due to traffic hazards.

Drivers are at a high risk as well. Traffic accidents are a cause of serious concern for transportation engineers and researchers. Road accidents result in significant social and economic costs. Fluctuations in the number of accidents have occurred on highways each year[8]. Some life-threatening consequences of pavement deterioration and defects are skidding, accidental driving off-road, and spontaneous maneuvering to eliminate road infrastructure problems [9] which places the driver and others at high-risk. Besides, poor surface macrotexture and microtexture could lead to hydro-planning and inconsistent tire pavement contact resulting in the reduction of tires gripping the pavement which can cause accidents [10]. Roads or pavement engineers usually inspect all types of cracks, distress, and unevenness routine manner by gathering road condition data. Gathering road data should be implemented in all weather and traffic conditions. This process may suffer from some human errors and consume time [11]. Therefore, it is important to have well-defined strategies for monitoring and maintaining roads [12].

The motivation behind this work can be described as a response to the thousands of needless deaths each year that occur due to pavement distresses all around the world. Maintenance workers put their lives on the line to perform manual inspections of roads. According to the United States Department of Labor Occupational Safety and Health Administration, out of 4,674 worker fatalities in private industry in 2017, 20.7% were in construction. In other words, one in five worker deaths last year were in construction. As a result, this motivates us to investigate the performance of the CNN method with PCA as a feature selection to detect pavement cracks inside images.

The objective of this work is to develop an intelligent approach based on CNN for road damage detection to achieve a trustworthy detection and classification of cracks from obtained 2D concrete and asphalt pavement images. This paper is organized as follows. Section II presents a literature review of the pavement crack detection research. In Section III, we discuss the PCA and CNN elements of the proposed method. A description of the dataset and the way we split the training and testing data is presented in Section VI. The experimental results based on a well-known pavement crack dataset is presented in Section VII. Finally, concluding remarks and future works are presented in Section VIII.

II. Literature Review

In the past, many research papers investigate the pavement crack problem as an image processing problem. For example, Sy *et al.* [13] applied three operations (i.e., bi-level threshold, morphological, and projection) to detect pavement cracks. The experimental data was on three kinds of images: laboratory images, static images, and AMAC reg images. Li *et al.* [14] studied this problem by proposing an approach as a thresholding method based on neighboring differential histogram statistics. Oliveira and Correia [15] handle pavement crack problems inside images by proposing a local thresholding approach based on non-overlapping blocks.

Recently, the deep convolutional neural network has been proven to show great advantages in image classification and an excellent classifier for pavement cracks. In [7], the authors extracted small patches from cracked pavement images as inputs to generate a large training database. Their proposed CNN network included 4 convolutional layers with 2 max-pooling layers and 3 fully connected layers. The proposed method had an accuracy of 91% and a recall rate of 91%. In the CrackIT project published in [16], the author used the mean and standard deviation for the unsupervised learning algorithm to distinguish blocks with cracks against blocks without cracks. They assigned severity levels to identify crack segments which relied on the computed measurement of the crack's width. The ratio between the crack segment area and the number of crack pixels belonging to the crack skeleton was computed. The results showed an accuracy of 97%, a recall score of 98.4%, and a precision of 95.5%. The drawback in their method was that they were dealing with extremely thin cracks (many of which were less than 2 mm), which proved to be a difficult task.

In [17], AlexNet created by Alex Krizhevsky used Rectified Linear Units (ReLU) instead of the $tanh$ function, which was standard at the time. ReLU's advantage over other methods is in its training time. CNN used a ReLU layer to provide a 25% error on the CIFAR-10 dataset which was six times faster than a CNN using the $tanh$ function.

In separate work, Honyan Xu *et al.* proposed an end to end crack detection based on the CNN with 28 layers, including 16 convolutional layers and earned a 90.19% test accuracy [18]. The Unmanned Aerial Vehicles (UAVs) was also used for crack inspection and monitoring. In [19], the authors proposed to simulate the pre-trained deep learning models with transfer learning methods to detect the pavement cracks based on UAV images of civil infrastructure. They employed small and complex UAV images for training and validation phases. The obtained results show that the accuracy of the proposed methods was 90% in finding cracks in practical situations with no need for augmentation and pre-processing.

In [20], the author introduced a CNN model structure to solve the crack detection and classification problems. They used a digital camera to collect images of various resolutions (i.e., 32 × 32 and 64 × 64). Two CNN networks were trained based on image resolution to detect if there was a crack or not. To achieve the second goal, the authors converted the image to binary ones with two types of crack, transverse, and longitudinal. The output from the first stage was feed to a second CNN to classify the type of crack. The finding was interesting since the images with low resolutions provided a higher classification accuracy. For 32 × 32 resolution images, the recall, precision, and accuracy calculated was 98.0%, 99.4%, and 99.2% respectively for crack and non-crack detection, while the performance for classification (i.e., transverse and longitudinal) reached the accuracy of 98% and 97%.

III. Methods

Developing an intelligent and trustworthy detection model based on CNN to detect cracks inside 2D concrete and asphalt pavement images is the main objective of this paper. The adopted database of concrete and asphalt pavement has images obtained by a 2D area digital scanning method. In this section, we shall describe the adopted methodology to solve the pavement crack detection problem.

A. Preprocessing: PCA

Principal Component Analysis commonly referred to as PCA, is a linear transformation of data. It is one of the most widely used methods of re-framing the data given [21], [22]. It measures the distances from the data to the line and tries to find the line that minimizes those distances or it can try to find the line that maximizes the distances from the projected points to the origin. It is a data transformation technique that can make it easier to use with reduction later. Data must be standardized. Dimensions will be centered around zero and have a standard deviation of 1. PCA will find a new axis, or a new attribute such that the data is maximized.

PCA works as a dimension reduction and data analysis tool. PCA has been applied successfully in a vast research area such as data mining, image processing, and artificial intelligence [23], [24]. PCA is one of the most well-known methods of factor analysis to project high-dimensional data (e.g., images) into low dimensional data based on a linear transformation without losing the value of original features [25]. So, the PCA method will reduce the number of variables and group these new variables into groups called *factors*, which improve the overall performance of machine learning classifiers such as execution time and memory usage.

The basic idea of PCA appeared in 1901 by Karl Pearson [26]. In 1936, Harold Hotelling [27] improved and developed the classical PCA. PCA is a method that aims in simplifying a multidimensional dataset to lower dimensions for analysis and visualization. In general, PCA works by converting the correlated feature variables into a new set of linearly uncorrelated features variables, which is called principal components. The main condition of PCA is that the number of PCA components should be less than or equal to the number of original features variables.

In this paper, we employed the PCA as a pre-processing step of the images before sending it to CNN to reduce the data size and improve the overall performance of CNN. Given a set of pavement crack image $\{x^1, x^2, x^3, ..., x^n\}$, the PCA works as follows:

- First: We calculate the covariance matrix of x and $\sum x$, using Equation 1.

$$\sum x = \frac{1}{m}\sum_{i=0}^{m}(x^i)(x^i)^T \tag{1}$$

- Second: We compute the eigenvectors of $\sum x$, and construct a matrix as shown in Equation 2, where u_1 represents the first eigenvector, u_2 represents the second eigenvector, and so on. Equation 3 shows the calculation that is used to construct the input features maps that are uncorrelated with each other.
  The covariance matrix for x_{rot} can be extracted from the diagonal matrix from U, whose diagonal elements λ_1, λ_2, λ_3,..., λ_n. Where λ_i presents corresponding eigenvalues of eigen vector matrix U.

- Finally: PCA is evaluated based on Equation 4.

$$U = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & ... & u_n \\ | & | & & | \end{bmatrix} \tag{2}$$

$$x^i_{rot} = U^T x^i \tag{3}$$

$$x_{PCA,i} = \frac{x_{rot,i}}{\sqrt{\lambda_i}} \tag{4}$$

B. What is ANN?

Artificial Neural Networks (ANNs) are computational processing systems that were inspired by how biological nervous systems function [28]. In reality, the neural system is a very complex one that consists of an extremely large number of neurons. Each neuron is designed to receive an input signal(s) from its dendrites and generate an output signal(s). the output signal(s) goes through the axon, which transfers the generated signal to the next neuron using synapses. Once a set of input signals reaches a predetermined threshold value, the neuron is triggered, which simulates the real functions inside the human brain (see Fig. 1 [29]).

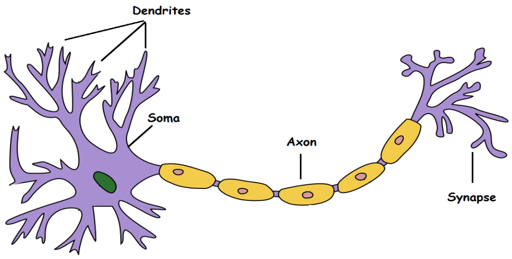


Fig. 1. A Biological Neuron [29].

ANNs consist of interconnected nodes, called neurons, that learn from given input to optimize to final output. These artificial neurons have numeric weights attached to them. These weights will be optimized through the training phase. The performance of well-trained ANN will show high performance with a piece of datum or pattern to recognize or identify (see Fig. 2). Using a suitable learning algorithm, these units are efficient in generating a function that maps a relationship between inputs and output training examples.

ANN uses a training dataset (i.e., images, row data, etc.) as input data. The input layer handles the training dataset, which is connected to the next layer (i.e., the hidden layer). The hidden layer will manipulate the data and tune the connection weights before sending it to the output layer. Each ANN should have a learning algorithm (e.g., BackPropagation, Convolutional Neural Network, Long Short-Term Memory, etc.) that tunes the ANN weights to enhance the overall performance of ANN by reducing the error between the real output (i.e., actual) and obtained output (i.e., predicted) from ANN [30], [31], [32].

$$S = \sum_{i=0}^{n} w_i x_i \tag{5}$$

$$\phi(S) = \frac{1}{1+e^{-S}} \tag{6}$$

Several tuning parameters should be designated before we can use ANN to be trained. They include the number of layers in the hidden layer, the type of sigmoid function for the neurons, and the adopted learning algorithm (see Fig. 2 [33])

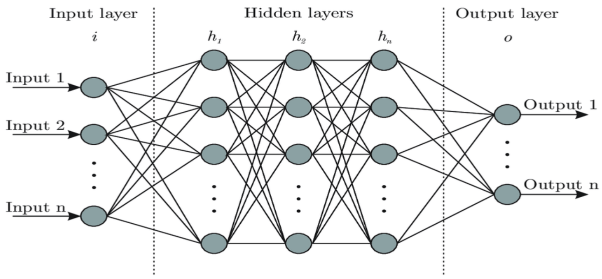


Fig. 2. Fully Connected ANN Architecture [33].

IV. What is CNN?

The earliest CNN model called the leNet-5 model was proposed by LeCun in 1998 [34]. CNN can be thought of as a close family member of the traditional ANN. The main structure of CNN is motivated by the discovery of the visual cortex in the brain, which contains a large number of cells that detect the light in the small receptive fields, and overlapping sub-regions of the visual field (see Fig. 3). These cells act as local filters over the input space, and the more complex cells have larger receptive fields.

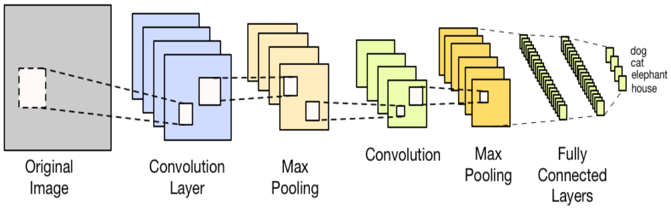


Fig. 3. An Overview of a CNN model.

Therefore, a simple CNN is a series of layers, and every layer of a CNN converts one volume of activations to another through a differentiable function. We use three main types of layers to build CNN architectures: Convolutional Layer, Pooling Layer, and Fully-Connected Layer (exactly as seen in regular Neural Networks). We will stack these layers to form a full CNN architecture.

A. *Convolution Layer*

Traditional neural networks are fully connected in every layer, while convolutional layers in CNNs use the convolutional operation [35]. The convolution layer in CNN operates the function that is performed by the cells in the visual cortex. The neurons in CNNs are self-optimize through learning. Each neuron receives input and operates. CNN's have an input layer, various hidden layers, and an output layer. These hidden layers use a mathematical model to pass on results to the following layer.

Convolution is the first layer used to extract features from an input image and preserves the relationship between pixels by learning image features using small tiles of input data [1], [36]. Essentially, the convolutional layer is a mathematical equation that takes two inputs such as an image matrix and a filter or kernel. Convolution uses a small square matrix, which preserves the spatial relationships among pixels, to learn image features [37]. The convolution layer can do quite a few operations with different filters including edge detection, blurring, and sharpening an image [38].

The convolution layer is the essential component of a convolutional neural network. The convolution layer includes of a set of independent filters. Each filter is individually convolved with the image, and feature maps are obtained. In general, if we convolve an image of size $N \times M$ with a filter of size $l \times k$, we get an output feature map of size $O_{width} \times O_{height}$ as given in Equation 7.

$$\begin{aligned} O_{width} &= \frac{N - l + 2p_l}{s_l} + 1 \\ O_{height} &= \frac{M - k + 2p_k}{s_k} + 1 \end{aligned} \tag{7}$$

where p_l and p_k are the padding in both width and height, respectively, and s_l and s_k represent the stride in both horizontal and vertical directions. Thus, if we apply a convolution operation with a filter of size 5 × 5 on an image with a size 32 × 32, the result will be a feature map of size 28 × 28, with a zero-padding and stride of one. The output feature map is acquired by the convolution of the input maps with a linear filter, adding a bias term and then applying a nonlinear function. The output can be generally denoted by the formula as in Equation 8

$$X_j^q = f(\sum_{I \in I_j} X_i^{q-1} \times W_{ij}^q + b_j^q) \tag{8}$$

where q represents the layer number, W_{ij} represents the convolutional kernel, b_j represents bias, I_j represents the set of input maps and $f(.)$ represents the activation function. Fig. 4 shows the output features collected from the third layer after doing the feature extraction.

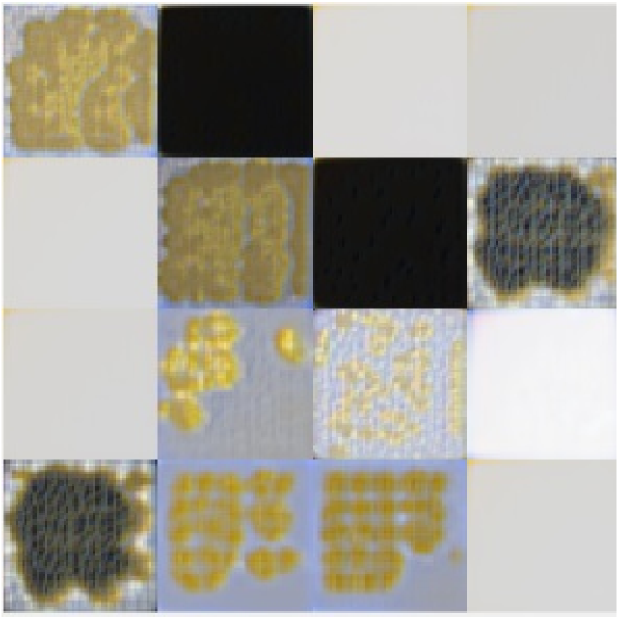
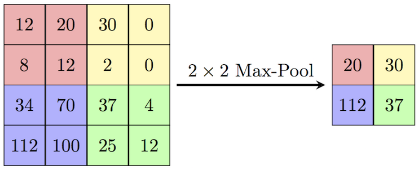


Fig. 4. The Output of the Third Convolution Layer.

B. *Polling Layer*

The main objective of the pooling layer is to decrease the spatial size of the representation, which will enhance the overall performance of the neural network, the number of CNN parameters, and reduce the probability of overfitting. In general, the pooling layer is located between successive convolutional layers on CNN. The pooling layer operates on sliding a two-dimensional filter over each channel of feature map and summarizing the features lying within the region covered by the filter.

The pooling layers in the CNN are used to reduce the number of parameters when the images are too large (see Fig. 5). Spatial pooling reduces the dimensionality of each map but keeps important information. It also can control over-fitting. There are two types of pooling 1) max and 2) average pooling. Max pooling is defined as a sample-based discretization process. The advantage of Max-Pooling is a massive edge detection-based matrix multiplication.



12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

20	30
112	37

Fig. 5. An Example of a Max Pooling Operation.

C. *Rectified Linear Unit*

The activation function is a very important element for the CNN design such that it can learn and perform more complex tasks. Activation functions are a nonlinear function utilized to the input. Several frequently used activation functions in the literature are sigmoid, logistic activation function, tanh, and hyperbolic tangent activation function. In this work, we are adopting a ReLU activation function. ReLU stands for the Rectified Linear Unit for a nonlinear operation. The rectified linear activation function is defined as a piecewise linear function. This function can produce either the same function input if the input is position and zero if the input is negative. Fig. 6 shows the ReLU function. Equations 9 and 10 demonstrate the computations of ReLU.

$$R(z) = max(0, z) \tag{9}$$

$$R(z) = \begin{cases} z & \text{if } z \leq 0, \\ z & \text{if } 0 < 0. \end{cases} \tag{10}$$

ReLU helps to backpropagate the errors and have multiple layers of neurons being triggered by the ReLU function. ReLU helps to overcome the vanishing gradient problem and allows models to learn faster. ReLU is widely recommended to use in CNN classification models [39].

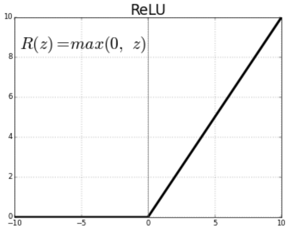


Fig. 6. Rectified Linear Unit (ReLU) Activation Function.

D. SoftMax Unit

SoftMax is another activation function like sigmoid, tanh, and ReLU. They are commonly used for the neurons in the output of the fully connected. It's defined as:

$$\sigma(z_i) = \frac{e_i^z}{\sum_{j=1}^{C} e_j^z} \tag{11}$$

where C is the number of classes, z is the input vector, and $\sigma(z_i)$ is the output class probability. SoftMax function provides a discrete probability distribution over all the given classes. The SoftMax function output is a probabilities $p_i \in [0, 1]$. The sum of the probability of all classes C is $\sum p_i = 1$.

E. Fully-Connected Layer

The fully connected (FC) means that every single neuron in the preceding layer is connected to every single neuron on the current layer. Each neuron shall have a summation followed by an activation function. The final layer of CNN is a FC layer that has a FC to all activation functions in the previous layer, as observed in the traditional ANN. These activation functions are used to compute the CNN final output via a matrix multiplication followed by a bias offset. In the CNN, the FC layer merges all the features obtained from the previous convolutional and sub-sampling layers.

F. Evaluation Metrics

The evaluation metrics that will be used in this paper are accuracy, precision, and recall. These metrics were chosen because they are commonly used in classification problems. They are defined as the following:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

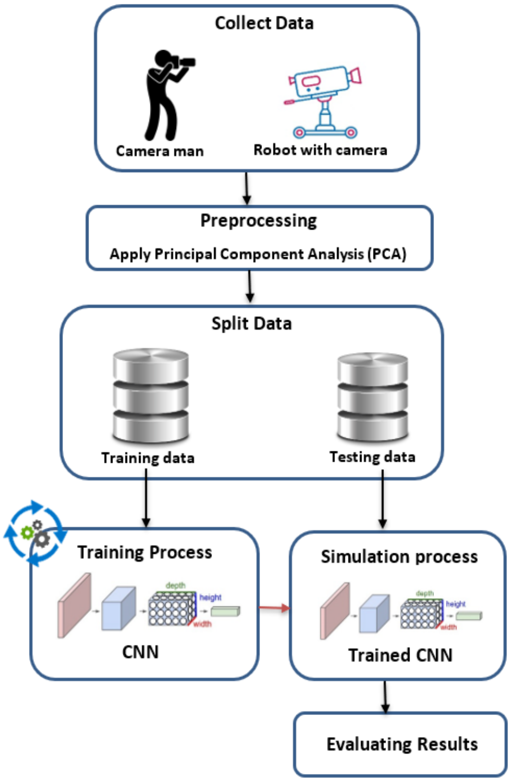


Fig. 7. Proposed CNN based-Detection Method.

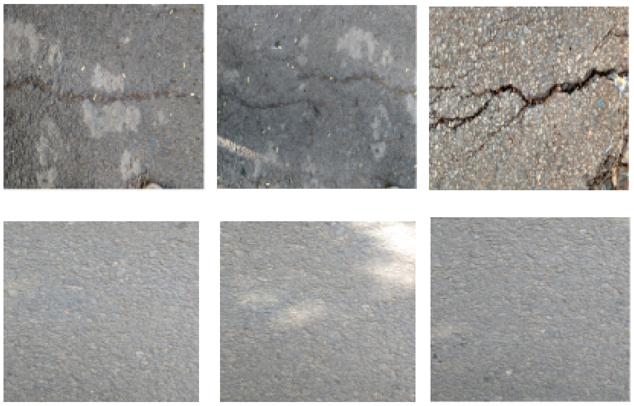


Fig. 8. Mendeley Asphalt Crack Dataset. Upper Row: Image Cracks, Lower Row: No Crack.

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

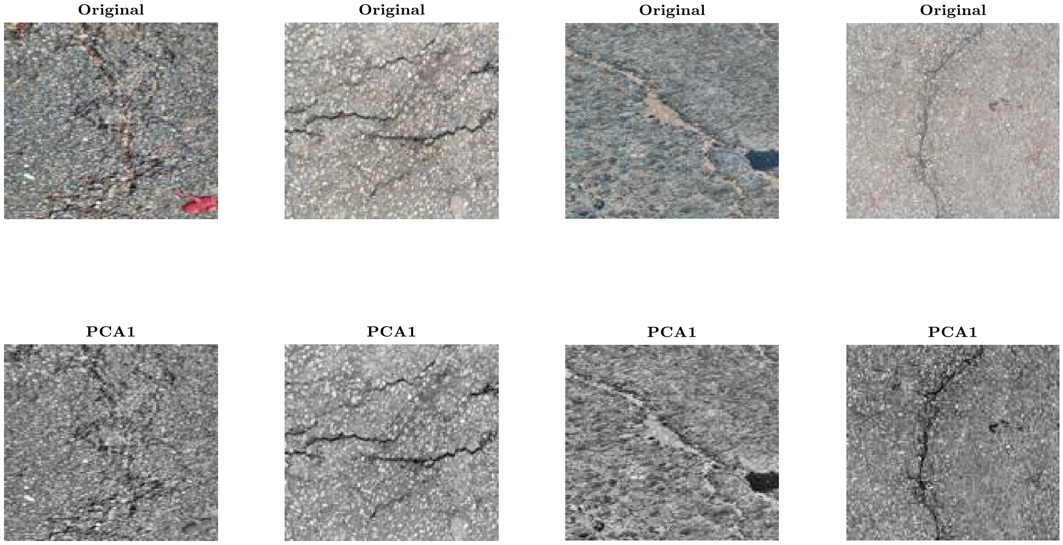
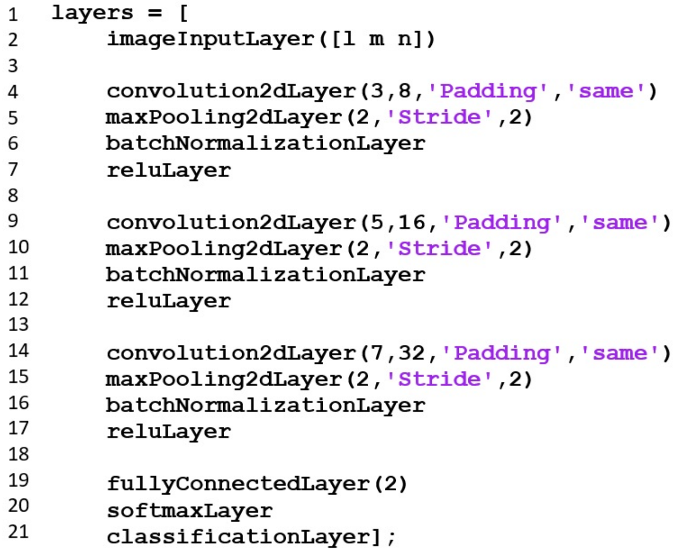


Fig. 9. First Row: Sample Images and Second Row: PCA-Crack Images.



```
layers = [
    imageInputLayer([l m n])

    convolution2dLayer(3,8,'Padding','same')
    maxPooling2dLayer(2,'Stride',2)
    batchNormalizationLayer
    reluLayer

    convolution2dLayer(5,16,'Padding','same')
    maxPooling2dLayer(2,'Stride',2)
    batchNormalizationLayer
    reluLayer

    convolution2dLayer(7,32,'Padding','same')
    maxPooling2dLayer(2,'Stride',2)
    batchNormalizationLayer
    reluLayer

    fullyConnectedLayer(2)
    softmaxLayer
    classificationLayer];
```

Fig. 10. Layer Structure for CNN Model.

TABLE I. OBTAINED RESULTS FOR TRAINING DATASET.

	PCA Component	Best	Worst	Avg.	Std.
Without PCA	—	97.78	85.02	93.13	2.92
With PCA	2	98.06	91.62	95.60	1.98
	5	95.08	90.11	93.18	1.51
	10	97.67	94.34	**96.62**	0.93
	15	98.13	93.75	96.03	1.11

where TP, FP, TN and FN are the true positive, false positive, true negative, and false-negative, respectively.

V. PROPOSED CNN-BASED METHOD

The proposed method used in this work is depicted in Fig. 7. The proposed method is a combination of PCA and CNN methods. In the first step, we collect data for pavement cracks using a traditional method (i.e., cameraman) or an intelligent method (i.e., a robot with a camera). After collecting data, it is important to analyze the most valuable features by extracting image features using the PCA method. The proposed method will enhance the performance of CNN convergence.

VI. DATASET

A pavement crack dataset called the Asphalt Crack dataset is used in this work. The data consists of 400 images. It is a contribution by Jayanth Balaji, Thiru Balaji, Dinesh M S, Binoy Nair, and Harish Ram D.S. The dataset was published on April 26, 2019 [40]. Fig. 8 depicts samples of the dataset (i.e., cracked and not cracked pavements).

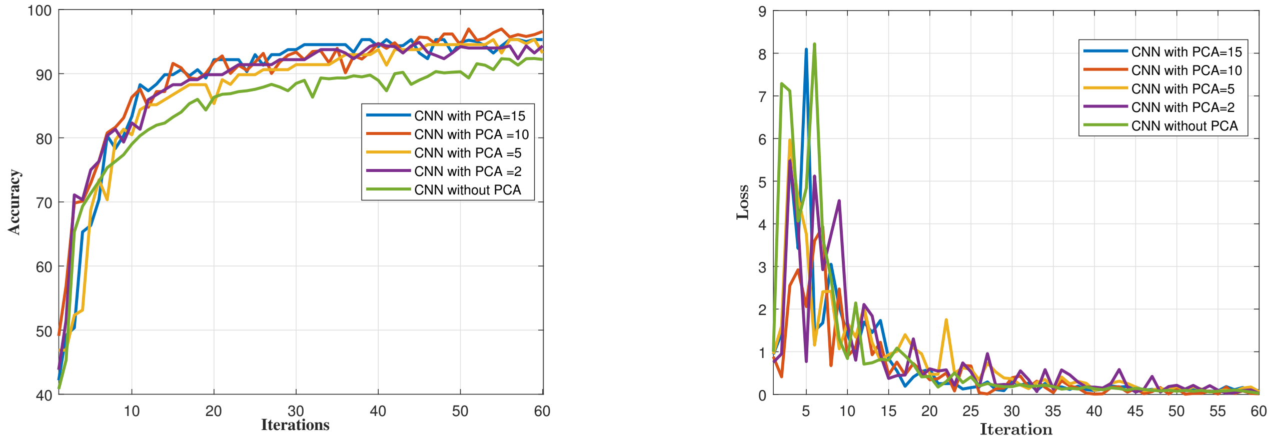
A. Training and Testing Data

The hold-out method is the simplest kind of cross-validation method. The main basic idea of the hold-out method is to divide them into two groups: Training and Test/Validation dataset. The training dataset is used to train the model (i.e., CNN) and evaluate the trained model using the Test/Validation dataset. This approach of splitting data is applicable to image processing applications. We adopted the classical holdout method for splitting data.

VII. EXPERIMENTAL RESULT

The proposed method for pavement crack detection was simulated on an Intel Core i7-7700HQ 2.8-GHz processor with 16 GB RAM and implemented using MATLAB R2019b environment [41]. First, we preprocessed the crack images using PCA. We created a set of feature images from the original crack images data set as given in Fig. 9. These are the first features created from PCA. These images were used as input to the CNN for further processing.

A sample code that shows the CNN architecture is presented in Fig. 10. In this work, we employed two types of experiments: (i) without PCA, and (ii) with PCA. Moreover, we explored the performance of PCA with several numbers



(a) Average Convergence Curves.

(b) Loss Convergence Curves.

Fig. 11. Experimental Results showing Convergence Curves.

of PCA component (i.e., 2, 5, 10, and 15). For each type of experiment, we executed our program eleven times. Table I shows the obtained results for the training dataset. The performance of CNN with PCA_{10} (i.e., PCA with ten components) outperforms all other methods based on average accuracy of 96.62 and standard deviation of 0.93. Fig. 11a explores the average convergence curves of the accuracy during the training process for all trained models. The performance of CNN is improved after employing the PCA. Fig. 11b demonstrates the performance of the CNN models with and without PCA. The performance of CNN with PCA is improved and able to detect the pavement cracks robustly compared to CNN without PCA.

In Fig. 12a we show the CNN performance based on eleven runs. It was found that the performance results of CNN without PCA is not the best. Table II shows the performance of the proposed method over the testing dataset. Again it is clear that the performance of CNN with PCA_{10} outperforms all other models based on average and standard deviation. Fig. 12b shows the performance of all trained models over the testing dataset. The PCA can enhance the performance of CNN. Moreover, the performance of CNN with PCA_{10} is the most suitable method for the pavement crack detection method.

TABLE II. Obtained results for testing dataset.

	PCA Component	Best	Worst	Avg.	Std.
Without PCA	—	91.36	84.36	88.61	2.35
With PCA	2	94.36	86.23	90.44	2.48
	5	94.39	87.42	91.97	2.04
	10	95.70	91.37	**94.01**	1.13
	15	95.31	88.23	91.55	1.92

Tables III and IV report the obtained average results (i.e., Accuracy, Precision, and Recall) for training and testing datasets, respectively. The performance of CNN with PCA component equals 10 outperforms all other models based on the average results for 11 independent runs.

TABLE III. Statistical analysis for all models for Training dataset.

	PCA Component	Accuracy	Precision	Recall
Without PCA	—	93.13	0.92	0.97
With PCA	2	95.60	0.91	0.93
	5	93.18	0.93	0.96
	10	**96.62**	**0.96**	**0.98**
	15	96.03	0.94	0.97

TABLE IV. Statistical analysis for all models for testing dataset.

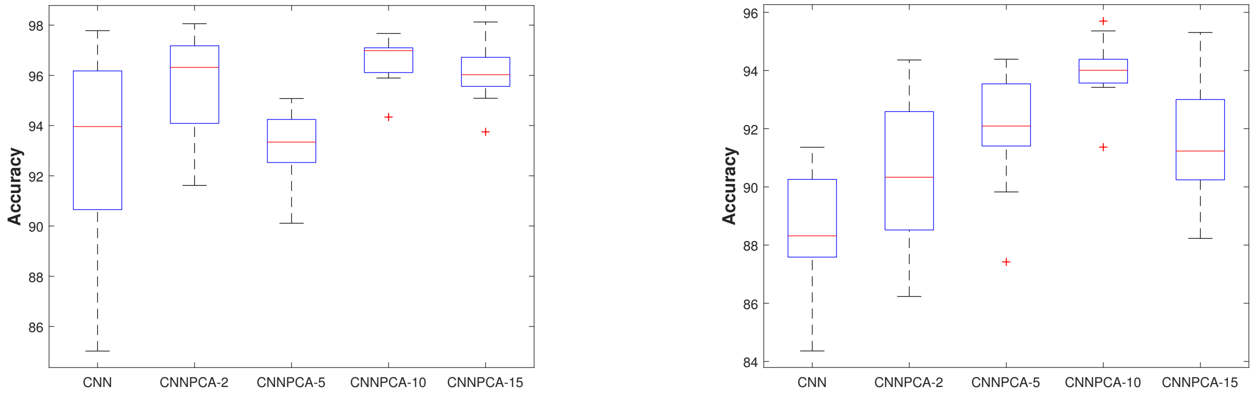
	PCA Component	Accuracy	Precision	Recall
Without PCA	—	88.61	0.82	0.84
With PCA	2	90.44	0.93	0.97
	5	91.97	0.89	0.90
	10	**94.01**	**0.94**	**0.93**
	15	91.55	0.87	0.84

To examine the obtained results and show how the number of PCA components affects the classifiers' performance, we employed a statistical analysis and comparison based on the Wilcoxon statistical test with a significance level of 0.05. Table V shows the p-values of the obtained results between a different number of PCA components. All the obtained p-values are less than 0.05, which means that there is a statistical difference between them. For example, the p-value between PCA_2 and PCA_{10} is 0.040, which means the performance of CNN is not similar.

TABLE V. P-value results based on Wilcoxon test.

	P-value
PCA_2 vs. PCA_5	0.031
PCA_2 vs. PCA_{10}	0.040
PCA_2 vs. PCA_{15}	0.027
PCA_5 vs. PCA_{10}	0.035
PCA_5 vs. PCA_{15}	0.049
PCA_{10} vs. PCA_{15}	0.039

Finally, from the obtained results, we can conclude that PCA as a feature extraction can enhance the performance of CNN. Moreover, the proposed approach can examine a huge



(a) Box-Plot for the Training Dataset.

(b) Box-Plot for the Testing Dataset.

Fig. 12. Experimental Results for the Training and Testing Dataset.

number of images automatically, which save time, cost for pavement cracks detection, and reduce risks for roads and pavements engineers.

VIII. Conclusions and Future Work

In this work, we proposed a CNN-based method to automate the pavement crack detection process. The main idea of the proposed method is to combine CNN with PCA to speed up the learning process. CNN was employed as a classification method, while PCA as a feature extraction one. We examined the performance of our proposed method on a public dataset that contains 400 images. We also explored several numbers of PCA components (i.e., 2, 5, 10, and 15). The obtained results show that CNN with PCA_{10} outperforms all other models. In future work, we will examine different parameters setting and employed an optimization algorithm such as a genetic algorithm, to optimize the PCA parameters.

Acknowledgment

The authors would like to thank the Computer Science Department, Southern Connecticut State University for the financial support of this research work.

References

[1] S. Chambon and J.-M. Moliard, "Automatic road pavement assessment with image processing: Review and comparison," *International Journal of Geophysics*, vol. 2011, 03 2011.

[2] A. Cubero-Fernandez, F. J. Rodríguez Lozano, R. Villatoro, J. Olivares, and J. Palomares, "Efficient pavement crack detection and classification," *Eurasip Journal on Image and Video Processing*, vol. 2017, 12 2017.

[3] F. Yang, L. Zhang, S. Yu, D. Prokhorov, X. Mei, and H. Ling, "Feature pyramid and hierarchical boosting network for pavement crack detection," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2019.

[4] ASCE, "2017 infrastructure report card: Roads. american society of civil engineers (asce)," https://www.infrastructurereportcard.org/wp-content/uploads/2017/01/Roads-Final.pdf, May 2017.

[5] A. Sheta, H. Turabieh, S. Aljahdali, and A. Alangari, "Pavement crack detection using a lightweight convolutional neural network," *Proceedings of 35th International Confer*, vol. 69, pp. 214–223, 2020.

[6] A. K. Schindler and B. F. McCullough, "Importance of concrete temperature control during concrete pavement construction in hot weather conditions," *Transportation Research Record*, vol. 1813, no. 1, pp. 3–10, 2002.

[7] Z. Fan, Y. Wu, J. Lu, and W. Li, "Automatic pavement crack detection based on structured prediction with the convolutional neural network," *arXiv preprint arXiv:1802.02208*, 2018.

[8] S. N. Baskara, H. Yaacob, M. R. Hainin, and S. A. Hassan, "Accident due to pavement condition–a review," *Jurnal Teknologi*, vol. 78, no. 7-2, 2016.

[9] M. Ergun, S. Iyinam, and A. F. Iyinam, "Prediction of road surface friction coefficient using only macro-and microtexture measurements," *Journal of transportation engineering*, vol. 131, no. 4, pp. 311–319, 2005.

[10] L. F. Beck, A. M. Dellinger, and M. E. O'neil, "Motor vehicle crash injury rates by mode of travel, united states: using exposure-based methods to quantify differences," *American Journal of Epidemiology*, vol. 166, no. 2, pp. 212–218, 2007.

[11] K. Gopalakrishnan, S. K. Khaitan, A. Choudhary, and A. Agrawal, "Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection," *Construction and Building Materials*, vol. 157, pp. 322 – 330, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950061817319335

[12] S. Saat, A. R. M. Kamil, M. Z. M. Tumari, and A. S. R. A. Subki, "Development of an autonomous robot for inspection system," in *2018 IEEE 14th International Colloquium on Signal Processing Its Applications (CSPA)*, March 2018, pp. 272–276.

[13] N. T. Sy, M. Avila, S. Begot, and J. C. Bardet, "Detection of defects in road surface by a vision system," in *MELECON 2008 - The 14th IEEE Mediterranean Electrotechnical Conference*, May 2008, pp. 847–851.

[14] Q. Li and X. Liu, "Novel approach to pavement image segmentation based on neighboring difference histogram method," in *2008 Congress on Image and Signal Processing*, vol. 2, May 2008, pp. 792–796.

[15] H. Oliveira and P. L. Correia, "Automatic road crack segmentation using entropy and image dynamic thresholding," in *2009 17th European Signal Processing Conference*, Aug 2009, pp. 622–626.

[16] H. Oliveira and P. L. Correia, "Crackit an image processing toolbox for crack detection and characterization," *2014 IEEE International Conference on Image Processing (ICIP)*, 2014.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105.

[18] H. Xu, X. Su, Y. Wang, H. Cai, K. Cui, and X. Chen, "Automatic bridge crack detection using a convolutional neural network," *Applied Sciences*, vol. 9, no. 14, p. 2867, 2019.

[19] K. Gopalakrishnan, H. Gholami, A. Vidyadharan, Alok, Choudhary, and A. Agrawal, "Crack detection in unmanned aerial vehicle images of civil infrastructure using pre-trained deep learning model," vol. 8, no. 1, 2018, p. 1—14.

[20] N. A. M. Yusof, M. K. Osman, M. H. M. Noor, A. Ibrahim, N. M. Tahir, and N. M. Yusof, "Crack detection and classification in asphalt pavement images using deep convolution neural network," in *2018 8th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, 2018, pp. 227–232.

[21] I. Jolliffe, *Principal Component Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1094–1096.

[22] I. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, p. 20150202, 04 2016.

[23] S. C. Ng, "Principal component analysis to reduce dimension on digital image," *Procedia Computer Science*, vol. 111, pp. 113–119, 12 2017.

[24] S. Yu, K. Yu, V. Tresp, H.-P. Kriegel, and M. Wu, "Supervised probabilistic principal component analysis," in *KDD 2006*, Max-Planck-Gesellschaft. New York, NY, USA: ACM Press, Aug. 2006, pp. 464–473.

[25] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37 – 52, 1987.

[26] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[27] H. Hoteeling, "Relations Between Two Sets of Variates," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 12 1936. [Online]. Available: https://doi.org/10.1093/biomet/28.3-4.321

[28] K.-l. Hsu, H. V. Gupta, and S. Sorooshian, "Artificial neural network modeling of the rainfall-runoff process," *Water resources research*, vol. 31, no. 10, pp. 2517–2530, 1995.

[29] W. Mcculloch and W. Pitts, "A logical calculus of ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, vol. 5, pp. 127–147, 1943.

[30] A. Nigrin, *Neural networks for pattern recognition*, ser. A Bradford book. Cambridge, Mass, London: The MIT Press, 1993. [Online]. Available: http://opac.inria.fr/record=b1126290

[31] J. Leonard and M. A. Kramer, "Improvement of the backpropagation algorithm for training neural networks," *Computer Chemical Engineering*, vol. 14, pp. 337–343, 1990.

[32] A. K. Jain, J. Mao, and K. Mohiuddin, "Artificial neural networks: A tutorial," *IEEE Computer*, vol. 29, pp. 31–44, 1996.

[33] F. Bre, J. Gimenez, and V. Fachinotti, "Prediction of wind pressure coefficients on building surfaces using artificial neural networks," *Energy and Buildings*, vol. 158, 11 2017.

[34] W.-q. Huang and Q. Wu, *Image Retrieval Algorithm based on Convolutional Neural Network: Selected Papers from CSMA2016*, 12 2017.

[35] H. D. Beale, H. B. Demuth, and M. Hagan, "Neural network design," *Pws, Boston*, 1996.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[37] A. de Bruijn, V. Muhonen, T. Albinonistraat, W. Fokkink, P. Bloem, and B. Analytics, "Detecting offensive language using transfer learning," 2019.

[38] H. H. Aghdam and E. J. Heravi, *Guide to Convolutional Neural Networks: A Practical Application to Traffic-Sign Detection and Classification*, 1st ed. Springer Publishing Company, Incorporated, 2017.

[39] M. Bhurtel, J. Shrestha, N. Lama, S. Bhattarai, A. Uprety, and M. Guragain, "Deep learning based seed quality tester," 11 2019.

[40] J. B. A, "Asphalt crack dataset," Apr 2019. [Online]. Available: http://dx.doi.org/10.17632/xnzhj3x8v4.2

[41] MATLAB, *(R2019b)*. Natick, Massachusetts: The MathWorks Inc., 2019.

Automatic Building Change Detection on Aerial Images using Convolutional Neural Networks and Handcrafted Features

Diego Alonso Javier Quispe[1], Jose Sulla-Torres[2]
Universidad Nacional de San Agustín de Arequipa
Arequipa, Perú

***Abstract*—In this article, we present a new framework to solve the task of building change detection, making use of a convolutional neural network (CNN) for the building detection step, and a set of handcrafted features extraction for the change detection. The buildings are extracted using the method called Mask R-CNN which is a neural network used for object- based instance segmentation and has been tested in different case studies to segment different types of objects obtaining good results. The buildings are detected in bitemporal images, where three different comparison metrics MSE, PSNR and SSIM are used to differentiate if there are changes in buildings, we used this metrics in the Hue, Saturation and Brightness representation of the image. Finally the characteristics are classified by two algorithms, Support Vector Machine and Random Forest, so that both results can be compared. The experiments were performed in a large dataset called WHU building dataset, which contains very high-resolution (VHR) aerial images. The results obtained are comparable to those of the state of the art.**

Keywords*—Bi-temporal images; convolutional neural network (CNN); building detection; building change detection; Mask R-CNN

I. Introduction

Building detection and change detection is a field that has been studied for a long time and attempts to solve different problems such as urban planning, cadastral updating, damage detection by natural disasters, among many others. Building change detection consists in differentiating the changes that occur over time in a building, considering that a building could be built, could be destroyed or could be modified.

Different researchers use different types of data to carry out their experiments. In their most basic form, aerial images or RGB satellite images are used, but these images have different drawbacks, starting with low resolution, perspective, lighting changes, shadows, and various variations that a building can have depends on the country, the city and the area where the images were captured. Other authors use different sensors to obtain more information such as multispectral sensors, synthetic aperture radar (SAR), light detection and ranging (LiDAR), digital surface models (DSM) and so on [1]. Using DSM allows us to obtain a 3D model of buildings, the advantage is that having altitude information allows us to better analyze changes in buildings, although the drawback is that it is difficult to obtain such information.

The related works will be divided into two sections, the works related to the detection of buildings, and those related to the detection of changes in buildings. The Morphological building / shadow index (MBI) [2] is a long-term building detection method used in different works related to building change detection [3] [4] [5] [6] [7], which consists of representing the spatial edges of the buildings in such a way that they can be distinguished from other objects. These properties are represented by brightness, size, contrast, directionality and shape. Other ways to detect the edge of a construction is through the use of shadows, Ali Ozgun [8] created a model to determine the relationship between the buildings and their shadows using a probabilistic approach. Saman Ghaffarian and Salar Ghaffarian [9] uses a different approach, instead of using the RGB image, they use another color space, the LUV, and through the FastICA algorithm they divide the image into three different regions: vegetation and shadows, firm ground and tracks, and buildings, in this way they try to avoid confusing some other object with buildings. Fadi Dornaika et al. [10] use a segmentation technique called statistical region mergin (SRM), which segments an image into small homogeneous regions based on their similar properties, considering the spectral information of shape and scale. After applying SRM they extract information using the local binary pattern (LBP) algorithm for each region segmented in the previous step. Finally they use four classifiers which are listed below: 1-NN, 3-NN, J48 and SVM, to determine if they are buildings or not buildings. Comparing these classifiers results in SVM being better than the other classifiers.

The works described so far only use machine learning and handcrafted features extraction techniques to segment buildings, but in recent years deep neural networks have obtained very good results in the field of object segmentation. Ji et al. [11], propose a deep neural network based on a neural network known as U-net, where instead of having a single input it requires two inputs so they call it Siamese U-net (SiU-Net), their proposal is compared to other networks, which are also used for segmentation of objects at the instance level, these networks are Mask R-CNN and U-net. The authors determine that SiU-net is slightly better than U-net and Mask R-CNN.

The detection of changes in buildings can be done at two different levels, detection of changes- at the pixel level and detection of changes at the object level, most of the related works were prepared based on changes at the pixel level. The work of Wen et al. [5], consists in classifying each pixel in 4 different categories: building, vegetation, water and soil. To determine the changes, each bi-temporal image is subdivided into quadrants, for each quadrant a histogram is calculated so that it can be compared with the histogram of its corresponding bitemporal image, finally the histograms are compared to determine if the quadrant is considered with changes or without changes. In addition to the four categories mentioned, a category also considered is the shadow which is taken into account by various authors [12] [7] [13], if any object generates shadow in a location where long ago there was no shadow, it is an indication of change, the disadvantage is that the shade is conditioned at the time of day and at the atmospheric condition. Considering only RGB information can be deficient, so other authors choose to use 3D information [14] [15], one of the advantages that this type of data contains is height, so that if we compare the height of buildings, it can be known if there have been changes or there have been no changes.

Deep Learning is also currently used in detecting changes in buildings. Considering that there are bi-temporal images, Debu et al.[16], propose three Fully Convolutional Neural Netwroks (FCNs) based on the concept of Siamese networks, they consider two FCN with double input. The third FCN has a single image as input, where the bitemporal images are concatenated. The results obtained are relatively low, this is because the data set contains low resolution images. Ji el al. [17] use a simple CNN for the detection of changes, their proposal is to use as input only the binary mask of the bitemporal images, and not the complete images, so that with this information the binary mask of changes is generated as output. Their experiments were carried out on a data set with a large number of buildings and an acceptable amount of changes. Furthermore, these images have a very high resolution.

In this article we propose a new framework for detecting changes in buildings using very high resolution aerial images (VHR). For the detection of buildings we use a convolutional neural network and for the detection of changes we use comparison metrics in different color spaces of the images, so that different characteristics of the buildings can be compared. Finally we use two SVM and RF classifiers to determine the buildings in which changes have been detected. The data set used is called the WHU building data set [11], which contains more than 220,000 independent buildings which vary in color and size.

This paper is organized as follows. In Section II the methodology applied for the automatic detection of changes in buildings is detailed. Section III describes the database used and presents the results obtained in the steps of building detection and building change detection with the proposed model. Finally, in Section V details the conclusions of the present paper.

II. Methodology

The general pipeline is shown in Fig. 1. First we take the bitemporal images as input, for each image a binary mask is created where it is determined if an area is a building or background, by means of a building extraction network. Next, each building is compared with its corresponding temporal image using comparison metrics as MSE, PSNR and SSIM, we use the HSV representation of each image. Finally we use two classifiers Support Vector Machine (SVM) and Random Forest (RF) to determine if there are changes in buildings.

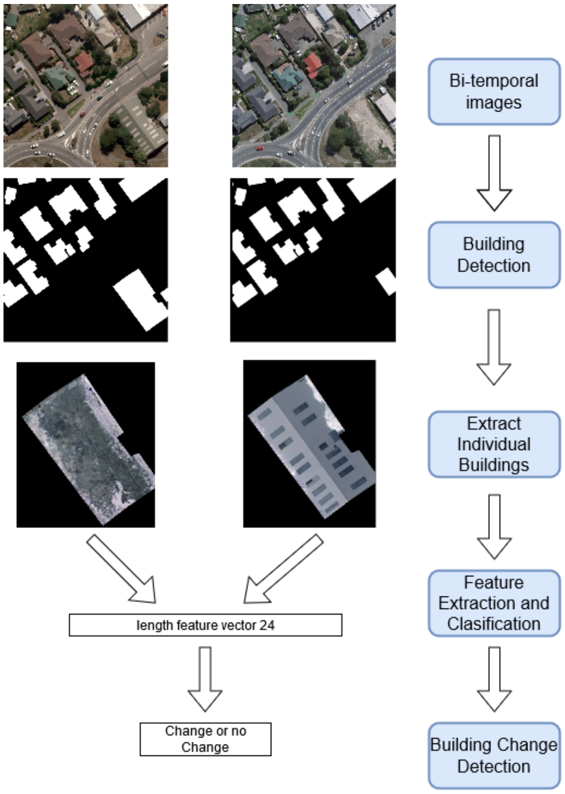

Bi-temporal images
Building Detection
Extract Individual Buildings
Feature Extraction and Clasification
Building Change Detection
length feature vector 24
Change or no Change

Fig. 1. Pipeline for Building Change Detection using a Neural Network for Building Detection and Handcrafted Features for Change Detection.

A. Building Detection

We use a convolutional neural network called Mask R-CNN [18], which is applied for object detection and we used it for detection of buildings. Based on the results obtained in [19]. This deep neural network was proposed in 2017 and is one of the most powerfull networks for object instance segmentation.

The advantages provided by this network is that it provides three different outputs for object detection, as the first output presents the classification of objects, the second output is the regression box and finally the prediction of the mask. While the classification section does not have to distinguish a large number of different objects, it must be able to differentiate a building from any other object. The main problem lies in the data because the bulding could have different sizes and shapes, also different objects can be confused as buildings, as is the case with large trucks.

In Fig. 2 shows the general architecture of Mask R-CNN, the network parameters suggested by the author were not modified.

B. Building Change Detection

1) Extract individual buildings: In an image of 512 x 512 we have different number of buildings. Each individual building is extracted, considering two cases, the first where the building has not changed and the second in the case that has changed. We obtained a total of 15005 sub - images with individual buildings.

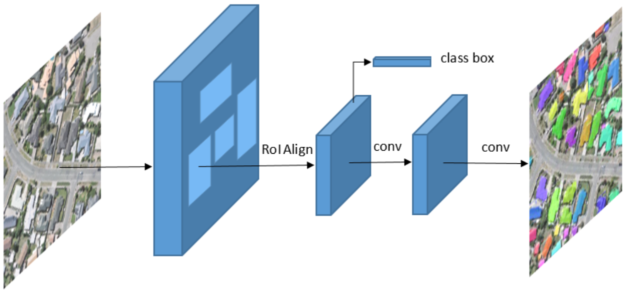


Fig. 2. Mask R-CNN Framework Arquitecture for Instance Segmentation.

2) Image Representation: To compare two images we use a different representation of the image, if we compare a RGB image with another RGB image we will get the same value for the red, green and blue bands, for that reason instead of using a RGB image we use the HSV representation where we have three different bands (Hue, Saturation, Value). In addition to using the HSV representation of the image, the grayscale representation is used.

3) Histogram of Oriented Gradient: The Histogram of Oriented Gradient (HOG) [20], is a feature descriptor that focuses on determining the shape of an object. This is determined by calculating the difference between the gradients of an image, to subsequently obtain a value and a direction for each pixel of an image until obtaining a histogram that represents its characteristics.

4) Comparison Metrics: We are going to compare the bitemporal images so that we obtain a unique value for each pair of images. We use three different comparison metrics: mse, psnr and ssim, its definition is detailed below:

a) MSE (Mean Square Error): This is the average square difference between two values. This is always no negative and if the value is very close to zero it means that there are no changes. We compare pixel by pixel the two images. The MSE formula is as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \tag{1}$$

b) PSNR (Peak Signal to Noise Ratio): It is represented as the relationship between the maximum value that a pixel can have and the noise that affects the representation of the entire image. The PSNR formula is as follows:

$$PSNR = 10\log_{10}\frac{R^2}{MSE} \tag{2}$$

The variable R corresponds to the maximum value that an image can take, depending on how it is represented, it can take the value of 1 or 255.

c) SSIM (Structure Similarity Index Method): It is used to measure the quality between two images. Considering that MSE tries to find the differences between the pixels of an image, SSIM does the opposite, comparing the pixels to determine their similarity, based on three different terms: luminance, contrast and structure. The final value is represented by multiplication of these terms. The SSIM formula is as follows:

$$SSIM = \frac{(2U_xU_y + C_1)(2\alpha_{xy} + C_2)}{(U_x^2 + U_y^2 + C_1)(\alpha_x^2 + \alpha_y^2 + C_2)} \tag{3}$$

One drawback to consider is the difference that exists between the 2012 images versus the 2016 images, due to the difference in the atmospheric conditions in which they were captured. Fig. 3 shows that there is a difference in the contrast between these images, therefore to solve this problem, a histogram equalization is applied to each representation obtained. Equalization is used as a complement and in addition, a feature vector is extracted using HOG for the gray images. Therefore we finally have 10 different representations which are the following: hue, value, saturation, gray, equalized hue, equalized value, equalized saturation, gray equalized, HOG of a gray image and HOG of a gray equalized image. Our final vector of characteristics is conformed by the concatenation of the 3 comparison metrics described (MSE, PSNR, SSIM) for each image representations obtained, by having 10 representations, our final final vector for each pair of images will have a length of 30.

5) Classifiers:

a) Support Vector Machine: The support vector machine (SVM) is a binary classifier which separates the classes into two different spaces by means of a hyperplane which is known as the support vector [21]. The characteristics can be separated in three different ways: by means of a linear nucleus for which the Euclidean distance is used to define the hyperplane, by means of a polynomial nucleus and by means of a Gaussian nucleus which is associated with the variance.

b) Random Forest: The random forest classifier consists of a large number of independent binary trees, the end result is a majority vortex [22]. Each binary tree generates a different prediction, it is expected that a set of them tends to a wrong result but a larger set of trees tends to the correct result, causing the global result to be correct. The tree mainly depends on two values, N that represents the number of trees, and p that indicates the depth of the tree .

III. Experimental Results

A. Data Set

We use the WHU building dataset proposed by [11], this dataset contains aerial and satellite images, but we only use aerial images for their high quality. This dataset consists in aereal images captured between 2011 to 2016, the area captured cover the city of Christchurch, New Zealand; have a total of 120000 buildings and covers an area between 450m^2 and 550m^2. The main drawbacks of the database are the noise caused by plants that obstruct sections of the roofs and large cars that are confused with buildings.

We divide this dataset in three sections, the first and second sections were used in the building detection step as training and validation data respectively. The third section was used as a test to building detection and all this section was used for the building change detection step.

In Fig. 3 we can see different images of the dataset, we take the same pattern proposed by the author of segmenting the image into sub-images of 512 by 512, in each sub image we can see that there are small and large buildings, as well as few changes in one image and many changes in another.

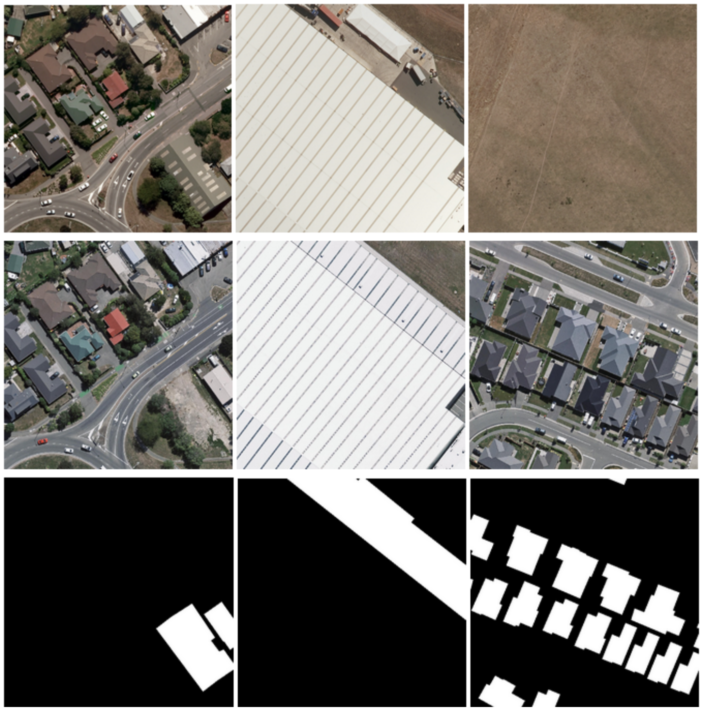


Fig. 3. Example of Database Images. First Row is Images of 2011. Second Row is Images of 2016. Third Row is Change Label.

B. Experiments on Building Detection

We used Mask R-CNN for building detection and use the same protocol used in the work of [19]. Mask R-CNN was pretrained with the COCO dataset, this converged after 30 epochs and the process took about 18 hours.

The test area consists of 1920 tiles of 512 × 512 pixels, in two different times, 2011 and 2016. Table I shows the results applying Mask R-CNN to the two bitemporal data sets to classify objects considering the buildings as objects. We compare our results with the results obtained in the work of [19].

TABLE I. Comparison of methods for building detection

Time	Method	Accuracy
2011	Mask R-CNN [19]	0.892
2011	MS-FCN [19]	0.922
2011	Our Mask R-CNN	0.866
2016	Mask R-CNN [19]	0.922
2016	MS-FCN [19]	0.939
2016	Our Mask R-CNN	0.897

Our results are similar to the results obtained in [19], although they are slightly lower, it may be because the training and validation data are generated without considering overlapping and in [19] it considers overlapping.

In Fig. 4 shows an example of building detection with their respective masks per building, it is observed that the mask is not perfect, the edges of the buildings are oval when they should be straight, this is a drawback of performing instance segmentation, due to the complexity it represents and to the different shapes and sizes that buildings can have, but this building extraction does not greatly affect the results obtained.

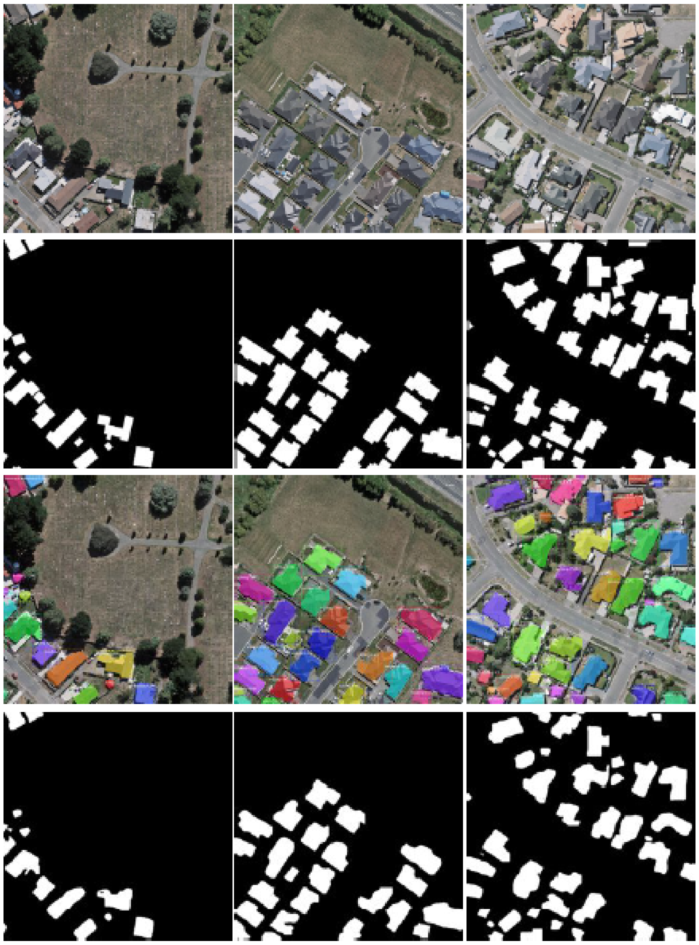


Fig. 4. Example of Building Detection. First Row is the Images of Different Zones. Second Row is the Ground Truth Masks. Third Row is the Title of the First Row with the Building Mask Detected. Fourth Row is the Detected Mask using Mask R-CNN

C. Building Change Detection Results

The binary building maps obtained in the previous step are taken and the pre-processing is carried out using the same criteria as [19], where all the buildings that are less than 500 pixels in size corresponding to a size of 4.8 x 4.8 m2 are eliminated, considering that buildings of that size are not usual, so they are considered a false detection.

Then, we extract the characteristics based on the comparison metrics MSE, PSNR and SSIM for each representation of the image, obtaining a feature vector of length 30 for each pair of images. Finally we use two classifiers, the first is Support Vector Machine for which we use a linear kernel and the second is Random Forest. For Random Forest we analyze which are the best parameters for number of trees N and the depth of each tree d, in the Fig. 5 it is observed that RF begins to converge with a value greater than 25 for N and in Fig. 6 it is observed that begins to converge with a value greater than 15 for d. For this reason we consider a value of 26 for N and a value of 16 for d.

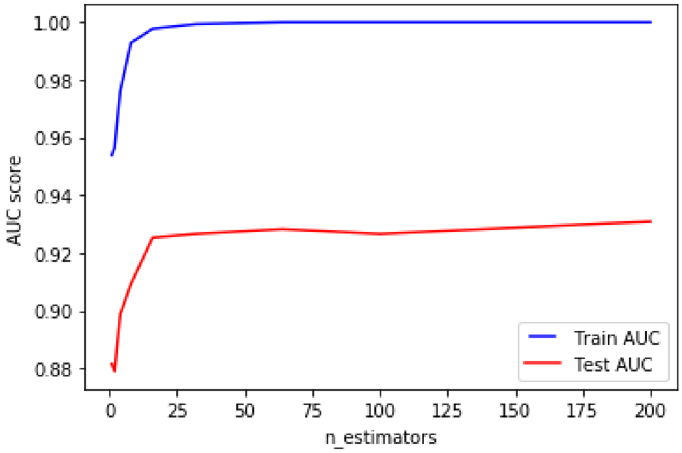


Fig. 5. Area under the Curve Considering a Number of Trees N from 1 to 200, to Random Forest.

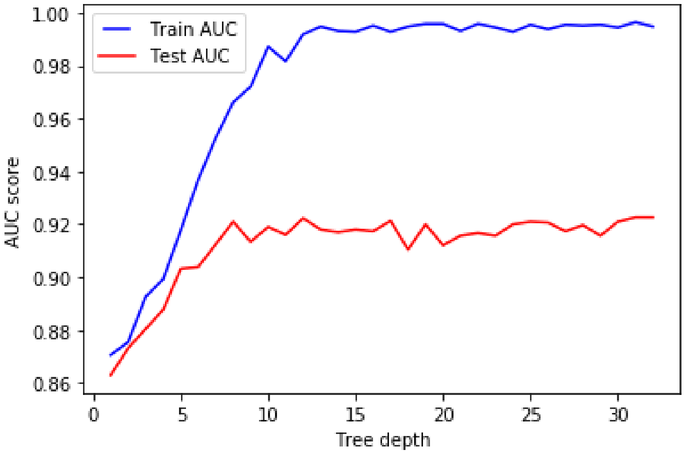


Fig. 6. Area under the Curve Considering a Depth per Tree d from 0 to 35, to Random Forest.

Table II shows a comparison of five different methods evaluated using three different metrics, for AP (counted on changed building instances), the best result is obtained with Our Mask R-CNN with SVM with a value of 0.854, in terms of recall the results obtained by almost all the methods are very similar close to 0.89 and in the same way for accuracy the results are similar close to 0.9, the FC-EF method is the only one that obtains results far below the other methods.

The method that obtains the best results in the state of the art is Mask R-CNN [17], so we can conclude that our proposal using Mask R-CNN with SVM obtains comparable results with that related work to which we obtain an accuracy of 0.911.

IV. Future Works

Explore different techniques for the detection of buildings, this step directly affects the building change detection, so an improvement in the accuracy of building detection, improve the precision of the general model proposed in this article.

Evaluate the proposed model in different data sets, in order to test its scalability.

TABLE II. Comparison of Methods for Building Change Detection.

Method	AP	Accuracy	Recall
Mask R-CNN [19]	0.814	0.910	0.883
MS-FCN [19]	0.796	0.891	0.872
FC-EF [16]	0.254	0.519	0.462
Our Mask R-CNN with SVM	0.854	0.911	0.891
Our Mask R-CNN with RF	0.852	0.891	0.899

V. Conclusions

In this article we propose a new framework for detecting changes in buildings using high-resolution images, for this we use a neural network for change detection and handcrafterd features for change detection. Experiments show that the results obtained by our proposal are comparable to those obtained in the state of the art. In this study we evaluate the detection of changes in buildings ignoring other types of objects such as bridges, tracks among others. Building detection directly affects change detection, so it is necessary to improve the precision of this step to improve the precision of the general model.

Acknowledgment

I would like to thank the Universidad Nacional de San Agustín de Arequipa for the financing provided with contract TP-015-2018, for which the development of this research work was possible.

References

[1] P. Sidike, D. Prince, A. Essa, and V. Asari, “Automatic building change detection through adaptive local textural features and sequential background removal,” 07 2016.

[2] X. Huang and L. Zhang, “Morphological building/shadow index for building extraction from high-resolution imagery over urban areas,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing - IEEE J SEL TOP APPL EARTH OBS*, vol. 5, pp. 161–172, 02 2012.

[3] Y. Tang, X. Huang, and L. Zhang, “Fault-tolerant building change detection from urban high-resolution remote sensing imagery,” *IEEE Geoscience and Remote Sensing Letters*, vol. 10, pp. 1060–1064, 09 2013.

[4] X. Huang, L. Zhang, and T. Zhu, “Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index,” *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 7, pp. 105–115, 01 2014.

[5] D. Wen, X. Huang, L. Zhang, and J. Benediktsson, “A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, pp. 1–17, 01 2015.

[6] B. Qi, Q. Kun, Z. Han, H. Wenjun, L. Zhili, and X. Kai, “Building change detection based on multi-scale filtering and grid partition,” 08 2018, pp. 1–8.

[7] C. Zhong, Q. Xu, F. Yang, and L. Hu, “Building change detection for high-resolution remotely sensed images based on a semantic dependency,” 07 2015, pp. 3345–3348.

[8] A. O. Ok, "Automated detection of buildings from single vhr multispectral images using shadow information and graph cuts," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 86, p. 21–40, 12 2013.

[9] S. Ghaffarian and S. Ghaffarian, "Automatic building detection based on purposive fastica (pfica) algorithm using monocular high resolution google earth images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 97, p. 152–159, 11 2014.

[10] F. Dornaika, A. Moujahid, Y. El Merabet, and Y. Ruichek, "Building detection from orthophotos using a machine learning approach: An empirical study on image segmentation and descriptors," *Expert Systems with Applications*, vol. 58, 03 2016.

[11] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, pp. 1–13, 08 2018.

[12] J. Tian, S. Cui, and P. Reinartz, "Building change detection based on satellite stereo imagery and digital surface models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, pp. 406–417, 01 2014.

[13] P. Sidike, D. Prince, A. Essa, and V. Asari, "Automatic building change detection through adaptive local textural features and sequential background removal," 07 2016.

[14] B. Chen, L. Deng, Y. Duan, S. Huang, and J. Zhou, "Building change detection based on 3d reconstruction," 09 2015, pp. 4126–4130.

[15] B. Chen, Z. Chen, L. Deng, Y. Duan, and Z. Jie, "Building change detection with rgb-d map generated from uav images," *Neurocomputing*, vol. 208, pp. 350 – 364, 06 2016.

[16] R. Daudt, B. Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," 10 2018, pp. 4063–4067.

[17] S. Ji, Y. Shen, and M. Lu, "Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples," *Remote Sensing*, vol. 11, p. 1343, 06 2019.

[18] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[19] S. Ji, Y. Shen, and M. Lu, "Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples," *Remote Sensing*, vol. 11, p. 1343, 06 2019.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," vol. 1, 07 2005, pp. 886–893.

[21] M. Hearst, S. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *Intelligent Systems and their Applications, IEEE*, vol. 13, pp. 18 – 28, 08 1998.

[22] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 10 2001.

ParaCom: An IoT based Affordable Solution Enabling People with Limited Mobility to Interact with Machines

Siddharth Sekar[1], Nirmit Agarwal[2], Vedant Bapodra[3]
Department of Computer Engineering
Mukesh Patel School of Technology Management and Engineering
Mumbai, India

***Abstract*—There are many people in this world who don't have the ability to communicate with others due to some unforeseen accident. Users who are paralyzed and/or suffering from different Motor Neuron Diseases (MND) like Amyotrophic Lateral Sclerosis (ALS), Primary Lateral Sclerosis etc, by making them more independent. Patients suffering from these diseases are not able to move their arms and legs, lose their body balance and the ability to speak. Here we propose an IoT based communication controller using the concept of Morse Code Technology which controls the smartphone of the user. This paper proposes a solution to give the user ability to communicate to other people using machine as an intermediator. The device will require minimal inputs from the user.**

Keywords—Internet of Things (IoT); Motor Neuron Disease (MND); Amyotrophic Lateral Sclerosis (ALS); Arduino

I. Introduction

Communication is a crucial requirement in people's everyday life, but patients suffering from MND, communication can be a very challenging task. Daily conversation is an effortless task for general people with no disabilities. However, for a person with limited bodily movement and speaking disabilities, a simple task of communication becomes an immense challenge. Creating a practical and straightforward communication medium for a person with hearing and speech disabilities can be a challenging task as each person is in different stages of ALS and the body posture of each patient varies. Thereby many solutions have been proposed, which are general. These earlier proposed solutions are either difficult to implement or are expensive to be affordable by many patients suffering from MND [9].

In today's world, we see many modes and form of communication. Mobile Phones are the most used mode of communication as it removes the barrier of communication in long distances and different regions. In today's world mobile phones are upgraded to smartphones where it's like having a computer in your pocket. These smartphones connect one to all the latest happenings, discovering and situations across the globe. One of the good things of a smartphone is customization where we can customize the user interface based on our needs and desires.

Our society has people with special needs, especially those who have been immobilized due to paralysis as a result of diseases or unfortunate events [17]. Many people suffer from severe stages of paralysis, making them completely immobile and unable to communicate with other people in any way possible. According to a study conducted by the World Health Organization, 250,000 – 500,000 people around the world suffer from spinal cord injuries [27]. People suffering from these injuries tend to die two to five times prematurely than those who do not have these injuries. People who suffer from injuries due to the accidents or violence tend to find that their entire world has changed, making them unable to communicate freely, unable to do actions on their own, unable to move their body and many other things which lead to depression and is followed by suicide [16].

We built an IoT based Device which connects an ALS patient to the smartphone, thereby connecting him/her to the world digitally. Even though there are many proposed methods and devices available like Equalizer, which was used by Dr Stephen Hawking, who was diagnosed by MND immobilizing from the neck downwards [22], but unfortunately, not many people with low and middle-class income can afford such devices or methods. In a study done by the World Health Organization, only 5-15% of people in low- and middle-income countries have access to such assistive devices which they need [16].

Here we have made an Arduino based controller using sensors as activation switches which would be connected to a smartphone can be used for these patients to have a medium of communication. The concept of our controller is Morse Code Technology. Although the Morse Code [4] was primarily invented to transmit telegraphic messages, there are new applications of the code, intended mainly for persons with disabilities. The simplicity of the Morse code demands minimal resources from the end equipment, requiring only the ability to send and receive dots and dashes, unlike other systems that depend on keyboards with 100 or more keys. In every Smartphone, we see a Morse Code style keyboard with dots and dashes [20]. Morse Code is also compact and easy to understand, since it assigns shorter codes for more frequently used characters in the English language, thereby making the overall function of the controller easy to use. Here the controller emulates the use of the smartphone for the user, making him/her use their smartphone smoothly.

II. Background

A. Amyotrophic Lateral Sclerosis (ALS)

Amyotrophic Lateral Sclerosis [4] is a spinal nervous disease. In this, the patient's motor movements gradually decrease and stop at some point in time. Necessary activities such as gripping, walking, speaking, swallowing, breathing etc. This is followed by slurred speech later leading to loss of voice and stiffening of the whole body to one posture. So, communication is very improbable for such patients [6] as they can't speak and are also incapable of using their hands for sign languages. It is also challenging to diagnose, as there is no independent test or tests to detect ALS. There are tests done to rule other conditions like blood tests, brain scan, spine scan, lumbar puncture etc. ALS or Amyotrophic Lateral Sclerosis is one of the most severe types of paralysis [21].

The number of ALS patients was arguably much higher than any other MND cases. The average age of onset is between 45 and 55 years which has been found a decade earlier than the Western world. An increased risk of development of ALS has been observed in the rural population, possibly due to a link between exposure to chemicals such as pesticides and the event of neuronal injury. ALS is a devastating disease; 50% of the patients succumb within the first 30 months of symptom onset, while 20% continue suffering from the disease for 5–10 years from its beginning [18].

There is no cure for ALS to date; however, few treatments are available to reduce the impact of the symptoms. This includes physiotherapy, exercises, a special diet, few medicines to relieve muscle stiffness and saliva problems etc.

Most of the people who have ALS lose their ability to speak in its early stages [19].

ALS can be classified in three stages; they are early stage, middle stage, last stage. In the initial stage, the patient faces muscle weakness, tightness, cramping, or twitching. This stage is also associated with muscle loss or atrophy. In the middle stages, muscle weakness and atrophy spread in the body of the patient. Some muscles also become paralyzed while others lose its strength leading to contractures in joints. In the last stage of ALS, almost all the voluntary muscles of the patient's body become paralyzed [7]. These muscles include the ones used in speaking, eating, breathing, walking, etc. Breathing ventilators and feeding tubes are used to assist the patient to keep them alive. Most people with ALS die due to respiratory failure after 3-5 years since the appearance of the first stage [19].

B. Morse Code

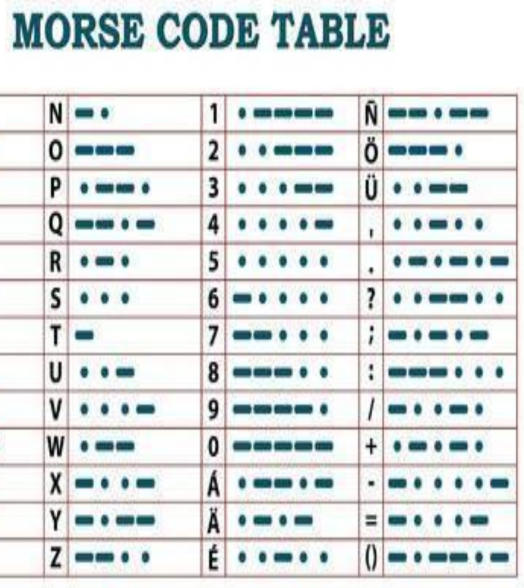
We help people suffering from ALS to connect to the digital world using the morse code as an interactor between our hardware device and smartphone. The Morse Code alphabet uses "dots" and "dashes" to encode the characters in the English language. Each letter is represented by a unique sequence of dots and dashes, e.g. the letter "a" is encoded as "dot-dash” [8]. Codes are also included for numerical characters and special characters. Amateur radio operators use the system. In HAM radio sets, a dot makes a short sound and is called a "dit" whereas a dash makes a longer sound and is referred to as "dash." The duration of a dash is three times the duration of a dot. Each dot or dash is followed by a short silence, equal to the dot duration. In Google Keyboards we can set the default keyboard to Morse keyboard making it very efficient to use and reducing the span of errors which used to cause in earlier Morse Code sequences [1][2].

When Samuel F.B Morse initially invented morse code, it consisted only to translate numbers in the beginning.

Alfred Vail expanded the morse code and included letters and few special characters such as punctuation marks. The Morse code has an assigned unique sequence of all the letters, numbers and special characters. The combination of signals was attached to each letter by conducting research on which letter is used the most by people and that letter has a shorter sequence then the characters that are used less frequently. The table below gives the morse code signal for all the letters, numbers and special characters [1] (Fig. 1).

C. Internet of Things

Internet of Things (IoT) is an integration of multiple devices connected through internet which shares or collects the data (Fig. 2). IoT is a platform where you collect the data through the sensors and even control the function of the devices. IoT technology transforms a simple device into smart computing peripherals that function to merge the digital and physical worlds of human beings. The equipment that is being controlled via the network transforms it into an IoT device. An example of using an IoT device is a simple light bulb which can be operated from anywhere using a controller as simple as a mobile phone. IoT device helps people in much more ways nowadays by reducing the error margin. It helps the user by giving them control, even being distant from the IoT device. For example, a User lives in a smart home which allows the user to turn off lights also if he is away.



MORSE CODE TABLE

A	• —	N	— •	1	• — — — —	Ñ	— — • — —
B	— • • •	O	— — —	2	• • — — —	Ö	— — — •
C	— • — •	P	• — — •	3	• • • — —	Ü	• • — —
D	— • •	Q	— — • —	4	• • • • —	,	• • — • •
E	•	R	• — •	5	• • • • •	.	• — • — • —
F	• • — •	S	• • •	6	— • • • •	?	• • — — • •
G	— — •	T	—	7	— — • • •	;	— • — • —
H	• • • •	U	• • —	8	— — — • •	:	— — — • • •
I	• •	V	• • • —	9	— — — — •	/	— • • — •
J	• — — —	W	• — —	0	— — — — —	+	• — • — •
K	— • —	X	— • • —	Á	• — — • —	-	— • • • • —
L	• — • •	Y	— • — —	Ä	• — • —	=	— • • • —
M	— —	Z	— — • •	É	• • — • •	()	— • — — • —

Fig 1. Morse Code Table.

Fig 2. IoT Devices.

Nowadays, the commonly known IoT devices are google home assistant, smart homes, smart shoes, Google Lens and many other accessories. IoT has become a commonplace which is used to integrate the digital world as well as the physical world of the humans to make the tasks more convenient.

III. Related Work

A. Eye Tracking (Laser & Twitch)

A more straightforward approach to Eye-Tracking includes hardware like two sets of IR LED sensors and an Arduino board. The process is such that the user's eyes are detected using the IR sensors by sensing the amount of light reflected from the eye and sends this data in the Arduino [23]. The Arduino then compares the data received from the IR sensor and determines whether the eye is open or closed with a predefined value. The closed eye is read as a blink [10], and then the flash and its duration are recorded temporarily, and these readings are taken at a rate of 10 per second. If the blink pattern matches a previously set design, then the alphabet is printed, or that function/command corresponding to that pattern is executed on the system [5].

A famous eye-tracking system for ALS patients is the Eyegaze Edge, which provides users with the ability to control devices like stereos and television, surf the internet, send emails, read books, and hold conversations. Many people cannot afford the equipment due to their expenses. The eye tracking device is based on the patient's knowledge, understanding the language as well as experience with the computers. The patient should be familiar with the recent technology to use this device. The eye tracking device is costly for the people who are having a low-class income or lives in a country which has low- and middle-class economy which is why most of the people cannot afford this device economically [22] (Fig. 3).

Fig 3. Eye Tracking Device.

B. Voice Amplification

This approach is used if the patient is suffering from the initial stages of ALS. Often after sometimes, patients with ALS in its later stages find themselves no longer able to project their voices due to complications in their respiratory system. So, there are devices made to help amplify the projection of their views can help to reduce fatigue by reducing the effort needed to speak. Some of these devices are Chatter Vox Voice Amplifier which comes with a headset microphone and can boost one's volume up to 18 decibels. I can also be comfortably fastened and worn at the front of your waist [22].

Voice Amplifier is a great device, but there is a big flaw for communication for ALS patients (Fig. 4). It can only be used by the patients that have not lost their voice [14] and according to a study conducted that around 80%-90% of the patients who have ALS lose their voice in the early stages which makes them unable to communicate. The reason mentioned above is not the only one a dumb person who is suffering from a severe paralysis also will not be able to communicate with any other person.

C. Voice Banking

This is a method used by ALS patients before they experience a disability to speak. Users either record typical phrases like "How are you today?" and "Let's go out to eat." etc. but some users also record a list of sentences and the sounds generated during these recordings are used to synthesize speech. One example is Samsung's and Google's initiatives to record one's voice in their intelligence assistant software like Bixby and Google Assistant, and these recordings are then used by this software whenever they are used.

Fig 4. Voice Amplifier Device.

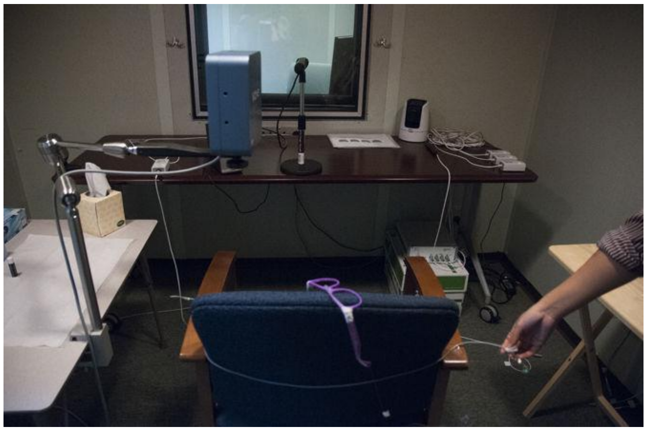
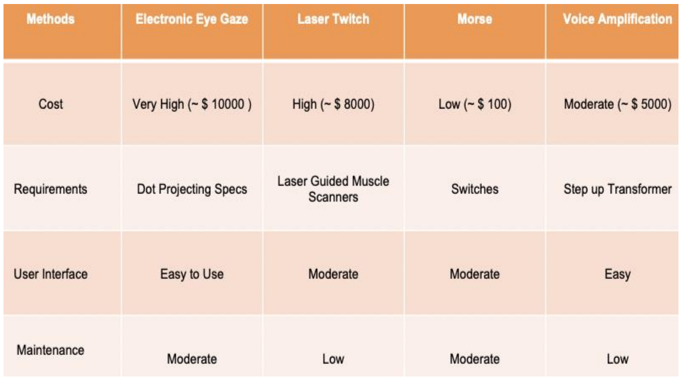


Fig 5. Voice Banking Device.



Methods	Electronic Eye Gaze	Laser Twitch	Morse	Voice Amplification
Cost	Very High (~ $ 10000)	High (~ $ 8000)	Low (~ $ 100)	Moderate (~ $ 5000)
Requirements	Dot Projecting Specs	Laser Guided Muscle Scanners	Switches	Step up Transformer
User Interface	Easy to Use	Moderate	Moderate	Easy
Maintenance	Moderate	Low	Moderate	Low

Fig 6. Market Survey.

Voice banking is a great communication device, but it has many flaws, some of which are sometimes it may not be able to express the correct thoughts of the user, it is costly, and it only helps if a person is communicating with another person right in front of him (Fig. 5). It is hard to use by patients who have their whole body paralyzed. It also does not help the patient to communicate with someone through the digital world [22].

Our experiment was inspired by an ongoing google experiments campaign wherein they are working on providing ALS ridden people with the ability to effectively communicate with others through the medium of Android device and a coupled morse input switches (Fig. 6). The user punches in the input as morse signals and there are further processed to be taken as keyboard input or device navigation gestures.

IV. List of Components

A. Arduino Leonardo

Arduino Leonardo is a microcontroller board based on the ATmega32u4 (Fig. 7). There are 20 digital I/O pins out of which seven can be used as PWM (Pulse Width Modulation) outputs and has 12 pins as analogue inputs. It also contains a 16 MHz crystal oscillator, a micro USB connection, power jack, ICSP header and a reset button.

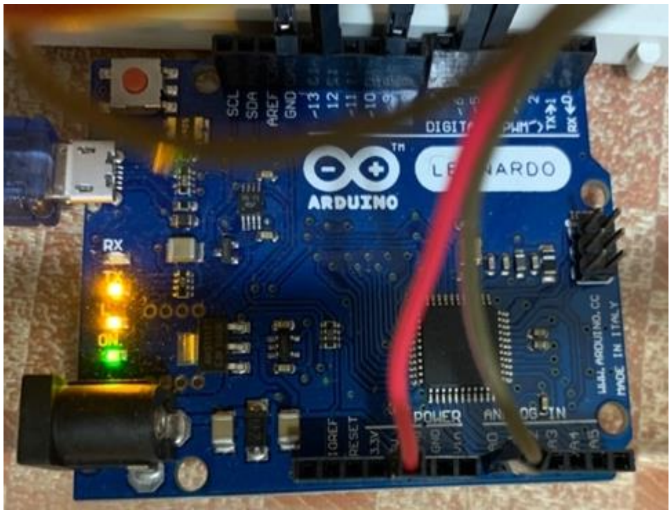


Fig 7. Arduino Leonardo.

The Leonardo board differs from all preceding boards in that the ATmega32u4 has built-in USB communication, which eliminates the need for a secondary processor. This allows Leonardo to appear to a connected computer as a mouse and a keyboard, in addition to a virtual (CDC) serial / COM port. The Arduino IDE software allows one to write the program on a digital device and helps them upload to your board. It is an open-source software which allows any user to code and upload the application to the board quickly. The environment of Arduino IDE is written in Java and is based on processing and another open-source software. The Arduino IDE can be used with any Arduino board. The Arduino IDE supports the languages like C and C++ using special rules of code structuring [24].

B. NodeMCU

NodeMCU is one of the most popular open-source IoT platforms. It provides access to GPIO (General Purpose Input Output) Pins and includes all the necessary firmware and hardware. Firmware is being run on an ESP8266 Wi-Fi SoC which consists of an ESP12 module. It contains 13 GPIO (General Purpose Input/Output) In all the 13 GPIO pin, only the GPIO 16 can be used for reading as well write operation. It does not support 1-wire, open drain, Interrupt or PWM [25] (Fig. 8).

C. Potentiometer

A potentiometer is 3 terminal variable resistors in which the resistance is manually varied to control the flow of the electric current. A potentiometer acts as an adjustable voltage divider.

It is a passive electronic component which works by varying the position of the sliding contact across a uniform resistance. The entire input voltage is applied on the whole length of the resistor, and the output voltage is applied across the entire range of the resistor [26] (Fig. 9).

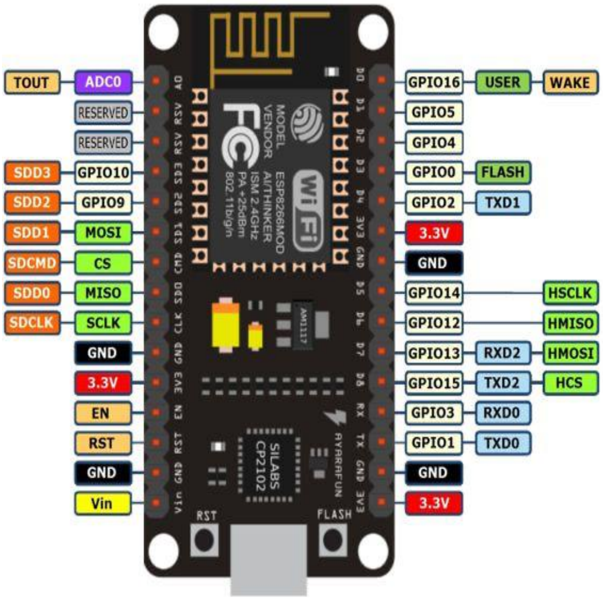


Fig 8. Schematic of NodeMCU.

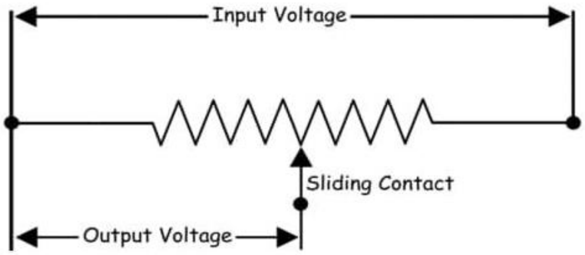


Fig 9. Circuit Diagram of Potentiometer [26].

Fig 10. Rotary Potentiometer.

Here we have used a Rotary type potentiometer in our device. It is used to obtain an adjustable supply voltage to a part of electronic and electrical circuits. It includes a rotary knob of the potentiometer which controls the supply to the amplifier (Fig. 10). This type of potentiometer has two uniform resistance terminal contacts places in semi-circular patterns. It also has a middle terminal through which the sliding contact on the semi-circular resistance is connected. When we rotate the knob, we move the sliding contact on the semi-circular resistance. It is used in substation battery chargers which adjust charging voltage of a battery.

V. Implementation

ParaCom helps to give a mode of communication to those people who do not have any pre-existing modes to communicate using the digital world as a medium, especially for the people who suffer from ALS. It helps the patient to use his/her digital device and give complete control of the device to them Morse code.

Fig 11. Relay Module (4 Channel).

Fig 12. NodeMCU.

Our devices brain can be considered as the Arduino Leonardo itself as it is the device which has the keyboard interfacing capabilities, which in turn are an integral and essential part of our product (Fig. 11).

The Arduino Leonardo, in turn, acts as an i/o device as it accepts input from the users and feeds them as an output to the corresponding device. This is done by the fact that the inputs are individually recorded as '.' and '-' and these, in turn, are interpreted as Morse code by the corresponding Morse keyboard. These inputs are taken by the switches which are assigned to either a '.' or a '-'. These purpose matched switches are correctly programmed for such an ideal operation wherein the user input may or may not be of the highest accuracy. Also, our setup includes a third switch as this is also an integral part to invoke the functioning of the system of the whole system [2] (Fig. 12).

This switch is the mode change switch, and it works by, as the name suggests changing the mode between the typing mode and the User Interface (UI) navigation mode. In the typing mode, the switches are used as '.' & '-' whereas in the UI navigation mode these are used as arrow keys for either going left or right throughout the menu [21].

The scanning of the screen is done entity wise from the top left of the screen to the bottom right of the screen. This intern ensures that none of the features is missing and we encompass every individual, and it is accounted for in the resulting film real estate (Fig. 13).

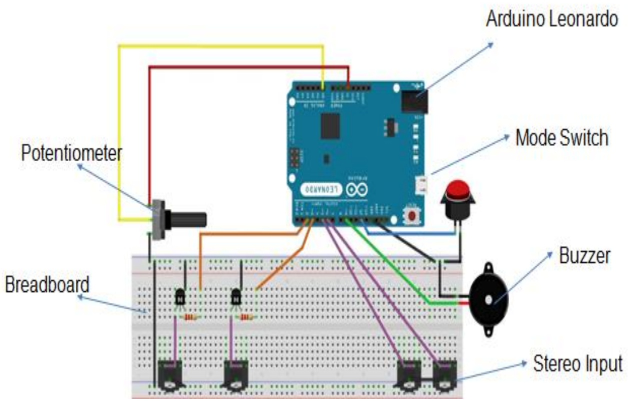


Fig 13. Connection of the Circuit.

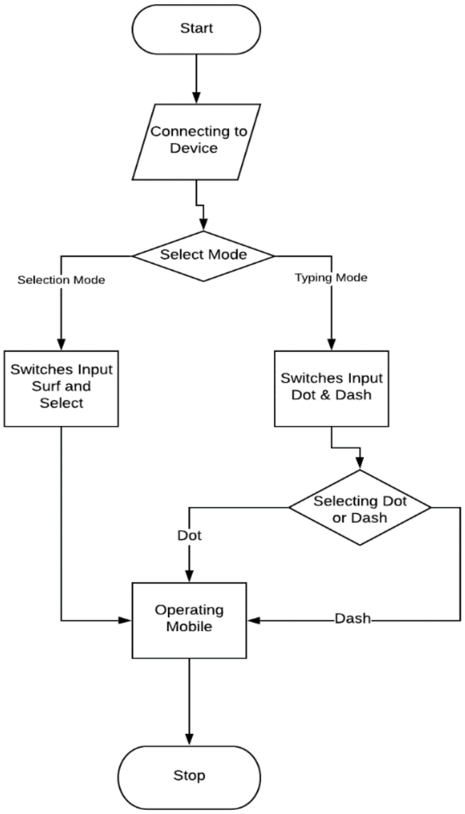


Fig 14. Flowchart of the Operations.

The other major component that helps less able people in using this whole system is the variable potentiometer. Since the patients have limited motor neuron skills, it only helps for them to have the system configured to their way of typing. In turn, this is the most any system can wish for, i.e. a system which adapts to the user itself as no system shall ideally present the user with any discomfort (Fig. 14).

This potentiometer works by reducing or increasing the input speed the input by lowering or raising the voltage resistance across the switches. This helps for higher accuracy as the rate at which the system gets an input signal can be made constant by adjusting such variables. For this to happen, the user at no time needs to adapt to the system, and the system takes care of this itself, which is what we strive for [26].

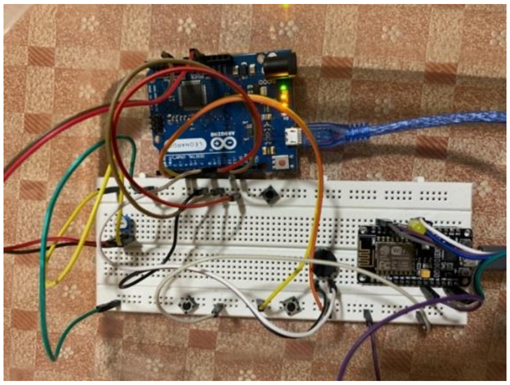


Fig 15. Architecture of the Device.

The next major component is the smart home automation environment of our device. This is done using a NodeMCU (ESP8266) and firebase protocol [25] (Fig. 15). The NodeMCU is hooked up to a relay which in turn can work as a switch to control several connected appliances of any magnitude.

All of this is made possible through a companion android application which helps us control the individual relays through virtual buttons remotely. The phone need not be on the same network and needs to just reconnect through the internet to the same firebase database virtual server as our NodeMCU.

The NodeMCU is connected through a preconfigured wireless network which is already defined in its Arduino programming code. The signals are sent from the phone to the server, and then these signals are in turn picked up by the NodeMCU as it keeps scanning for changes in its connected server and these changes are reflected in its output pins and turn to the relays (Fig. 16).

A module has been added in our device, i.e. the module of home automation. The patient can control the electrical devices in their homes, i.e. using our IoT device through their smartphones, they can remotely monitor their electrical appliances, i.e. other IoT devices like television, fans, led lights, geysers etc. This module will provide more independence to the patient using our devices by making them in charge of their home appliances [11].

The user interacts with the device by virtue of the buttons placed in the headrest of the wheelchair ridden user. These buttons would help the user control the whole device mounted in front of them in the way mentioned above and will hence allow them to control the corresponding smart home services too.

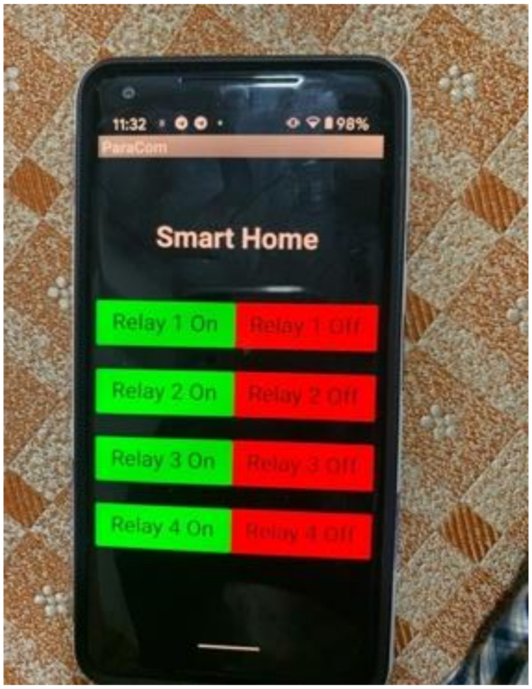


Fig 16. User Interface of Android App.

VI. Results

As mentioned earlier our device will act as a medium for any ALS or severe conditioned patient who is unable to speak and unable to use his hands for gestures. We will use a sensor-based headrest acting as the communication medium for a device to the user. Based on the user's head movements (which can be calibrated to any movement) the user will be able to use the smartphone.

Since there are three buttons in our system as we are using the concept of Morse code which includes a dot(.), a dash (-), and a third button which changes the mode of operation on the device i.e. either typing or operating the phone. Using the head movements, the user will choose dot(.) or dash (-) and the mode button would be operated remotely where each sensor is connected to the respective button.

So, for example when the user turns his head left corresponding bit that goes to the controller board is a dot (.) and when the user turns his head to the right a dash (-) will be the bit that is sent to the controller. These two bits can be used by the patients to communicate with other people for sending a message in any social device. This will help the patient to convey his or her own thoughts to other people. The third bit helps the patient to change the mode of the digital device when the user nods, the device will send the third corresponding bit to the controller switch that is to change the mode of the device it helps the patient to switch between apps by himself/herself.

The user will also be able to control compatible devices in the house when the user switches to the software application that we made. For example, the user wants to turn on the light in the house, he/she will simply turn the head left sending the dot bit to the controller and it will turn on the light whereas turning the head right will turn off the light [15]. It will help the user to control home making the patient more independent. It will also be convenient for the user to control the device using a simple dot(.) and dash (-).

The control of these devices is finally handled by corresponding relays that are physically connected to each of the devices' line of communication. The relays act as a remote switch which can be operated by electrical signals and need not be accessed physically by a person [13]. We take advantage of this basic concept allowing the patient to control the corresponding relay-controlled devices remotely by the virtue of their smartphone [12] (Fig. 17).

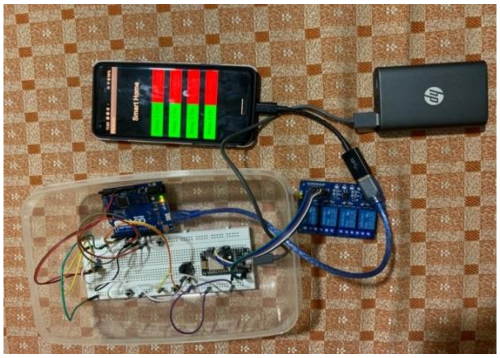


Fig 17. Complete ParaCom Setup.

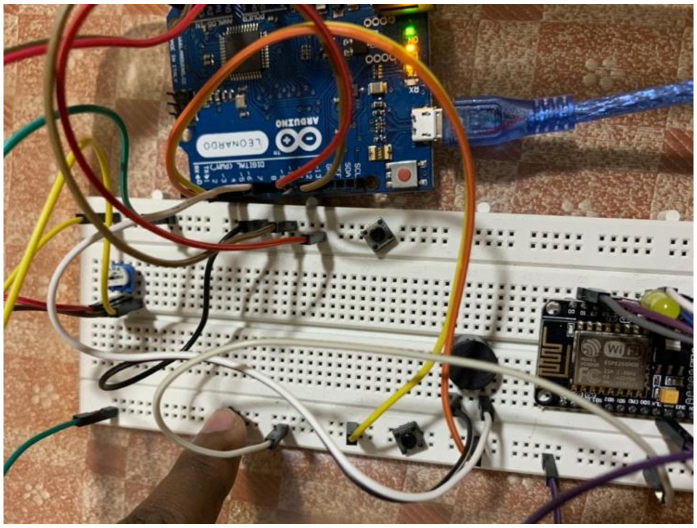


Fig 18. Using Dash (-) button of ParaCom.

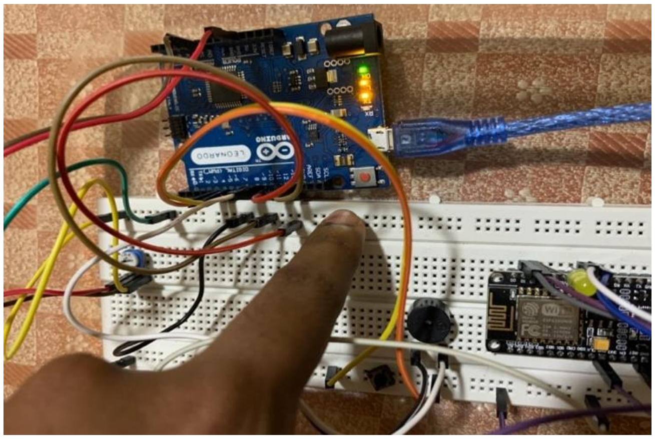


Fig 19. Using Mode Switch Button in ParaCom.

The pressing of an "ON" or "OFF" switch on the preinstalled application, in turn, sends a signal to the corresponding relay. This is achieved by using Google's Firebase as a backend database and the NodeMCU (ESP8266) reads data from firebase and sends the appropriate signals to the relays thereby enacting the user's desired command. This whole setup can also be controlled through voice activated control through Google assistant.

Here are some implementation images although they are using direct pressure buttons instead of sensor-based headrest but the implementation overall is the same (Fig. 18 and 19).

VII. Challenges

The ParaCom device is a capable connectivity device built for paralyzed and ALS patients that is still in its initial phases. There is much room for further improvement and enhancement of the device. Below we have listed a few limitations of ParaCom.

A. Device Dependent: One of the significant limitations of ParaCom is that it needs a smartphone or tablet device to work. It was mentioned earlier it acts as a medium for the patient and the Smartphone [3].

B. Slow Usage Speed: Using the smartphone through ParaCom decreases their speed of usage of the phone by ALS patients compared to the average rate of smartphone usage by ordinary people as they can directly use the touchscreen of the phone whereas when using through ParaCom it is like using a cursor in the smartphone to operate it [3].

C. Pre-setup and Calibration: ParaCom being a hardware device it needs to be placed near the patient's wheelchair or bed and near the smartphone to which it will connect to either using a USB cable or using Wi-Fi Direct. The Input rate of the sensors also needs to be adjusted according to patient's comfort of using his/her ability to use the sensors smoothly. Also, the sensors need to be calibrated according to the movement of the patient and also be placed in an effective position [3].

VIII. Future Scope

As of the future scope of our device, we can implement many add on functionalities to it. One of them is the implementation of patient monitoring where we can measure the temperature, blood pressure, heart rate, breathing rate etc. of the patient using the device. All these recorded details of the patient are sent to their doctors and family members as a daily health status report. In case of any unusual or unpredictable readings being measured, an SOS can be sent to their doctors and family members regarding the health status and immediate measures can be taken accordingly.

A new companion application could be added to our project setup which would in turn allow the patient to navigate the device with ease and hence mitigate the frustration of going through complex menus while having limited accessibility to the device user interface. This application would encompass functionalities such as medication reminder, remote physician database updating, accessibility features to aid in device usage and much more.

IX. Conclusion

People who have severe paralysis or ALS feel separated from society as they would be unable to communicate and interact with other people. ParaCom is a solution for all those paralyzed individuals to perform on-screen tasks, especially those people who want to do something valuable with their own lives and contribute something to the world. ParaCom has several functions, especially for paralyzed people like.

A. Operating Gadgets: It helps the paralyzed person to control their digital device, giving them a sense of independence to manage their digital device.

B. Entertainment: paralyzed people can even play games or watch movies on their digital device. It can also help them to communicate to text with other person become more active on social media.

C. Information Seeking: Helps the paralyzed person to be up to date with the current world affairs. It helps them to get the information that they seek, to study and learn about various things where their interest lies.

References

[1] Alex Rupom Hasdak, Istiaq Al Nur, Adnan Al Neon and Hasan U. Zaman, “Deaf-Vibe: A Vibrotactile Communication Device Based on Morse Code for Deaf-Mute Individuals”, 2018 9th IEEE Control and System Graduate Research Colloquium (ICSGRC 2018), 3 - 4 August 2018, Shah Alam, Malaysia

[2] Sérgio Silva, António Valente, Salviano Soares, M.J.C.S. Reis (a,c),Jean Paiva, Paulo Bartolomeu, Morse Code Translator Using the Arduino Platform: Crafting the Future of Microcontrollers, SAI Computing Conference 2016 July 13-15, 2016 | London, UK

[3] Haroon Malik, Anam Mazhar, “EyeCom: an IoT based affordable wearable solution for paralyzed people to interact with machines”, Journal of Ambient Intelligence and Humanized Computing

[4] Franklin Rosado, Iliana Rumbo, Felicia Daza, Hernis Mercado, Diana Mier, "Morse code-based communication system focused on amyotrophic lateral sclerosis patients."

[5] Kingshuk Mukherjee, Debdatta Chatterjee, “Augmentative and Alternative Communication Device Based on Eye-Blink Detection and Conversion to Morse-Code to Aid Paralyzed Individuals”, 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Jan. 16-17, Mumbai, India

[6] Yurdagül KARAGÖZ, Sevda GÜL, Gökçen ÇETøNEL, “An EOG Based Communication Channel for Paralyzed Patients”

[7] Md. Masud Rana, Md. Wahidul Islam, Md. Jubayer Hossain, “EyeWriter Based PC Control System for Paralyzed and Disabled Patients”, International Conference on Advancement in Electrical and Electronic Engineering 22 - 24 November 2018, Gazipur, Bangladesh

[8] Cheng-Hong Yang, Li-Yeh Chuang, Cheng-Huei Yang, and Ching-Hsing Luo, “Morse Code Application for Wireless Environmental Control Systems for Severely Disabled Individuals”, IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING, VOL. 11, NO. 4, DECEMBER 2003

[9] Dr Kirti Wankhede1, Sayali Pednekar, "Aid for ALS Patient Using ALS Specs and IOT", 2019 2nd International Conference on Intelligent Autonomous Systems (ICoIAS)

[10] Asfand Ateem, Mairaj Ali, Muhammad Asad Bashir, Zeeshan Ali Akbar, "Eye Monitored Device for disabling People", 2017 20th International Conference of Computer and Information Technology (ICCIT), 22-24 December 2017

[11] Md. Mohaiminul Islam, Md. Nahiyan Farook, S. M. G. Mostafa, Yasir Arafat, "Design and Implementation of an IoT Based Home Automation", Advances in Science Engineering and Robotics Technology (ICASERT) 2019 1st International Conference on, pp. 1-5, 2019.

[12] Shih-Chung Chen, Chung-Min Wu, Yeou-Jiunn Chen, Jung-Ting Chin, Yu-Yin Chen, “Smart Home Control for the People with Severe Disabilities”, Proceedings of the 2017 IEEE International Conference on Applied System Innovation IEEE-ICASI 2017 - Meen, Prior & Lam (Eds)

[13] Progress Mtshali, Freedom Khubisa, “A Smart Home Appliance Control System for Physically Disabled People”, 2019 Conference on Information Communications Technology and Society (ICTAS)

[14] Mrs Paul Jasmin Rani1*, Jason Bakthakumar2, Praveen Kumaar.B3, Praveen Kumaar.U4 and Santhosh Kumar5, "VOICE CONTROLLED HOME AUTOMATION SYSTEM USING NATURAL LANGUAGE PROCESSING (NLP) AND INTERNET OF THINGS (IoT)", 2017Third International Conference on Science Technology Engineering & Management (ICONSTEM)

[15] Moravec P., Krumnikl M., Olivka P., Seidl D. (2016) Connecting Household Weather Sensors to IoT World. In: Saeed K., Homenda W. (eds) Computer Information Systems and Industrial Management. CISIM 2016. Lecture Notes in Computer Science, vol 9842. Springer, Cham

[16] World Health Organization - https://www.who.int/news-room/fact-sheets/detail/spinal-cord-injury

[17] Christopher and Dana Reeve Foundation Statistics about paralysis, https://www.christopherreeve.org/living-with-paralysis/stats-about-paralysis

[18] Indian Brand Equity Foundation - https://www.ibef.org/industry/healthcare-presentation

[19] ALS News Today on Stages of ALS - https://alsnewstoday.com/stages-of-als/

[20] Los Angeles Times - https://www.latimes.com/archives/la-xpm-1987-10-25-me-16256-story.html

[21] Google Experiments - https://experiments.withgoogle.com/collection/morse

[22] Crossroads Hospice and Palliative Care.com on available approaches of communication for ALS patients Posted on Wednesday, May 2, 2018 - https://www.crossroadshospice.com/hospice-palliative-care-blog/2018/may/02/new-tools-for-patients-with-als/

[23] Arduino IDE – https://www.arduino.cc/en/main/software

[24] Arduino Leonardo - https://www.arduino.cc/en/Main/Arduino_BoardLeonardo

[25] NodeMCU - https://nodemcu.readthedocs.io/en/master/

[26] Potentiometer and its uses - https://www.electrical4u.com/potentiometer/

[27] Spinal Cord.com on Types of Paralysis 19 November 2013 - https://www.spinalcord.com/types-of-paralysis

Security of a New Hybrid Ciphering System

Mohammed BOUGRINE[1], Fouzia OMARY[2], Salima TRICHNI[3]
Faculty of Sciences, Mohammed V University in Rabat
Department of Computer Science
Rabat, Morocco

***Abstract*—The protection of privacy is a very sensitive subject and comes into force in all areas. They represent the first priority in the development of new technologies. In fact, opt for a new Big data or IOT technology is a very difficult decision for organizations and calls into question the confidentiality, integrity, authenticity and non-repudiation of their data. Convincing these organizations to adhere to technological intelligence is tantamount to providing them with powerful tools and mechanisms of security that are resistant to new types of vulnerability. However, the problem today is that most security tools are based on old cryptographic primitives. Certainly; they have proved their resistance until today but the need to have others becomes crucial in order to meet the new technological requirements. In this paper, we propose a new hybrid encryption alternative based on two encryption systems, the first one is an evolutionary encryption system and the second one is based on an asymmetric encryption system. To present this work we begin with a description of our evolutionary cipher system. Then, we present the principle of proposed hybridization and its contribution compared to other existing systems. Finally, we perform a detailed study on the safety of this system and its long-term resistance.**

Keywords—Security; confidentiality; hybrid encryption; evolutionary algorithms; symmetrical encryption; cryptography

I. Introduction

Symmetrical encryption systems, although invented long before the asymmetric encryption systems, are still the most commonly used type of cryptosystems used in applications and information systems [1].

The widespread use of symmetrical encryption systems is mainly due to their simplicity, speed and security strength compared to asymmetric encryption systems [2]. This is the case as long as an attacker cannot discover the secret key, which represent a critical criterion for the application. Therefore, the main difficulty lies in the distribution and the agreement over the keys to enable the entities concerned by this communication to share the same initial secret without any potential attacker intercepting it [1][2]. The delivery of the secret key must take advantage of all possible means of protection to ensure the authentication, the integrity and the confidentiality of all the information exchanged.

Whitfield Diffie and Martin Hellman in [3] were able to put an end to this problem and avoid the pitfall of symmetrical systems using a new mechanism based on two keys, one public and one private [3]. The emergence of asymmetric cryptosystems, or public key cryptosystems, provide an indubitable answer to the key exchange problem. The robustness of this type of algorithms is based on the difficulty and complexity of resolution of certain mathematical problems [4]. However, these algorithms lack speed and are practically unusable especially for an online exchange with large volumes of data. However, this kind of cryptosystems is used in hybrid cryptosystems. Hybrid cryptosystems are a new approach that consists of a combination of symmetric and asymmetric algorithms in order to take advantage of the benefits of each of them and make them complementary.

In this paper, we took inspiration from the hybrid cryptosystems approach [5][6][7][8], to design a new Hybrid Evolutionary Cryptosystem. The goal of the present work is to describe the process of this system and then to show his strength against other existing hybrid cryptosystems which are in widespread use.

II. Related Work

Philip Zimmermann was the first to introduce hybrid cryptographic systems. He managed to combine the IDEA symmetric encryption system with the RSA asymmetric encryption system. His work gave birth to the Pretty Good Privacy (PGP) cryptosystem [5]. PGP was the first hybrid encryption system created. Since then, PGP has incorporated other cryptographic concepts to cover not only the data confidentiality but also the different security requirements for the exchange, storage and disclosure of data for private use (signing, compression, etc.).

III. Background

A. Description of the Advanced Symmetrical Evolutionary Ciphering (ASEC)

Brief History:

In 2006, the Symmetrical Evolutionary Ciphering (SEC) was created. It was one of the first systems introduces evolutionary algorithms [22][24] as an encryption process in [9][10] and [11]. It is based on a simple principle. First, the plaintext is encoded and each character is linked to its positions' list. Then, a search through the different iterations of the genetic algorithm is done in order to find the most powerful combination of these lists to realize a well secured encryption [11].

In each step of this algorithm, a set of mathematical mechanisms and methods is applied in order to find the solution that meets the need of confidentiality. In 2011, in order to respond to new security requirements, we developed an advanced version of this system called "Advanced Symmetrical Evolutionary Ciphering" (ASEC), and we introduced the partition problem in the stage of mutation [12] and also at the level of the evaluation function [13].

Ciphering Algorithm:

To explain the ciphering, let's T be the plaintext.

T is an input of our system.

Step 1: Encoding

This is the stage of coding the plaintext as a chromosome.

T contains the following characters: c1, c2, c3, ... , cm.

Each character occurs at least in one position in the text, then we define for each character his list of positions in this text called Li (0 <i <m+1). So, the plaintext is represented by the vector T = {(c1, L1) ... (cm, Lm)} which will be the initial chromosome of all the populations.

Also, with this representation, they are two important properties of this population, which are:

① Li ∩ Lj = Ø, for i, j ∈ [1, m], with i ≠ j.

② L1, L2,..., Lm is a partition of the set {1, 2 ..., n}

Step 2: Generating the initial population

Let:

- q be the population size
- CHj (j ∈ [1, q]) be the representation of each chromosome
- and P1 be the representation of the initial population

Then, the first population will be represented by: P1 = {CH11, CH12, CH13,…, CH1q}

The second one is: P2 = {CH21, CH22, CH23,…, CH2q}, and so on. Each chromosome CHj (j ∈ [1, q]) is defined as a new combination between the ci and Li.

The first generated population must not follow a well-defined function but it must rely on random events to generate it because more the initial population generation is random more the algorithm is efficient.

Step 3: Evaluation

In this step, we evaluate a random partition Ej constructed from each chromosome Xj such as:

Ej ={ej1 , ej2 , … , ejm}.

And then you have to assign a value to each chromosomal partition in order to evaluate its effectiveness using the fluid formula [13]:

$$F(\mathrm{Xj}) = \sum_{i=1}^{m} |Card\,(eji) - [n/m]|$$

Through this function, we try to find the partition that all his elements has a similar cardinality.

Step 4: Selection

Using a selection method the roulette wheel selection. As its name suggests, the principle of this function is based on the casino roulette performance [14]. It can transform the performance of each parent to a probability that will be distributed later on the roulette of the game. We randomly choose the value of the parameter "r" that can be considered as the ball to be cast on the wheel in order to choose the elected chromosome [15].

Step 5: Genetic operators

There are two steps: a crossover and a mutation.

MPX Crossover method:

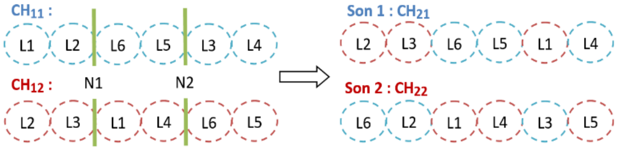
In this case, we must use the crossover that maintains the characteristics of the original population which are:

① Li ∩ Lj = Ø, for i, j ∈ [1, m], with i ≠ j.

② L1, L2, ..., Lm is a partition of the set {1, 2, ..., n}

This is why the MPX crossover is used [16] where the coding of the child has a strong analogy with that of the parents.

This crossover was developed specifically for the TSP problem by Gorges-Schleuter and Mülhelenbein [17] in 1988. The MPX operator is illustrated in the example below:



Mutation:

Contrary to the old SEC encryption system, in ASEC the mutation step is the most important step in the algorithm.

In fact, at this level we try to create the new partition of the positions lists [12].

It is noted that the new lists must absolutely respect the properties of the original text. That said:

- ✓ they must be independent: L'$_j$ ∩L'$_i$ = Ø
- ✓ they must be ordered

Construction of the new generation:

The new generation is built keeping the same chromosomes of the population of the crossover except that this time it is based on the new lists.

In other words, instead of having the child: L3-L1- L5 -L10-L7-L2-L4-L9-L6-L8

We will have L'3-L'1- L'5 -L'10-L'7-L'2-L'4-L'9-L'6-L'8 and so on.

Discussion:

From the new design of the lists of positions, we can see that a character can replace 1, 2 or even more characters as it can be replaced by several other characters instead of a single character. The relationship between the initial character and the replacement character becomes more complex.

Encryption Key:

Finally, to encrypt plaintext, the key is represented as follow:

- The sequence of numbers with the permutation of the elected child.
- The sequence of numbers with the permutation of the elected child in the new partition of lists.
- The sequence of numbers with the cardinals of the elected child lists.
- And ultimately, the final permutation of encryption.

Decryption:

To Decrypt message, we applied the same Key in inverse order

IV. Hybrid Evolutionary Cryptosystem

Problematic:

ASEC can be considered as a symmetric encryption system because it uses the same key for encryption and decryption. The only difference is that the evolutionary encryption resembles the disposable mask encryption mechanism. In fact, the encryption key is not exchanged once but changes from an execution to another. It is then considered a session key. The problem that arises in this case is that this key must absolutely be secured whenever we wish to establish a communication using this system of encryption.

Solution:

To address this problem, we propose to use the principle of hybrid cryptosystems using the symmetric ciphering ASEC that allows to include the session keys generation step. In this new cryptosystem, the keys are generated implicitly by the system of encryption.

The principle is simple and can be illustrated by the following diagram (Fig. 1).

The question that arises is: what is the benefit of ASEC for a user compared with other symmetrical systems used in the PGP cryptosystem?

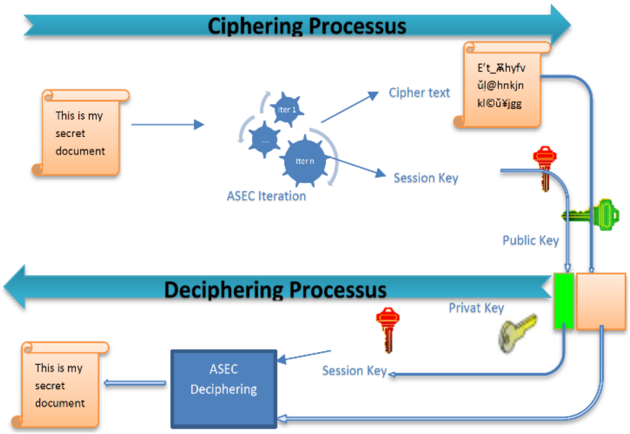


Fig. 1. Principal of Hybrid Evoltionnary Cryptosystem.

For this, we propose to begin with the security study of this system because it represents the first factor to choose it. Then we give a comparative study with the others symmetrical systems used in the PGP cryptosystem.

To answer this question, we need to specify the selection criteria for these symmetrical systems and see how our system can meet these criteria.

We can then distinguish the following criteria:

- ✓ Security degree
- ✓ Time: Fast or slow
- ✓ Material: the capacity of the equipment used to encrypt and decrypt messages.
- ✓ Setting: the size of the key and the blocks, allowing to increase the strength of the algorithm against brute-force attacks.
- ✓ Power: ability to resist the different possible attacks according to the setup of this algorithm in the PGP system.
- ✓ Reputation: it is related mostly to the seniority of the algorithm in the field.
- ✓ Patented.

In fact, in this study, we can't rely on all criteria to demonstrate the effectiveness of this system because obviously ASEC has not yet been released to talk about his reputation and accessibility. However, we're going to focus on security study of this system because it represents one of the most interesting factors in this phase and then we will do a comparison study based on the other criteria such as the execution time, the setting and the Material.

V. Security Study

A. *System Setting and Brute Force Attack*

Symmetric ciphers systems setting is a prerequisite essential for his security [17]. The setting comprises the key size and the block size needed to ensure the resistance of the system to brute Force attacks [18][19].

Two major factors increase the resistance to this type of attack:

- ⇨ The size of the key, which must be as high as possible [27].
- ⇨ The representative sequence of the key which must be undistinguishable from a true random output by a third person [19].

Key Length:

The size of this key depends on the number of different characters of the plaintext to be encrypted using the following relation: (8*n)*4, with n being the number of different characters of the plaintext.

An experimental study is performed on several plaintexts of different sizes and from different sources. Table I and Fig. 2

and 3 show the progression of this key depending on the size of the text to be encrypted.

TABLE I. DEPENDENCE BETWEEN THE KEY AND THE SIZE OF THE TEXT

Size of the message (characters)	Size of the message (bits)	Different characters	the key size
642	5136	40	1280
864	6912	31	992
1204	9632	52	1664
1516	12128	41	1312
2893	23144	55	1760
4543	36344	66	2112
5514	44112	72	2304
6097	48776	80	2560
6181	49448	110	3520
5514	44112	72	2304
9162	73296	82	2624
14250	114000	83	2656
20531	164248	85	2720
23396	187168	100	3200
24280	194240	91	2912

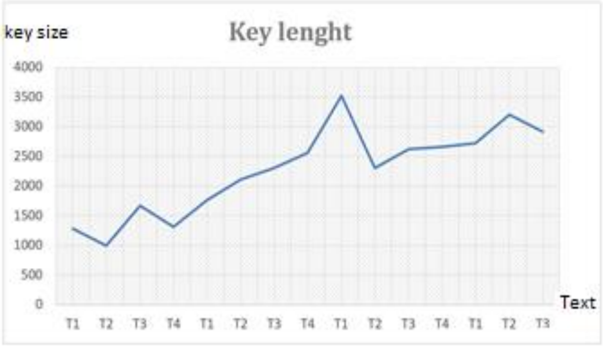


Fig. 2. Dependence between the Key and the Size of the Text.

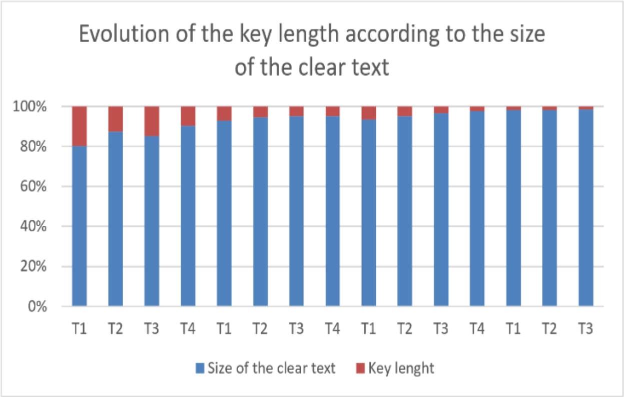


Fig. 3. The Evolution of the Key by Contribution to the Clear Text.

Following this experimental study, we can see that the increase in the size of the key is very small (constant for the larger plaintexts) relative to the size of the message to be encrypted. These makes sense because having a larger text does not imply necessarily that it contains more characters than a small text. As a result, the average size of the key in our case is 2231 bits. However, in theory, the maximum size that can be reached is 8192 bits if we consider that the text to be encrypted contains all 256 possible characters. We can then say that the ASEC key size ensures resistance against brute force attacks and offers long term security.

Key generation:

The security of ASEC is not only related to the size of its key but to other strong points which lie primarily in the way in which it is generated, namely:

- ⇨ It does not use any key generation system.
- ⇨ The key is automatically generated by the system.
- ⇨ It is built through a non-deterministic algorithm [23].
- ⇨ It uses several probabilistic mechanisms that rely on random choices to decide the optimal solution [25].
- ⇨ Its size is variant.

Session key:

Each plaintext encrypted by the ASEC system has one and only one key which depends on its structure, its size and its nature. A change in one of these criteria gives birth to a new key. As a result, the same plaintext can lead to two different ciphertexts and this is achieved by changing the initial population based on the evolutionary algorithm [26].

Having a session key in our system allows extending authentication across the communication medium and preventing different attacks seeking to know the key [19]. Indeed, finding the key won't be very useful because it will be only used in the current transaction.

B. *Algorithm Performance*

Complexity:

The principle of evolutionary cryptographic algorithm is based on the idea of creating equiprobable partitions whose size is almost the same. This introduces the partition problem which is a difficult problem to solve. Normally, this kind of design is used in asymmetric ciphers. It Increases the complexity of solving this encryption exponentially. This makes the cryptosystem much more resistant to different types of attacks.

Avalanche Test:

To test the randomness of ASEC ciphering result, we are applied hamming distance between the input and output messages as in [20][21]. For each message Mi, We execute the ASEC Encrypting Algorithm with different Key bits changed. Then, we calculate the average of Hamming distance value between the message and all his Cipher text.

In fact, the Hamming distance of the cipher obtained should be a half of the output size (see Fig. 4).

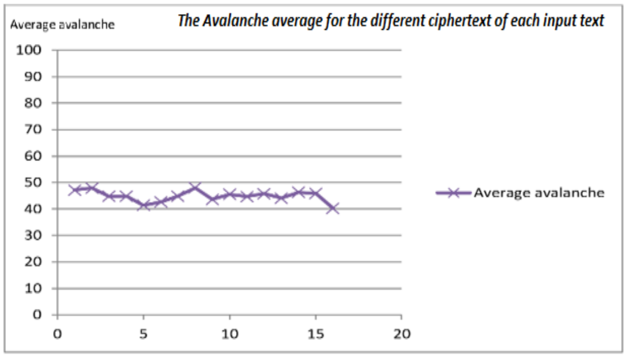


Fig. 4. Show the Obtained Hamming Distance for the different Cipher Text for each Message.

We conclude that the Hamming distance of the cipher obtained converge to the half of the output size average (Hamming(H(x),H(y)) ≈ n/2.

This result prove the randomly of output cipher.

Statistical test: Diehard

In order to bypass statistical attacks, we applied various statistical tests included in the DIEHARD package [28]. This platform offers all verifying tools of the different statically tests to demonstrate the strongest and efficacy of bit sequence ciphering generated by our system. It's also checking the randomness up to an extreme level.

Table II shows the result given by execution of DIEHARD tests on file that contains all ASEC ciphers message:

TABLE II. EXECUTION OF DIEHARD TESTS

Test	*P-value*	*Interpretation*
Diehard birthdays	0.50646335	PASSED
Diehard operm5	0.50241355	PASSED
Diehard_rank_32x32	0.65910598	PASSED
Diehard_rank_6x8	0.73626155	PASSED
Diehard_bitstream	0.75444424	PASSED
Diehard_opso	0.84942117	PASSED
Diehard_opso	0.59952027	PASSED
Diehard_dna	0.09103884	PASSED
Diehard_count_1s_str	0.94765782	PASSED
Diehard count1s byt	0.77998540	PASSED
Diehard parking lot	0.69671967	PASSED
Diehard 2d sphere	0.03413893	PASSED
Diehard 3d sphere	0.09723242	PASSED
Diehard squeeze	0.28015448	PASSED

VI. CONCLUSION AND PERSPECTIVES

The most common obstacles for the exchange of the keys is their generation and their transmission. In this work, we are designed a new Hybrid Cryptosystems that we called Hybrid Evolutionary Cryptosystems because it uses the Symmetric Evolutionary Ciphering ASEC. As we are shows and experiment it in this paper, the robustness of this system is lies to several factors that can be reduced as following:

- Key size: the key is so large and is sufficiently secure.
- No blocks: No need to split the message into blocks, the encryption and decryption are not based on the entire message. It can be likened to encryption algorithms block, where the block has a large dimension equal to its size, which increases its level of security and it allows also to avoid the propagation of errors likely by sending block by block.
- Random secret key generation.
- ASCII coding can be used and offer the compatibility with ASCII systems.
- The statistical attacks tests are conclusive because of the randomness of bit sequence ciphering generated by this system.

REFERENCES

[1] Florin G. Et Natkin S: Techniques Of Cryptography. Cnam 2002.

[2] Menezes A.J., Oorschot P.C. Van Et Vanstone S.A.: Handbook Of Applied Cryptography.(Crc Press, 1997).

[3] W. DIFFIE, M. E. HELLMAN, "New Directions in Cryptography " IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. IT-22, NO. 6, NOVEMBER 1976 Pp 644 –654

[4] Rivest, R., Shamir A., and Adleman L. A Method for Obtaining Digital Signatures and Public Key Cryptosystems, Communications of the ACM, 21(2) 1978 120-126.

[5] Zimmermann, P. R. (1991) PGP User's Guide, 5th June 1991, Version 1.0, Phil's Pretty Good Software.

[6] Adedeji Kazeem B. and Ponnle Akinlolu A. "A New Hybrid Data Encryption and Decryption Technique to Enhance Data Security in Communication Networks: Algorithm Development "- International Journal of Scientific & Engineering Research, Volume 5, Issue 10, October-2014

[7] P. Kuppuswamy , S. Q. Y. Al-Khalidi "Hybrid Encryption/Decryption Technique Using New Public Key and Symmetric Key Algorithm" - MIS Review Vol. 19, No. 2, March (2014), pp. 1-13

[8] A. Naser, H. Fatemeh and K. Riza, "Developing a new hybrid cipher using AES, RC4 and SERPENT for encryp-tion and Decryption", International Journal of Computer Applications, vol. 69, no. 8, pp.53-62, 2013.

[9] F.Omary, A.Tragha, A.Lbekkouri, A.Bellaachia, A.Mouloudi: An Evolutionist Algorithm To Cryptography- Brill Academic Publishers – Lecture Series And Computational Sciences Volume 4, 2005, Pp.1749-1752.

[10] F.Omary : Application Of Evolutionary Algorithms To Cryptography (Applications Des Algorithmes Evolutionnistes À La Cryptographie).Doctoral Thesis, University Mohammed V Agdal , Faculty Of Science - Rabat Marocco. (July 2006).

[11] Omary F., Mouloudi A., Tragha A., Bellaachia A. (2006) A New Ciphering Method Associated with Evolutionary Algorithm. In: Gavrilova M.L. et al. (eds) Computational Science and Its Applications - ICCSA 2006. ICCSA 2006. Lecture Notes in Computer Science, vol 3984. Springer, Berlin, Heidelberg.

[12] S.TRICHNI and al: A New Approach Of Mutation Operator Applied To The Ciphering System Sec. Iccit 2011,vol 63, no. 9;sep 2013.

[13] M. Bougrine, F. Omaiy, S. Trichni and B. Boulahiat, "New evolutionary tools for a new ciphering system SEC version," 2012 IEEE International Carnahan Conference on Security Technology (ICCST), Boston, MA, 2012, pp. 140-146, doi: 10.1109/CCST.2012.6393549.

[14] Deb K. Introduction To Selection. in: "Evolutionary Computation 1: Advanced Algorithms And Operators". Editor: Bäck T., Fogel D.B., Et

Michalewicz Z.; Institute Of Physics Publishing, Bristol And Philadelphia, 331 P. 2000.

[15] T. Back, ‘Evolutionary Algorithms In Theory And Practice’, Oxford University Press, Oxford, 1996.

[16] F.Omary : Application Of Evolutionary Algorithms To Cryptography (Applications Des Algorithmes Evolutionnistes À La Cryptographie).Doctoral Thesis, University Mohammed V Agdal , Faculty Of Science - Rabat Marocco. (July 2006).

[17] European Union Agency for Network and Information Security (enisa) : Algorithms, key size and parameters - report of 2014

[18] Arjen K. Lenstra, Eric R. Verheul : Selecting Cryptographic Key Sizes, Journal of Cryptology (2001) 14: 255–293 DOI: 10.1007=s00145-001-0009-4

[19] Agence nationale de la sécurité des systèmes d’information, Référentiel Général de Sécurité, version 2.0 : Choix et dimensionnement des mécanismes cryptographiques

[20] Echandouri, Bouchra & Omary, Fouzia & Ziani, Fatima Ezzahra & Sadak, Anas. (2018). SEC-CMAC A New Message Authentication Code Based on the Symmetrical Evolutionist Ciphering Algorithm. International Journal of Information Security and Privacy. 12. 16-26. 10.4018/IJISP.2018070102.

[21] Christophe Caux- Henri Pierreval- Marie-Claude Portmann : Genetic Algorithms And Their Application To Scheduling Problems (Les Algorithmes Genetiques Et Leur Application Aux Problemes D’ordonnancement). Apii Volume 29-N◦ 4-5/ 1995, Pp 409-443.

[22] Goldberg D.E: Genetic Algorithms In Search Optimisation & Machine Learning. Addison-Wesley Publishing Company,Inc,1989.

[23] Grefenstette J.J: Optimization Of Control Parameters For Genetic Algorithms. Ieeetrans. On Smc, Vol. 16, N◦ 1, Jan/Feb. 1986, Pp. 122-128.

[24] Khan Phang C: Heuristic And Evolutionary Algorithms. Doctoral Thesis, University Of Lille. (Octobre 1988).

[25] Shaul Drukmann: Evolutionary Algorithms.Encyclopedia Of Computational Neuroscience 2014, Pp 1-7.

[26] Mühlenbein H., And Schlierkamp-Voosen D."Predictive Models For The Breeder Genetic Algorithm-I, Continuous Parameter Optimization. Evolutionary Computation,1(1),25-49".1993

[27] Karthik .S, Muruganandam .A-“Data Encryption and Decryption by Using Triple DES and Performance Analysis of Crypto System”. International Journal of Scientific Engineering and Research (IJSER) - Volume 2 Issue 11, November 2014

[28] Georges Marsaglia. Diehard test suite. Online : http ://www. stat. fsu. edu/pub/diehard/. Laste visited, 8(01) :2014, 1998.

Factors Influencing Practice of Human Resource Information System in Organizations: A Hybrid Approach of AHP and DEMATEL

Abdul Kadar Muhammad Masum[1], Faisal Bin Abid[2]
ABM Yasir Arafat[3]
Department of Computer Science and Engineering
International Islamic University Chittagong
Chittagong, Bangladesh

Loo-See Beh[4]
Department of Administrative Studies and Politics
University of Malaya
Kuala Lumpur
Malaysia

***Abstract*—This paper blends the development of the Technology-Organization-Environment (TOE) framework and Human-Organization-Technology (HOT) fit model to identify the factors that influence the administration choice in embracing human resource information system (HRIS) in the organizations. Here, a hybrid Multi-Criteria Decision Making (MCDM) model combining the Decision Making Trial and Evaluation Laboratory (DEMATEL) and Analytic hierarchy Processes (AHP) is used to achieve the objective of the study. In this study, the experts agree that the staffs IT skill is most significant than other factors for the Human dimension. Similarly, IT infrastructure, top level support, and competitive pressure are the most vital factors for Technology, Organization and Environment dimensions respectively. Moreover, this paper will help the managers to take care of some factors that are vital for HRIS implementation in the organizations.**

Keywords—Analytic Hierarchy Processes (AHP); Decision Making Trial and Evaluation Laboratory (DEMATEL); factor; Human Resource Information System (HRIS); Multi-Criteria Decision Making (MCDM) Model

I. Introduction

Conventional human resource management (HRM) processes have been moved to HRIS in order to achieve the organizational objectives [1]. Specialists realized the magnitude of HRIS applications and investigated an expansive number of persuasive elements for the choice and usage of HRIS among business organizations [2]. Experts contended that the relative weight of investigated factors might be changed alongside gradual development and its settings. Besides, researchers unveiled that the technological innovation is consistently unpredictable in a competitive setting. And, it is crucial to comprehend the possible factors impacting the choice of HRIS selection and usage in the associations for a specific setting [3].

The technology innovation adoption concept has evolved globally [4] though there are still some constraints for appropriateness in developing settings of innovation models of Western countries [5, 6]. In compared to the West, Bangladesh is incredibly unique considering innovation, cultural, economic conditions as a developing nation. Thus, investigating the applicability model for HRIS adoption in developing countries is imperative. Past research indicates that there is a research gap in connection to the potential factors influencing the use of HRIS adoption in developing nations [7]. Therefore, the prime research target of this study is to exploring the most important factors of HRIS adoption among organizations of Bangladesh. Following this goal, we have some novelties in this paper. Firstly, the current study suggests an IT adoption model in the context of Bangladesh as there is a lack of theories of IT adoption for developing nations. Secondly, this paper reveals a thorough investigation assessing the essential level of interdependency among basic variables for the selection of HRIS usage in developing nations. Moreover, this study proposes an MCDM combining the DEMATEL and AHP approaches to assess and discover the significance level of the determining elements for HRIS usage in Bangladesh.

The remaining part of the paper is organized as follows. Section II labels the factors of human resource information system adoption based on Human-Organization-Technology (HOT) fit model and technological, organizational, and environmental (TOE) model. The research model and research methods and techniques that are adopted to achieve the objectives of the study are presented in Section III. Section IV includes data analysis along with some interesting results. A brief discussion on findings is provided in Section V, before concluding the paper in Section VI.

II. Literature Review

Researchers identified some factors for technological innovation in different settings. In this paper, we examined a few variables for IT innovation adoption. Both subjective and quantitative methodologies were utilized in technological adoption research. However, the subjective is a thoroughly detectable approach. It is apparent from exploring past researches that, the most significant and widely used fifteen variables are adopted dividing into four dimensions e.g. human (the characteristics of senior executives, Staffs IT skill), technology (comparative advantage, perceived compatibility, perceived complexity, and IT infrastructure), organization (top level support, organizational culture, organizational structure, organizational size, perceived cost), and environment (competitive pressure, support from technology service provider, government support)). The related research findings regarding the above variables are summarized below. Remarkably, in most past studies, HRIS and electronic human resource management (e-HRM) are used interchangeably [8].

In this study, the three variables of human dimension is characteristics of senior executives (innovativeness and IT knowledge), staffs IT skill, and employee Behavioral Characteristics. Adoption of a new system is an strategic decision of an organization [9]. Therefore, the decision towards adopting or rejecting an innovation is depends on personal attributes of top managers and skill on IT [10]. On the contrary, a couple of studies asserted that senior officials' IT learning and creativity does not influence IT adoption in organizations [11]. Previous research of Alam, Masum, Beh and Hong [5] revealed that organizations having staff with IT learning background create probability of IT enabled HRM applications. In line with previous research, Teo, Lim and Fedric [4] and Bian [12] showed IT ability and employee intention to use IT applications are strong indicators to choose HRIS in organizations of China. However, a few scholars found that IT abilities of staff was unimportant factor for technological innovations considering multiple settings [2, 13].

The organization dimensions include five variables such as top level support, organizational culture, organizational structure, organizational size, and perceived cost. Some researchers identified the support of top management as a vital factor for IT application use [2, 4]. Conversely, few studies on adoption of IT have claimed that the support from top level is not always a subject of influence to adopt IT in organization. [11, 14]. Various scholars explored the variable - Organizational culture to be an important factor for the selection of IT applications [15, 16]. In addition, Cooper and Schindler [17] resolved that incase of any clashes in information system (IS) of an organization the software is become abandoned or customized so that it matches the existing culture of business organizations. Conversely, some researchers argued that organizational culture is not an essential divider amongst adopters and non-adopters of modern IT application [18]. Moreover, earlier studies on 110 manufacturing firms in Singapore confirmed that organizational size is the only constantly accepted factor among the possible causes for HRIS use in the organizations [4]. Scholars contended that if expenses are perceived to be high, people will be less inclined to invest in the selection of HRIS [5]. A contemporary study on the adoption of internet business in Iranian SMEs uncovered that the budgetary viewpoint (high expenses) is the sole issue behind not embracing web-based business applications [19].

The four variables of technology dimension such as perceived compatibility, perceived complexity, comparative advantage, and IT infrastructure are widely established technological factors in IT innovation research. Teo, Lim and Fedric [4] indicated that choice of HRIS adaption is strongly connected to an amiable impression of HRIS in human resources (HR) department. Similarly, Al-Dmour [2] claimed the significance of perceived relative advantage on HRIS usage. Whereas perceived compatibility was evident to be highest influencer on the use of HRIS [4]. Likewise, Ghobakhloo, Arias-Aranda and Benitez-Amado [19] recommended perceived compatibility as key component of IT application in adoption stage. In this regards, researchers concluded that the successful adoption of new technology suffers from a high extent of uncertainty and risk due to perceived complexity For this reason Gutierrez, Boukrami and Lumsden [14] mentioned that, in the UK, perceived complexity has a strong influence in the implementation of distributed computing services. Conversely, Teo, Lim and Fedric [4] and Bian [12] found it as a non-discriminate factor among the firms. And, few researchers have claimed that compatibility does not influence innovation adoption[11]. In additions, researchers uncovered that perceived complexity is not a vital factor for HRIS or related IT appropriation choice in firms or organizations [4, 11]. Recent studies indicated that IT infrastructure is the most common barrier to adopt IT enabled applications in organizations of the developing countries [1, 20]. Conversely, some current research showed that IT infrastructure is an insignificant factor for IT innovation adoption [5, 21].

Lastly, environmental dimension consist three variables like competitive pressure, support from technology service provider and government support. Ghobakhloo, Arias-Aranda and Benitez-Amado [19] recommended that competitive pressure is one of the most important variables taken into account for deciding to implement IT strategies within an organization. Additionally, Bian [12] discovered that pressure from competitor was a critical factor for using HRIS in China. In Jordanian firms, Al-Dmour [2] stated that the availability of IT providers and their sponsorship is the most worthy factor in an environmental context for using HRIS. Conversely, various studies on IT adoption have concluded that support from technology service provider is not identified to be a major factor [11, 22]. A present study on the adoption of HRIS in public organizations of Australia uncovered that the government rule and regulations is very important determinants for HRIS adoption and implementations [23]. Similarly, Al-Dmour [2] claimed that government policy regarding IT applications adoption is one of the significant factors for HRIS implementation in firms of Jordan.

In contemporary research on organizations of Bangladesh, researchers claimed that organization, environment and technology are observed to be the most persuasive dimensions for HRIS adoption [5]. The authors obtained five most critical factors for HRIS adoption in organizations. And, these factors are IT infrastructure, Staffs IT capabilities, top level support, perceived cost, and competitive pressure. Likewise, Masum [1] stated that the foremost obstruction for executing HRIS in developing countries is the lack of management knowledge, experience, training, and fear of expense. But, the advantages of the HRIS is outweigh the limitations [24]. The past studies indicate that research on HRIS in Bangladesh is in early-stage for Bangladesh setting [25].

Amalnick, Ansarinejad and Nargesi [26] adopted DEMATEL and ANP evaluated some important factor for successful implementation of ERP systems. They explored that ERP vendor selection, project team, project/business plan and business model, management, and budgeting made up the top five success factor for ERP implementation in the organizations. In a recent research, scholars employed AHP with Fuzzy AHP (FAHP) and group decision-making (GDM) to identify the diversified factors from different dimensions of the web-based E-Learning system in university context [27]. A study on an Iranian steel company, Rouhani, Ashrafi and Afshari [28] evaluated CSFs in ERP using a hybrid model based on fuzzy DEMATEL and fuzzy AHP. The results exposed that project champion, clear project plan, project team competence, training and education and organizational culture were among the most significant 8 aspects to be considered.

Interestingly, most of the past studies depicted descriptive statistics, absence of robust inferential statistics, and advanced artificial intelligence (AI) tools and techniques are not used. So, research gap is visible and that need be resolved by different types of AI techniques such as MCDM Model Approach. MCDM is one of the famous techniques to manage complex issues that show high vulnerability, conflicting goals, different interests and numerous points of view [29]. Moreover, MCDM approaches are viable in decision making, weighing and choosing the most fitting options.

III. Research Methods

Technology-Organization-Environment (TOE) structure is the widely used model in technology adoption at organizations. The model consists of triple dimensions such as technological, organizational and environmental. But, the important limitation of this model is no consideration of human factors such as IT skill and innovative knowledge of managers and employees. Reversely, the latest model of IT application adoption named Human-Organization-Technology (HOT) fit at organizational level also suffers from the limitation to explaining the total modalities of IT enabled applications such as HRIS adoption. Particularly, the model not recognized external (environmental) factors though environmental factors greatly influence the organizational systems. So, theoretical gap is identified. Thus, the most relevant factors of HRIS adoption at the organizational level may significantly predict combining four dimensions (i.e., human, organizational, technological, and environmental) in context of Bangladesh.

Through extensive literature review, 15 factors were chosen. The factors are characteristics of senior executive (CHAR_SEC), staffs IT skill (SIS), Employee Behavioral Characteristics (EMPB_CHAR), Comparative Advantage (COM_ADV), Perceived Compatibility (PER_COMPA), Perceived Complexity (PER_COMPL), IT Infrastructure (IT_INFRAS), Top level support (TLS), Organizational Culture (OC), Organizational Structure (OST), Organizational Size (OSIZE) , Perceived Cost (PER_COST), Competitive Pressure (CP), support from technology service provider (TVS), Government Support (GOV_SUP). A hybrid MCDM model is utilized to identify the interconnections of variables with the AHP and DEMATEL strategies.

Data was collected using questionnaire survey. And, the questionnaire attempted to obtain information relevant information related to the research such as factors of HRIS adoption decision, extent of usage HRIS application etc. We have selected 15 human resource managers from 15 reputed manufacturing organizations of Bangladesh, who have adequate knowledge and expertise on HRIS. In order to attain anonymity and free from biasness, the self-administered structured questionnaire was employed. For furtherance of the research the survey method was used as the sampled elements and the variables were treated having no to control, influence, or manipulate them.

AHP, one of the famous MCDM techniques is used to analyze complex decisions for broader application areas. In the literal context, DEMATEL is adjusted for various subjective and factor-related points including modern industrial arranging, basic leadership, feasible development and other world issues [30]. Realizing the benefits of hybrid MCDM especially integrating DEMATEL and AHP techniques, scholars used them in different settings such as product selection, supplier selection, software selection, and so on [31]. DEMATEL is a helpful strategy in order to analyze cause-effect connections between factors and dimensions. But DEMATEL itself can't decide the weights of individual criteria, where AHP proves to be useful. The two techniques offer help in managing complex issues, as decision producers can have a superior understanding of the issues to be unraveled. As indicated by Horng, Liu, Chou, Yin and Tsai [32], combining DEMATEL and AHP techniques altogether can give a supportive tool for recognizing the basic traits of policy arrangement usage as well as computing the weights of the business decision criteria. Thus, DEMATEL and AHP strategies are embraced to fulfill the goals of this study.

IV. Results

As we discussed in the previous section, a conceptual research strategy has been developed in this study. Consequently, in this segment, we constructed a combination of MCDM models for the procedure of HRIS adoption selection. The proposed MCDM model includes two principle stages namely AHP and DEMATEL.

A. *Analytic Hierarchy Process (AHP)*

Based on the inputs, taken from the respondents, AHP is used to derive ratio scales from pairwise assessments. The consistency ratio was the parameter used for finding out the accuracy of the respondents. In case of being inconsistent, the respondents were told to submit their responses once again. The stepwise procedure is given below:

Step 1: Respondents provide pairwise comparison of dimensions and variables. The ranking scale is depicted using a scale of 1 to 9, where 1 indicates same preference and 9 represents the intense preference of dimensions, and variables. Meanwhile, ranking 3, 5, and 7 indicate little, strong and very strong preference. Based on the given input by the respondents, an n×n reciprocal matrix has been conducted where n represents dimensions of HRIS adoption. Sample input for the respondent and the matrix has been given in Fig. 1. The dimensions are represented in a short form such as Org, Hum, and Tech, Env for organization, human, technology and environment respectively. All elements in the reciprocal matrix should be greater than zero, i.e $a_{ij}>0$ where a_{ij} represents an element inside the matrix of row i and column j.

Step 2: Sum up each and every column of the reciprocal matrix. Divide every component of the matrix with the summation of its column. As a result, normalized relative weight matrix has been obtained. Whenever an average of the rows is being done, we can find the normalized principal

Eigenvector or priority vector. The priority vector can be depicted by multiplying by 100 to get the percentage of preference of the elements.

Step 3: Find out the consistency (CI) of the respondents: CI=$\frac{\lambda\max-n}{n-1}$ where λ_{max} is the principal Eigen value determined by summing up products between each element of Eigen vector and summation of columns of reciprocal matrix.

λ_{max} can be computed using the following formula:

$\sum w_i x_j$ where x_j= sum of column of comparison matrix and w_i = priority vector , here i=j=n.

Step 4: If n>2, calculate consistency ratio (CR) using CI and Random consistency index (RCI) proposed by Prof Satty given solution using 500 matrices as sample size. The formula of CR is given as below:

$$CR = \frac{CI}{RI}$$

Step 5: If CR<10%, the judgment is considered to be consistent, otherwise the respondent was requested to reconsider preferences once again.

$$\text{Expert} = \begin{array}{c|cccc} & \text{Org} & \text{Hum} & \text{Tech} & \text{Env} \\ \hline \text{Org} & 1 & 3 & 7 & 9 \\ \text{Hum} & 0.33 & 1 & 5 & 7 \\ \text{Tech} & 0.142 & 0.2 & 1 & 3 \\ \text{Env} & 0.11 & 0.14 & 0.33 & 1 \end{array}$$

Fig. 1. Sample Matrix for four Dimensions Given by the Expert11.

TABLE I. USING AHP TO FIND OUT THE RANKING OF DIMENSIONS

Experts	Technology	Organization	Environment	Human
Expert 1	9.02	57.39	4.44	29.13
Expert 2	8.39	65.77	5.17	20.66
Expert 3	10.78	56.28	6.21	26.71
Expert 4	11.03	52.54	6.33	30.08
Expert 5	12.18	55.78	5.68	26.33
Expert 6	12.18	55.78	5.68	26.33
Expert 7	9.99	67.15	5.98	16.85
Expert 8	12.18	55.78	5.68	26.33
Expert 9	9.02	57.39	4.44	29.13
Expert 10	8.39	65.77	5.17	20.66
Expert 11	10.78	56.28	6.21	26.71
Expert 12	11.03	52.54	6.33	30.08
Expert 13	12.18	55.78	5.68	26.33
Expert 14	12.18	55.78	5.68	26.33
Expert 15	9.02	57.39	4.44	29.13
Sum of score	158.44	867.48	83.21	390.85
Average	10.56%	57.83%	5.54%	26.05%
Rank	3	1	4	2

Step 6: Average the priority judgment of the 15 respondents (Expert 1 … Expert 15) and find out the highest value to be ranked 1 and so on for dimensions and respective variables in Table I. Thus, the highest and lowest priority of the dimensions as well as variables can be found out. The rankings are represented from higher ranking to lower ranking. It can be clearly seen that the organization dimension receives the highest average value (57.83%) whereas the environment receives the lowest average value (5.54%). Thus, organization is ranked as the most preferred one (ranked 1st) whereas environment is the least preferred one (ranked 4th) among the respondents. Moreover, human is ranked at 2nd with 26.05% and technology is ranked at 3rd with 10.56%.

B. *Decision Making Trail Evaluation Laboratory (DEMATEL)*

DEMATEL algorithm has been used in the next section in order to find out the most important causes as well as the correlations among variables. It reveals the influence between the variables. The steps have been described as follows:

Step 1: Respondents will provide the influence based on the factors (dimensions and variables). The influence will be based on the assumptions given in step 1 of Section 3.1. Based on the ranking, answer matrices (let's name it as x) will be constructed. If there are H respondents and n factors, then there will be H matrices each comprised of n factors. Thus there will be H matrices each composed of n×n factors.

Step 2: Construct an initial direct relation matrix. It is also known as average influence qualification. Each and every component will be measured as a_{ij}= $\frac{\sum x_{ij}\text{H}}{\text{n}}$ where i and j represent row and column of answer matrix and n is the number of factors. Sum up all the components of each and every row and take the largest sum of the row as G. Now the value of G will be used to generate normalized direct relation matrix (D).

$$\text{G} = \max_{1\leq i\leq n} \sum_{j=1}^{n} a_{ij}$$

Step 3: Direct relation matrix's Normalized form D will be computed by dividing all the elements of A by G.

Step 4: Total relation matrix will be worked out using the following formula: T=D (I-D)-1. The total relation matrix of the four variables has been shown below.

Step 5: (special condition): If i= = j, then ,ri+cj indicates the total effect given and received by factor i and ri– cj indicates net effect factor i has on the system.

If ri– cj>0, factor i is a causer, else Factor i is a receiver, it is also known as effect.

The Total relation matrix for the Variables of Technology dimension is:

$$\text{Tech=} \begin{array}{ccccccc} & \text{V1} & \text{V2} & \text{V3} & \text{V4} & r_i+c_i & r_i\text{-}c_i \\ \text{V5} & 0.1072 & 0.1913 & 0.0687 & 0.0278 & 1.1936 & -0.4036 \\ \text{V6} & 0.0646 & 0.1064 & 0.0266 & 0.0164 & 1.6767 & -1.2487 \\ \text{V7} & & & & & & \\ \text{V8} & 0.1879 & 0.4559 & 0.1074 & 0.0708 & 1.2031 & 0.4409 \\ \text{V9} & 0.4389 & 0.7091 & 0.1783 & 0.1066 & 1.6543 & 1.2113 \end{array}$$

Where, V1= COM_ADV, V2= PER_COMPA, V3= PER_COMPL, V4= IT_INFRAS

The Total relation matrix for the Variables of Organization dimension is:

$$\text{Org} = \begin{array}{cccccccc} & V5 & V6 & V7 & V8 & V9 & r_i+c_i & r_i-c_i \\ V5 & 0.0899 & 0.6443 & 0.4056 & 0.1565 & 0.2379 & 1.7515 & 1.3168 \\ V6 & 0.0136 & 0.0899 & 0.0515 & 0.0197 & 0.0183 & 2.0489 & -1.6629 \\ V7 & 0.0220 & 0.1679 & 0.0897 & 0.0244 & 0.0486 & 1.4268 & -0.7215 \\ V8 & 0.0545 & 0.4502 & 0.3518 & 0.0895 & 0.1590 & 1.4501 & 0.7599 \\ V9 & 0.0373 & 0.5037 & 0.1757 & 0.0549 & 0.0898 & 1.4152 & 0.3078 \end{array}$$

Where, V5= TLS, V6= OC, V7= OST, V8= OSIZE V9= PER_COST

The Total relation matrix for the Variables of Human dimension is:

$$\text{Hum=} \begin{array}{cccccc} & V10 & V11 & V12 & r_i+c_j & r_i-c_j \\ V10 & 0.1745 & 0.0839 & 0.4764 & 1.3845 & 0.0851 \\ V11 & 0.4013 & 0.1742 & 0.9675 & 1.8370 & 1.2489 \\ V12 & 0.0738 & 0.0359 & 0.1755 & 1.9045 & -1.3340 \end{array}$$

Where, V10= CHAR_SEC, V11= SIS, V12=EMPB_CHAR

The Total relation matrix for the Variables of environment dimension is:

$$\text{Env} = \begin{array}{cccccc} & V13 & V14 & V15 & r_i+c_j & r_i-c_j \\ V13 & 0.1326 & 0.4735 & 0.8160 & 1.6171 & 1.2272 \\ V14 & 0.0402 & 0.1328 & 0.3137 & 1.1518 & -0.1783 \\ V15 & 0.0222 & 0.0587 & 0.1320 & 1.4747 & -1.0489 \end{array}$$

Where, V13=CP, V14=TVS, V15=GOV_SUP

C. *Net Effect and Total Effect of Variables*

In the technology dimension, the value of r-c is greater than 0 for V3 and V4. More specifically, the value of perceived complexity and IT infrastructure are 0.4409 and 1.2113. So, these are the major causes of the technology dimension and should not be overlooked. The higher value of r+c indicates higher prominence. The Value of r+c is highest for variable V2, namely perceived compatibility, revealing the fact that this variable has major interaction with other variables compared to the interaction of other variables with this variable.

Same rule applies for organization, human and environment dimensions' variables. It can be unequivocally seen in the organization dimension that variable V5, top level support is the most important cause while variable V6, organizational culture has the strongest correlation than any other criteria for the organizational dimension. The variable V4, IT infrastructure is the most important cause in the technological dimension. Variable V11, Staffs IT skill is the most significant cause that has impact on other variables whereas V12, employee behavioral characteristics have major correlations with all other factors in the system in terms of human dimension. Turning to dimension environment, the variable competitive pressure is the significant causer affecting other variables.

In nutshell, top level support is the most crucial factor in the organization dimension and Staffs IT skill is the most significant in terms of human dimension. Moreover, IT infrastructure is the most vital factor in the technology dimension. Turning to environment dimension, variable competitive pressure has been considered to be having the prominent significance.

V. Discussions

This research adds new ideas to existing HRIS literature. The investigation explores the impacting variables on decision making for adopting HRIS in the industries of Bangladesh. The exploration of factors based on the theoretical model overcomes the limitations explained in the HOT-fit model and the TOE model. Analyzing the respondents' opinion, the research explored that organization is ranked as the most important one (ranked first) whereas environment is the least important one (ranked fourth). Moreover, human is ranked at second and technology dimension is ranked at third.

In human dimension, IT expertise of staff is recognized as the most significant driver that inducing HRIS usage in the organizations. This outcome is consistent with past studies [5, 12, 33]. In order to continue consistent growth and keeping long term focus in business organizations, the availability of expert HRIS is a vibrant factor in the IT-dependent HRM applications. In a contemporary study, researchers quantified that personnel who have both technical and managerial skills are considered as the most valuable asset for a firm [25]. It also shows the importance of technology readiness for any technological acceptance. So, HR professionals should be enriched with multidisciplinary knowledge of application based IT and the functions of HR to ensuring effective HRIS operations in organizations. However, this result rejects some contemporary studies where IT expertise of staff was identified as an insignificant driver for IT innovation adoption [2, 13].

In organization dimension, this exploration finds that the top level support is factually critical determiner for HRIS utilization. In this way, administration support from higher authority is vital for distribution of recourse and to motivate people to use the system. The finding also supports past studies [2, 4, 5, 12]. So, active support of top management speeds up the HRIS adoption in the organizations. Furthermore, to success the adoption project, top management provides sufficient resources such as people, materials, capital, and related support. The outcome of this study indicates that adopting a new technology in organizations will be easier to the organizations when the senior executives support the innovation adoption. Moreover, sometimes, innovation adoption encounter barrier within the organizations. At that time, top management helps to overcome these problems. So, this study strongly recognized the importance of top level support that significantly influences the incumbents to use HRIS in routine HRM tasks Conversely, this result rejects some contemporary studies where top level support was identified as an non-significant driver for HRIS implementation and usage [11, 14].

In technology dimension, the outcome demonstrates that IT infrastructure is positioned as highest -ranked factor to HRIS selection in the organizations of Bangladesh. Also, this variable was discovered to have huge impact in earlier research on IT application selection [34, 35]. Likewise, Masum [36] additionally uncovered IT infrastructure as a critical success factor in organizations of Bangladesh. The contemporary research shows that it progressively becomes important to ensure that the HRIS fits in with the existing IT infrastructure for information systems used in an organization. Also, HRIS modules need easy to understand and user friendly interface. In almost all developing countries, related costs to IT, necessary infrastructure, and the quality of these infrastructures hinders the adoption of IT applications in the organizations. Nevertheless, this result rejects some current studies where IT infrastructure was acknowledged as an insignificant factor for IT innovation adoption [5, 21].

From environment dimension, competitive pressure is signified to be dominant drivers for HRIS use in organizations. It implied that organizations are feeling pressure for using new technologies such as HRIS to gain competitive advantage and to achieve its goals .The results of previous research support the finding of the current study [14, 37, 38]. So, this study suggests that organization should adopt proper strategies to handle the competitive pressure as; nowadays, new technology adoption crying need of survival in the industry. In earlier research, researchers stated that a company cannot perceive competitive advantage without properly managing their human resource and IT applications [25]. Currently, organizations are adopting HRIS to support in getting maximum results from their employees, making better-informed decisions and streamlining HR processes and better distribution of human resources. Therefore, these things encourage other organizations to adopt and use the HRIS as well as related IT applications are being competitive in the holistic spectrum of business. However, the present research confronts findings of Ahmad, Abu Bakar, Faziharudean and Mohamad Zaki [39], Ahmadi, Nilashi and Ibrahim [11], and Teo, Lim and Fedric [4]. The researchers stated that competitive pressure is an insignificant factor for HRIS adoption and usage.

VI. Conclusions

This paper combines two theories of adoption such as HOT-fit model and TOE framework to recognize the factors influencing the organizations of emerging country like Bangladesh to implement and usage of HRIS for managing human resources efficiently and effectively. The findings of the research will enhance the managers to identify the issues that are related to adoption, implementation, and usage. As far our knowledge goes, organization-level adoption behavior of HRIS using MCDM techniques has never been examined in Bangladesh. Thus, this paper added value to the HRM discipline by improving present understanding of HRIS adoption issues, which is an unsearched field in Bangladesh. This study only applied 15 determinants to inspect the decision of use HRIS. Importantly, some others relevant factors may also influence HRIS usage such as pressure from trading partners, data security, government policy, and information intensity etc. For further research, the findings of the study might be used as a generalized model for the developing countries to take decision adopting HRIS in the organizations.

References

[1] K. M. Masum, “Adoption Factors of Electronic Human Resource Management (e-HRM) in Banking Industry of Bangladesh,” Journal of Social Sciences, vol. 11, no. 1, pp. 1, 2015.

[2] R. H. Al-Dmour, “An integration model for identifying the determinants of the adoption and implementation level of HRIS applications and Its effectiveness in business organisations in Jordan ”, Computer Science, Brunel University London, UK, 2014.

[3] A. S. Narayana, and R. L. Bhusal, "Adoption and Use of Human Information System Digital Technology for Organizational Competitiveness: An Exploratory Study in the Context of Nepal," Handbook of Research on Social and Organizational Dynamics in the Digital Era, pp. 250-275: IGI Global, 2020.

[4] T. S. Teo, G. S. Lim, and S. A. Fedric, “The adoption and diffusion of human resources information systems in Singapore,” Asia Pacific Journal of Human Resources, vol. 45, no. 1, pp. 44-62, 2007.

[5] M. G. R. Alam, A. K. M. Masum, L.-S. Beh, and C. S. Hong, “Critical Factors Influencing Decision to Adopt Human Resource Information System (HRIS) in Hospitals,” PloS one, vol. 11, no. 8, pp. e0160366, 2016.

[6] A. Davarpanah, and N. Mohamed, “Human Resources Information Systems Implementation and Influences in Higher Education: Evidence From Malaysia,” International Journal of Asian Business and Information Management (IJABIM), vol. 11, no. 3, pp. 65-84, 2020.

[7] A. K. M. Masum, L.-S. Beh, M. A. K. Azad, and K. Hoque, “Intelligent human resource information system (i-HRIS): a holistic decision support framework for HR excellence,” Int. Arab J. Inf. Technol., vol. 15, no. 1, pp. 121-130, 2018.

[8] T. Bondarouk, E. Parry, and E. Furtmueller, “Electronic HRM: four decades of research on adoption and consequences,” The International Journal of Human Resource Management, vol. 28, no. 1, pp. 98-131, 2017.

[9] C. R. Greer, Strategic human resource management, 2001.

[10] Y. Alshamaila, S. Papagiannidis, and F. Li, “Cloud computing adoption by SMEs in the north east of England: A multi-perspective framework,” Journal of enterprise information management, vol. 26, no. 3, pp. 250-275, 2013.

[11] H. Ahmadi, M. Nilashi, and O. Ibrahim, “Organizational decision to adopt hospital information system: An empirical investigation in the case of Malaysian public hospitals,” International journal of medical informatics, vol. 84, no. 3, pp. 166-188, 2015.

[12] L. Bian, “An Empirical Study on Factors that Influencing the Adoption of Electronic Human Resource Management (E-HRM) Among Firms in Northeast of China,” University of Malaya, 2012.

[13] C. Low, Y. Chen, and M. Wu, “Understanding the determinants of cloud computing adoption,” Industrial management & data systems, vol. 111, no. 7, pp. 1006-1023, 2011.

[14] A. Gutierrez, E. Boukrami, and R. Lumsden, “Technological, organisational and environmental factors influencing managers’ decision to adopt cloud computing in the UK,” Journal of Enterprise Information Management, vol. 28, no. 6, pp. 788-807, 2015.

[15] S. Jackson, “Organizational culture and information systems adoption: A three-perspective approach,” Information and Organization, vol. 21, no. 2, pp. 57-83, 2011.

[16] Y.-M. Wang, Y.-S. Wang, and Y.-F. Yang, “Understanding the determinants of RFID adoption in the manufacturing industry,” Technological forecasting and social change, vol. 77, no. 5, pp. 803-815, 2010.

[17] D. R. Cooper, and P. Schindler, Business research methods, 8th ed., Irwin, Boston: McGraw-Hill, 2003.

[18] T. S. Teo, S. Lin, and K.-h. Lai, “Adopters and non-adopters of e-procurement in Singapore: An empirical study,” Omega, vol. 37, no. 5, pp. 972-987, 2009.

[19] M. Ghobakhloo, D. Arias-Aranda, and J. Benitez-Amado, “Adoption of e-commerce applications in SMEs,” Industrial Management & Data Systems, vol. 111, no. 8, pp. 1238-1269, 2011.

[20] O. Mothobi, and L. Grzybowski, “Infrastructure deficiencies and adoption of mobile money in Sub-Saharan Africa,” Information Economics and Policy, vol. 40, pp. 71-79, 2017.

[21] G. Kannabiran, and P. Dharmalingam, “Enablers and inhibitors of advanced information technologies adoption by SMEs,” Journal of Enterprise Information Management, 2012.

[22] R. Rahayu, and J. Day, “Determinant Factors of E-commerce Adoption by SMEs in Developing Country: Evidence from Indonesia,” Procedia-Social and Behavioral Sciences, vol. 195, pp. 142-150, 2015.

[23] I. Troshani, C. Jerram, and S. R. Hill, “Exploring the public sector adoption of HRIS,” Industrial Management & Data Systems, 2011.

[24] A. K. M. Masum, M. J. Kabir, and M. M. Chowdhury, “Determinants that influencing the adoption of E-HRM: An empirical study on Bangladesh,” Asian Social Science, vol. 11, no. 21, pp. 117, 2015.

[25] A. K. M. Masum, M. G. R. Alam, M. S. Alam, and M. A. K. Azad, "Adopting factors of electronic human resource management: Evidence from Bangladesh." pp. 1-4.

[26] M. S. Amalnick, A. Ansarinejad, and S.-M. Nargesi, “New perspective to ERP critical success factors: Priorities and causal relations under fuzzy environment,” Journal of Mathematics and Computer Science, vol. 2, pp. 160-170, 2011.

[27] Q. N. Naveed, M. R. N. Qureshi, N. Tairan, A. Mohammad, A. Shaikh, A. O. Alsayed, A. Shah, and F. M. Alotaibi, “Evaluating critical success factors in implementing E-learning system using multi-criteria decision-making,” Plos one, vol. 15, no. 5, pp. e0231465, 2020.

[28] S. Rouhani, A. Ashrafi, and S. Afshari, “Segmenting critical success factors for ERP implementation using an integrated fuzzy AHP and fuzzy DEMATEL approach,” World Applied Sciences Journal, vol. 22, no. 8, pp. 1066-1079, 2013.

[29] J.-C. Pomerol, and S. Barba-Romero, Multicriterion decision in management: principles and practice: Springer Science & Business Media, 2012.

[30] C.-Y. Huang, J. Z. Shyu, and G.-H. Tzeng, “Reconfiguring the innovation policy portfolios for Taiwan's SIP Mall industry,” Technovation, vol. 27, no. 12, pp. 744-765, 2007.

[31] F.-Y. Pai, “Analyzing consumers' decisions to select micro-invasive aesthetic service providers using a hybrid method,” Applied Mathematics & Information Sciences, vol. 8, no. 6, pp. 3071, 2014.

[32] J.-S. Horng, C.-H. Liu, S.-F. Chou, Y.-S. Yin, and C.-Y. Tsai, “Developing a novel hybrid model for industrial environment analysis: A study of the gourmet and tourism industry in Taiwan,” Asia Pacific Journal of Tourism Research, vol. 19, no. 9, pp. 1044-1069, 2014.

[33] A. K. M. Masum, M. J. Kabir, and M. M. Chowdhury, “Determinants that Influencing the Adoption of E-HRM: An Empirical Study on Bangladesh,” Asian Social Science, vol. 11, no. 21, pp. p117, 2015.

[34] H. Sulaiman, and N. Wickramasinghe, “Assimilating healthcare information systems in a Malaysian Hospital,” Communications of the Association for Information Systems, vol. 34, no. 1, pp. 66, 2014.

[35] W.-H. Hung, I. Chang, D. C. Yen, and C.-M. Lee, “Critical Factors of Adopting Enterprise Application Integration Technology: An Empirical Study on Larger Hospitals,” Communications of the Association for Information Systems, vol. 36, no. 1, pp. 31, 2015.

[36] A. K. M. Masum, “Adoption Factors of Electronic Human Resource Management (e-HRM) in Banking Industry of Bangladesh,” Journal of Social Sciences/Sosyal Bilimler Dergisi, vol. 11, no. 1, 2015.

[37] N. Al-Qirim, “The adoption of eCommerce communications and applications technologies in small businesses in New Zealand,” Electronic Commerce Research and Applications, vol. 6, no. 4, pp. 462-473, 2008.

[38] B. Ramdani, D. Chevers, and D. A. Williams, “SMEs' adoption of enterprise applications: A technology-organisation-environment model,” Journal of Small Business and Enterprise Development, vol. 20, no. 4, pp. 735-753, 2013.

[39] S. Z. Ahmad, A. R. Abu Bakar, T. M. Faziharudean, and K. A. Mohamad Zaki, “An Empirical Study of Factors Affecting e-Commerce Adoption among Small-and Medium-Sized Enterprises in a Developing Country: Evidence from Malaysia,” Information Technology for Development, vol. 21, no. 4, pp. 555-572, 2014.

A Review on Virtual Machine Positioning and Consolidation Strategies for Energy Efficiency in Cloud Data Centers

Nahuru Ado Sabongari[1], Dr. Abdulsalam Ya'u Gital[2], Prof. Souley Boukari[3]
Badamasi Ja'afaru[4],Muhammad Auwal Ahmed[5], Dr. Haruna Chiroma[6]
Dept.of Mathematical Sciences
A.T.B.U Bauchi
Bauchi, Nigeria

Abstract—**The cloud data center consumes massively more and more energy which is considered inacceptable. Therefore further efforts are needed to improve the energy efficiency of such data centers by using Server Consolidation to minimize the number of Active Physical Machines (APMs) in a data center setting. Strategies for positioning and transformation of VM maintain their usefulness as a roadmap to maximum consolidation. The latest techniques do complex restructuring, thus optimizing VM's positioning. The paper provides a detailed state-of - the-art strategies for VM positioning and consolidation that help improve energy efficiency in cloud data centers. A comparison is provided here between the strategies that revealed the worthiness, limitations and suggestions of strengthening other methods along the way.**

Keywords—Energy efficiency; optimization; cloud data centers

I. Introduction

Cloud data centers result in high energy consumption and a significant amount of carbon footprints are generated which can be described as the 21st century's biggest challenge. The data center environment is the network that physically houses Cloud computing resources and services (L. Zhang, Yin, Li, & Wu, 2015). One of the main reasons for cloud computing's diverse views is that while new technical ideas, new technology, cloud computing has a traditional operating model that brings together a variety of current business management technologies [1, 2]. With a lot of affordable cloud services, and pay as you go. There is a need to at all times have a cheaper, secure, open service with a high demand for cloud service infrastructure and pay as per you go service[3]. Three significant services are made available by the cloud to the user via the Internet. Computing infrastructure as a service (IaaS), with services such as Amazon Elastic Compute Cloud, being supported. Platform as a service (PaaS) to an application of runtime applications, like the Google App Engine[4]. Though Salesforce.com, for instance[5], Software as a Service (SaaS).Although most of these services are provided through virtualization. Virtualization enables the multiple occurrences method to run on one computer[6].Making them shareable among multiple physical users is an abstraction over physical resources [7]. Physical resources are homogeneously virtualized and are therefore efficient for parallel and distributed computing [8]. Distributed cloud systems usually consisting of distributed interconnected data centers, thus using virtualization technologies to provide computing and storage resources for each request on demand [9].

Although cloud computing makes it easier for companies to benefit greatly from lowering operational and administrative costs, the situation suffers from the issue of high energy usage, which could reduce its benefits [10, 11]. Current studies suggest that data centers produce 78.7 million tons of CO2, 2% of global emissions [12]. CDCs used up to 100 billion kilowatt hours (kWh) in 2015, sufficient for Washington City in the United States alone[12, 13]. By 2022, this high electricity consumption will spread 150 billion kWh, with a 50 per cent increase[14, 15]. If measuring instruments are not, this energy consumption will increase by 2030 in CDCs to 8,000 terawatt hours (TWh). Several prominent cloud providers, including Google, Amazon, Microsoft and IBM, are positive about achieving zero carbon footprint growth and are looking for new ways to render environmentally friendly CDCs and cloud-based services [16, 17]. Such extraordinary energy consumption will lead to excessive carbon dioxide (CO2) emissions which contribute to global warming. VM consolidation is one of the most successful and enabling strategies to reduce energy footprints in cloud data centers [18].Virtualization offered support with the coming cloud computing, by which it further corroborated the energies for energy-efficient computing. Reducing power usage and energy indulgence had thus become imperative considerations for developing environmentally friendly cloud services [19]. For data centers, major causes of energy inadequacy are the lack of idle power as ICT devices, such as servers, run as long as the processing and storage space is poor in use [20]. The main objective of cloud service providers is to provide a cost-effective and energy-efficient solution for the virtualization of ICT infrastructure for end-user applications following the Service Level Agreement SLA Quality of Service QoS. However, establishing a specific model of energy consumption for VMs remains an open challenge [21].In these devices, however, energy consumption is more desirable and more so. Although the advent of cloud computing has led to massive resource virtualization, due to growing demand for cloud services, their energy cost remains real and rapidly rising [22]. Some of the big challenges of cloud data center days are the amount of electricity consumed in a network to complete the application deployment, and the number of workloads executed to the total energy expended by CDC to execute those workloads. The use of energy can be improved if the amount of mechanisms can be increased or reduced to the dynamic power range.. Must of the researcher primarily committed to optimizing the processor and memory energy consumption. Therefore, to spend this research on covering other workloads, such as storage, network bandwidth etc. at the same time, a great deal is needed to facilitate the request of users. This contest of energy efficient cloud services needs to be solved by the future researcher, this can be done by having a good proposal on resource management policies, algorithms, and architectures.

Therefore, because data centers are powered on the combination of grid and renewable energy, it would be wise for cities to save much of their electricity. Consequently, there is a need to handle both resources and QoS effectively together to provide effective Virtual Machine Placement. Most of the current energy-aware resource management techniques and policies focus primarily only on energy-reduction VM server placement, without considering other resources such as networks, storage, memory, and cooling system which consumes huge amounts of energy. This problem can be solved if the energy consumption and SLAs are handled at the same time. Although researchers are currently doing their best on the issue at hand, more is needed to ensure that the Energy-efficient and Service Level Agreement (SLA) is reached at the same time to reduce operating costs and meet the needs of consumers.

II. Related Work

Cloud computing is growing rapidly, hence the need to look at the data center's cost and efficiency. Service providers draw customers who provide this service with high quality at a lower cost and could be achieved if that physical machine energy is also a bargain in demand to fulfill SLA. Indeed my researcher has already started this data center's energy-efficiency policies.

The author in [23] argued that successful energy management is indeed crucial in cloud data centers and therefore appropriate techniques remain important for energy efficient allocation of VMs.

In [24], the author suggested that a constructive way of consolidating would be primarily to explain the VM placement algorithms and procedures used to find an optimal solution to the VM placement issue. These approaches, whichever minimizes power consumption or provides QoS, might be the biggest conflicting target. Ranking these algorithms or selecting the best one could be a very difficult task to suggest since all other placement approaches have specific targets, such as relocation, resources, and powerful parameters. Although it was suggested that these methods might seem outwardly appropriate, some or the other kind of trade-offs still occur when measured in depth.

The author in [25] proposed that the system, given the significant advantages of cloud computing, is still not mature enough to reach its full potential. The various key challenges this area faces, which include automated resource provisioning, power management, and security management that are just starting to get the attention of the research community. For now, huge potential for researchers to make creative contributions in this field will save substantial impact on the growth of the industry.

Author in [19] suggested that energy efficient allocation of resources remains an open challenge. This where it discusses software and hardware-based techniques. The research proposes a taxonomy aspect namely on objective purpose, allocation process, resource adaptation policy, allocation operation, and interoperability.

Author in [26] explore state-of-the-art techniques for maximizing bandwidth, DVF facilitates power management, server consolidation schemes, and methods for optimizing efficiency across WAN connections. Virtual machine migration work critical through an extensive analysis of existing schemes. To conclude, open research questions and trends in the VM migration domain need to be considered to improve further.

It's said that in [27], the extraordinary impertinence that genuine cloud markets are mostly thousands to millions of dynamically generated and destroyed VMs. No agreed criterion for issues with VMP studies depending on the study. And picking up a test question during experimental research should be useful.

The author in [26] proposed a data management and indexation of the big data taxonomy techniques. The aim is to study the indexing needs of big data for the current state-of -the-art probability indexing techniques by providing researchers with a basis for designing improved solutions for a specific field to support heterogeneity, scalability and accuracy of data as a major concern. The study is based on the precision of collaborative artificial intelligence techniques for extracting information. The proposed method is based on indexing techniques for easy indexing and retrieval, which as the major issue of BD-MCC is acceptable for large size data. The methods provide acceptable data recovery rate and accuracy in the cloud, and end users always use and capture data wherever they are.

In [28], the author applied the method cumulative energy efficiency CEE provides for a direct comparison of servers and IT devices used in data center, taking into account all infrastructures and the different stages in its lifecycle, and various operating conditions. Evaluate maximum energy consumption of data center facilities, establishes a resource metric efficiency that allows a comparison of products throughout their entire life cycle in a data center. Where the result can be used to improve design, operation, and end-of -life strategies for decision makers.

The author in [29] used energy-saving strategies at the data center, with an emphasis on the energy efficiency effect of airflow distribution. Bearing in mind the formation of the thermal environment, multi-scale factors affecting the thermal environment simplify and validate thermal models. This would lead to accurately predict and evaluate the thermal environment and to optimize data center thermal environment have become problems that is important to the data center lifespan.

Author in [14] proposes a Total Energy Management forProfessional Data CentersTEMPRO Analytics framework approach will allow for the preliminary evaluation of the energy efficiency of data centers by means of a visualization with consideration of conformity testing of accredited KPIs. The result shows in some certain areas of data centers, they will be used to optimize overall energy effectiveness. Moreover, it suggested the solution would include different means of visualization for a preliminary evaluation, such as the Sankey diagrams.

III. Taxonomy of Virtual Machine Placement Techniques

There are a number of virtual machine techniques for energy efficiency which have centered in a cloud environment on the subject of energy efficient and resource management. Accordingly, the section contrasts the following dimensions: energy policy adaptation, allocation process and energy usage.

Number of VMP taxonomy for the definition in the literature presented here has been studied. Around 40 research article related to the current study where chosen, with various possible question of formulation. Indeed, selecting the best host to deploy a virtual machine known as VMP is a procedure [30]. These formulations can be either power conscious, or service quality. Consequently, it has also broken down into whether it is an artificially intelligent, non-artificially intelligent or collaborative process of power consumption or service quality.

A. Virtual Machine Placement Policy

Several researchers have tried to work on successful solutions aimed at reducing data center energy consumption while maintaining preferred QoS (Service Quality) [31]. QoS and power saving are two key VM consolidation goals [24]. This type of VM placement method differs from one cloud service provider to another according to the placement target, a VM placement algorithm can generally be divided into two types: Power-based approach with goals of achieving a VM-PM mapping resulting in a system that is energy-efficient with the highest use of resources [32]. While QoS-based a VM placement method varies from one cloud service provider to another.

Indecisions arise from a number of issues that may be resource volume demand (e.g. bandwidth, electricity, and storage space), while failure (e.g., network connection failure and CPU hosting instance failure) and user load configuration (e.g. number of users and location) may occur. The paper is based on the strategy of Virtual Machine Placement Techniques divided into two categories: Power consumption and quality of service as contained in Fig. 1.

Cloud-based Hardware resource status can track e, g. Network, virtual server, and storage while software resources such as application servers, web servers, database servers, etc. all constitute basic functionality and virtual machine placement techniques implementation policies. Monitoring operation includes dynamically profiling the QoS parameters that are connected to the hardware and software resources, the physical resources that are shared while the applications run on them or storing the data. Monitoring services can help to position a virtual machine with respect to: maintaining the energy efficiency level at peak activity for applications and cloud resources, monitoring the energy efficiency and service quality (QoS) provided to the host application, and tracking resource and device failures.

Fig. 2 is a tool that helps you get a perfect understanding of categorizing your organization. Based on the above existing methods, categorized in three categories: NAI, AI, and Hybrid. The latest virtual machine placement strategies that are being analyzed in this survey to see their energy efficiency suitability. Placement strategies are listed in categories as Non-artificial intelligence, artificial intelligence, and hybrid are power consumption and quality of service with subcategories. From the above sub-categorization, AI is based on artificial intelligent power consumption or service quality techniques, for example. Ant colony algorithm, Firefly algorithm, Particle swan algorithm, etc., though NAI is such an algorithm that is not, for example, based on artificial intelligent. Greedy algorithm, Heuristic, Best Fit Decreasing etc. but Hybrid is the combination of artificial and non-artificial intelligent or either called hybrid algorithm.

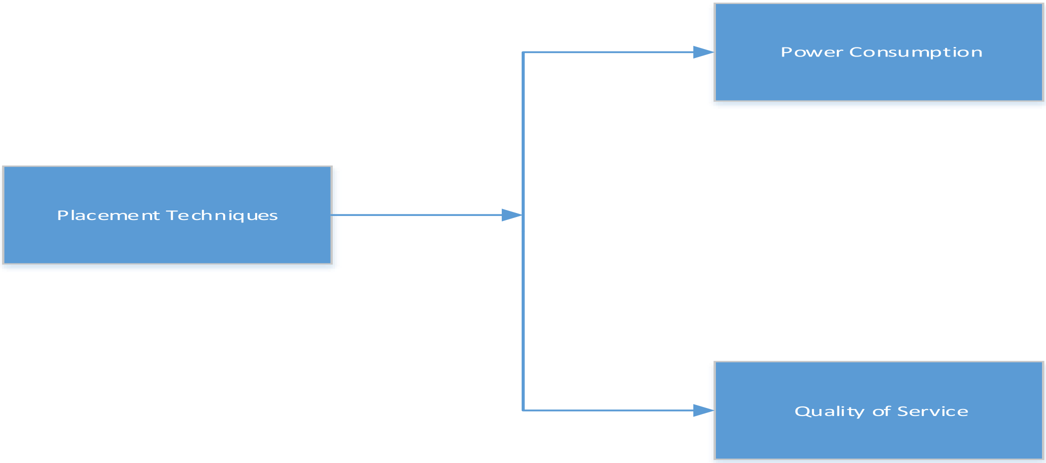


Fig. 1. Types of Placement Techniques.

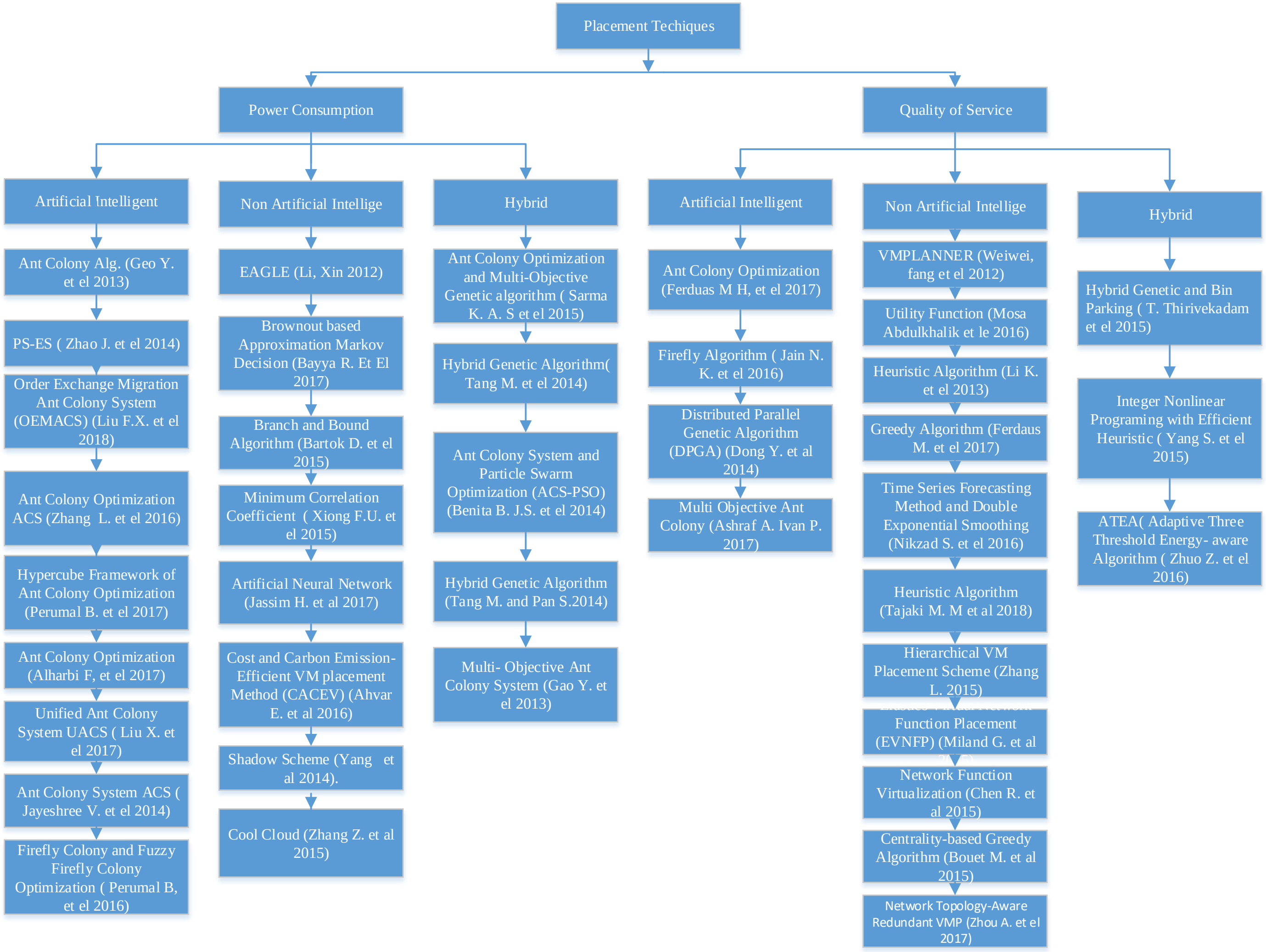

Placement Techiques
Power Consumption
Quality of Service
Artificial Intelligent
Non Artificial Intellige
Hybrid
Artificial Intelligent
Non Artificial Intellige
Hybrid
Ant Colony Alg. (Geo Y. et el 2013)
PS-ES (Zhao J. et el 2014)
Order Exchange Migration Ant Colony System (OEMACS) (Liu F.X. et el 2018)
Ant Colony Optimization ACS (Zhang L. et el 2016)
Hypercube Framework of Ant Colony Optimization (Perumal B. et el 2017)
Ant Colony Optimization (Alharbi F, et el 2017)
Unified Ant Colony System UACS (Liu X. et el 2017)
Ant Colony System ACS (Jayeshree V. et el 2014)
Firefly Colony and Fuzzy Firefly Colony Optimization (Perumal B, et el 2016)
EAGLE (Li, Xin 2012)
Brownout based Approximation Markov Decision (Bayya R. Et El 2017)
Branch and Bound Algorithm (Bartok D. et el 2015)
Minimum Correlation Coefficient (Xiong F.U. et el 2015)
Artificial Neural Network (Jassim H. et al 2017)
Cost and Carbon Emission-Efficient VM placement Method (CACEV) (Ahvar E. et al 2016)
Shadow Scheme (Yang et al 2014).
Cool Cloud (Zhang Z. et al 2015)
Ant Colony Optimization and Multi-Objective Genetic algorithm (Sarma K. A. S et el 2015)
Hybrid Genetic Algorithm(Tang M. et el 2014)
Ant Colony System and Particle Swarm Optimization (ACS-PSO) (Benita B. J.S. et el 2014)
Hybrid Genetic Algorithm (Tang M. and Pan S.2014)
Multi- Objective Ant Colony System (Gao Y. et el 2013)
Ant Colony Optimization (Ferduas M H, et el 2017)
Firefly Algorithm (Jain N. K. et el 2016)
Distributed Parallel Genetic Algorithm (DPGA) (Dong Y. et al 2014)
Multi Objective Ant Colony (Ashraf A. Ivan P. 2017)
VMPLANNER (Weiwei, fang et el 2012)
Utility Function (Mosa Abdulkhalik et le 2016)
Heuristic Algorithm (Li K. et el 2013)
Greedy Algorithm (Ferdaus M. et el 2017)
Time Series Forecasting Method and Double Exponential Smoothing (Nikzad S. et el 2016)
Heuristic Algorithm (Tajaki M. M et al 2018)
Hierarchical VM Placement Scheme (Zhang L. 2015)
Function Placement (EVNFP) (Miland G. et al
Network Function Virtualization (Chen R. et al 2015)
Centrality-based Greedy Algorithm (Bouet M. et al 2015)
Network Topology-Aware Redundant VMP (Zhou A. et el 2017)
Hybrid Genetic and Bin Parking (T. Thirivekadam et el 2015)
Integer Nonlinear Programing with Efficient Heuristic (Yang S. et el 2015)
ATEA(Adaptive Three Threshold Energy- aware Algorithm (Zhuo Z. et el 2016)

Fig. 2. Taxonomy of Placement Techniques.

Artificial intelligence relates to highly technical and specialized techniques which use the knowledge base when placing a virtual machine to deliver energy efficiency and quality of service. This category is similar to ant colony optimization algorithm, particle swarm algorithm, firefly algorithm, genetic algorithm, etc. This is one of the highly effective methods of power consumption and quality of service in a cloud datacenter.

Non-Artificial Intelligence means placing strategies that are transparent in terms of energy use and quality of service. These techniques are developed mostly for the rapid and efficient placement of virtual machine data. They deploy techniques that include Heuristic optimization algorithm, Greedy algorithm, can best fit, optimizing utility function algorithm, etc. Such non-artificial intelligence strategies are classified as the rule-based automated placement applies only cover-known patterns and cannot identify the data center's unknown behavior.

Hybrid Placement based techniques Hybrid develops energy efficiency and service quality. Such approaches combine AI in a data center to obtain a better supporting solution for virtual machine placement indexing. Some of the techniques in this group are Hybrid Genetic Bin Parking Algorithm, Ant Colony System Particle Swarm optimization, ATEA etc. The significant advantage of the hybrid virtual machine placement classifier is flexibility, so it can be applied to any result of classification (Fig. 3 to 6 graphical representation of the techniques).

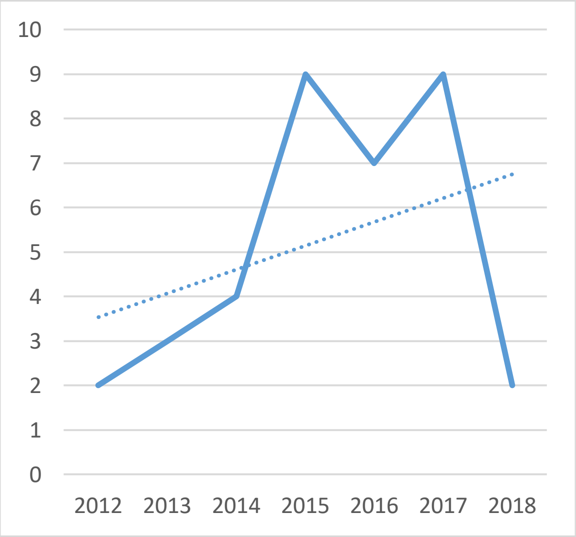


Fig. 3. Virtual Machine Placement Techniques Review Per Year.

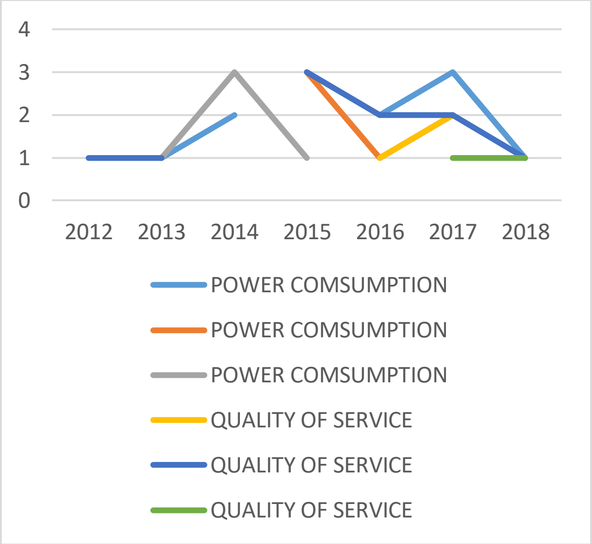


Fig. 4. Per Year Virtual Machine Techniques.

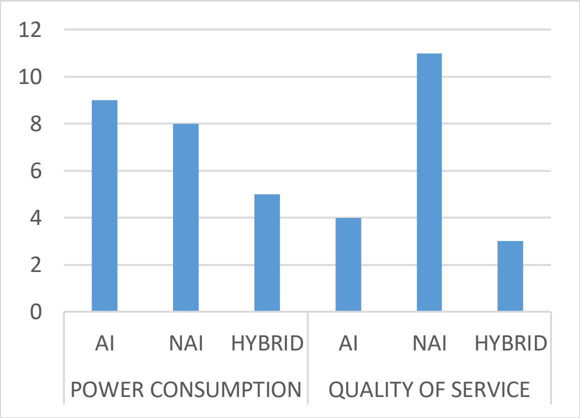


Fig. 5. Types of Virtual Machine Placement Techniques.

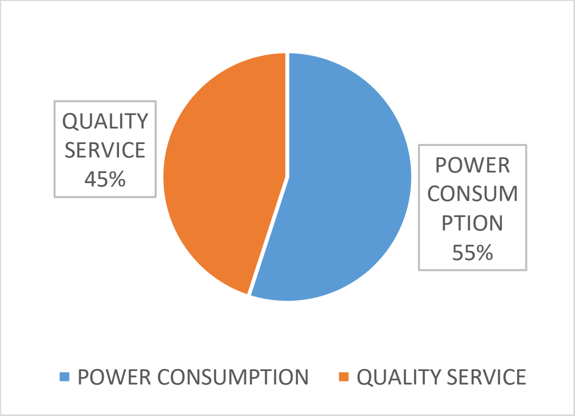


Fig. 6. Virtual Machine Placement Techniques.

IV. Performance Evaluation

As contained in Table I, power-based restructuring allows the efficient use of resources available thereby violating QoS and breaching SLA constraints. Most algorithms therefore find CPU only as their primary resource and should therefore be extended to take into account other important resources which cannot be relegated.

TABLE I. Performance Evaluation

Author/Date	Metrics	Metrics Parameter	Optimization Parameter	Method Applied	Method objective	Open Research Challenge	Method compared
[33]	Power Optimization	Multi-Objective	CPU and Memory	Order Exchange and Migration Ant Colony System (OEMACS)	Resource efficiency and energy saving	Did not consider the quality of service of the virtual machine resources	Heuristic algorithm
[34]	Quality of Service (QoS)	Multi-Objective	CPU and Memory	Unified Ant Colony System (UACS)	Efficient Virtual Machine Migration and Quality of service	The method does not consider network bandwidth	Heuristic algorithm
[35]	Power Optimization	Multi-Objective	CPU and Memory	Ant Colony System (ACS)	Total energy consumption	The method did not consider network bandwidth quality of service of the resources	Compared with existing Ant Colony Algorithm

[36]	Power Optimization	Multi-Objective	CPU and Memory	Firefly Colony and Fuzzy Colony	Energy efficiency and resource wastage	The approach fails to consider the standard of resource service	Better than Heuristic and metaheuristic algorithms
[37]	Power Optimization	Multi-Objective	CPU and Memory	Multi-objective device for ant colony	Waste of resources, and energy consumption	Didn't consider data center combination issues	Multifunctional genetic algorithm, Bin parking algorithm and Ant Colony algorithm max-min
[38]	Power Optimization	Multi-Objective	CPU and Memory	Hypercube Framework Ant Colony System	Power consumption and resource wastage	Did not take care of Quality of Service of resource	Similar to the ant system Max-Min and Ant colony
[39]	Power Optimization	Hybrid	CPU and Memory	Multi-Objective and Ant Colony System	Power efficiency and resource utilization	Not consider network traffic	Compared with Multi-Objective genetic algorithm
[40]	Quality of Service	Multi-Objective	Network bandwidth	VMPlanner	Traffic demand and capacity supply	The algorithm did not Consider resource utilization	NA
[41]	Quality of Service	Multi-Objective	CPU and Memory	EAGLE	Balance the resource utilization, reduce the running PM and lower the energy consumption	The algorithm did not take care of Network communication of the data center	The algorithm outperforms the first-fit algorithm.
[42]	Quality of Service	Multi-Objective	CPU and Memory	Utility function / Genetic algorithm	Energy optimization and SLA	Did not consider Network traffic of the data center	Capered with Heuristic based algorithm
[43]	Power Optimization	Multi-Objective	CPU and Memory	Ant Colony Optimization (ACO) Metaheuristic	Reduce energy consumption, resource wastage, and migration overhead	The algorithm did consider SLA	Compared with ACO metaheuristic
[44]	Quality of Service	Hybrid	CPU and Memory	Hybrid Genetic algorithm	To reduce resource utilization and SLA	The algorithm did not take care of energy consumption efficiently	Compared with algorithms first fit, best fit and round robin.
[45]	Quality of Service	Hybrid	CPU and Memory	Heuristic algorithm	To reduce job completion time	The proposed algorithm does not consider SLA	Compared with the heuristic algorithms of best fit and first fit
[46]	Quality of Service	Multi-Objective	CPU and Network bandwidth	Network and Data-aware Placement (NDAP)	The proposed algorithm aims to reduce energy consumption and improved network performance by reducing delay on the transmission packet.	The proposed algorithm does not consider SLA	NA
[47]	Quality of Service and Power Optimization	Multi-Objective	CPU and Memory	Modified Best Fit Decreasing Algorithm (MBFD) with clustering	The algorithm lower the energy consumption, Service Level Agreement Violation SLAV, and performance degradation	The proposed algorithm did not consider network traffic in the cloud data center	NA

[48]	Power Optimization	Multi-Objective	CPU and Memory	The proposed a custom branch-and-bound algorithm	The algorithm is to improve the effectiveness of cost, application performance, and energy consumption	The algorithm does not take care of network traffic of the data center	Compared with integer linear programming (ILP)
[49]	Quality of Service and Power Optimization	Multi-Objective	CPU and Memory	Adaptive Three-Threshold Energy-Aware Algorithm(ATEA)	The algorithm aims to improve the energy and Service; level agreement effectively	Is not network traffic aware algorithm	NA
[50]	Power Optimization	Multi-Objective	CPU and Memory	The Proposed Integer Nonlinear Programming (INLP)	The algorithm has better in effectiveness in node ratio and performance	The running time is significantly larger than all the heuristics.	The algorithm is compared with ordinary integer nonlinear programming
[51]	Power Optimization	Multi-Objective	CPU and Memory	The proposed Firefly algorithm	The algorithm have shown better energy efficiency and migration	The weakness of the algorithm does not consider the Service Level Agreement	The algorithm is compared with Ant Colony Optimization (ACO) and First Fit Decreasing (FFD) algorithms
[52]	Power Optimization	Hybrid Objective	CPU and Memory	A hybrid genetic algorithm is proposed	The algorithm is aimed at improving efficiency	The communication network is not taken care	Compared with existing heuristic algorithms
[53]	Power Optimization	Multi-Objective	CPU and Memory	A proposed minimum correlation coefficient	The algorithm aims to reduce energy consumption, migration policy and service level agreement in a cloud data environment	The result shows that network traffic is not considered	The result is compared with PABFD
[33]	Power Optimization	Multi-Objective	CPU and Memory	A proposed Order Exchange and Migration Ant Colony System algorithm	The algorithm minimizing the number of active servers, improving resource utilization, balancing different resources, and reducing power consumption.	Service Level Agreement and Network Traffic is not considered	The algorithm is compared with the Heuristic algorithm
[54]	Power Optimization	Multi-Objective	CPU, Memory, and Bandwidth	A Virtual Machine Placement biogeography-based optimization (VMPBBO) algorithm is proposed	The algorithm takes care of power consumption and resource waste at the same time	The Proposed algorithm does not consider Service Level Agreement (SLA)	Compared with Modified General Greedy Algorithm and Virtual Machine Placement Ant Colony System (MGGA and VMPACS)

[55]	Power Optimization	Multi-Objective	CPU and Memory	The proposed Ant Colony System with Particle Swarm Optimization algorithm (ACS-PSO)	The proposed algorithm aim at minimizing resource wastage, minimizing power consumption and for load balancing in physical servers.	The algorithm does not take care of SLA and Network Traffic	compared with the multi-objective and ant colony system algorithm
[56]	Power consumption and Quality of Service	Multi-Objective	Network bandwidth	Network-topology aware redundant VM placement approach to minimizing the consumption of network resources	The proposed approach is a three-step process: host server selection, optimal redundant VM placement, and recovery strategy decision	The algorithm does not take care of SLA	The algorithm where compared with the heuristic algorithm
[57]	Quality of Service	Multi-Objective	Network bandwidth	Proposed near optimal approximation algorithms	Provide bi-criteria solutions reaching constant approximation factors with respect to the overall performance, and adhering to the capacity constraints of the networking infrastructure by a constant factor as well	The algorithm does not take care of SLA	Compared with many realistic algorithms
[58]	Power consumption and Quality of Service	Multi-Objective	CPU and Memory	A Cool Cloud algorithm is proposed	The algorithm provides better power consumption and load balancing	The algorithm does not take care of SLA	The algorithm compared with ILP and Heuristic
[59]	Quality of Service	Multi-Objective	CPU and Memory	A shadow routing based dynamic	The algorithm shows more energy efficiency	The algorithm does not consider network infrastructure	The algorithm is compared with the heuristic algorithm
	Power Consumption	Multi-Objective	CPU, Bandwidth and Response time	Cielo algorithm	The algorithm theoretically ensures that every application has an evolutionarily stable positioning strategy, which is a constant solution under a given workload and availability of resources in a cloud	The algorithm has less SLA consideration	an Algorithm is compared with the Heuristic algorithm

V. Challenges

Cloud computing is still at a timely stage, as the industry generally agrees, where a researcher can work. There are many issues that have not been fully addressed, as the application of industry poses new challenges. Find this section summarizing some of the difficult issues related to cloud computing science. Virtual machine placement algorithms need to tackle various challenges that come from cloud energy efficiency model characteristics. Discussing these issues, and their role in shaping the versatility and convenience offered by cloud environments. The Clouds ' Quality of Service perspective has been well described as a large pool of virtualized resources (such as hardware, platforms for development, and/or services) that are available and accessible only. Cloud computing's main problems are allocating resources to different end-users who have unique resource demands based on their application patterns. Capacity management is either aimed at improving the application's QoS or Energy efficiency, thus improving energy use. QoS seeks to provide optimized parameters which can be measured in terms of space, time, communication delay and budget.

There is a lot of work that has been checked for the purpose of seeing functionality and restricting them when putting VM in the cloud environment. The techniques were divided into a group which included power consumption and service quality. Since power consumption is aimed at a VM-PM mapping that result in an energy-efficient system with the best use of resources, while QoS-based approach is VM-PM mapping that certifies the highest satisfaction of the requirements for service quality. Verily, almost all of the algorithms differ in the goal, but most of them are common in the achievement of power efficiency, while the rest are for service quality. The energy efficiency is highly required in all data center operators to minimize operating costs as well as there is a need to meet user needs where service quality is subject to consumer SLA requirement. One of the hardest tasks is determining which of the sub-techniques is most important.

The energy-efficiency resource management can be described as a major challenge in managing virtualized resource pools effectively for us cloud computing service providers. Physical resources, such as CPU cores, disk space, and network bandwidth need to be cut and shared for virtual machines running potentially heterogeneous workloads. In addition to enhancing productivity processes, integrated resource management can also improve the use of data centers and thus reduce energy usage. Achieving these can be achieved by carefully consolidating workload on a smaller number of servers and turning off unused resources. This research area therefore has an endless opening for the researcher, and is still in need of more.

A. Security

Safety in cloud computing ensures anonymity for the services offered. This model allows data encryption in order to increase the reliability. The' data encryption software' is used to encrypt and decrypt. Service providers are recommended to have the accuracy and reliability of data encryption. Through requesting the key from the key cloud server (KCS), this function can be accomplished through increasing the reliability and the encryption processes. One of the functionality of access control management (ACM) is to approve and authenticate users who access the cloud. Just approved users are allowed to access the cloud to prevent it from attacking. To order to avoid these issues, an intrusion detection program must be used to detect and only allow users to reduce the difficulty.

B. Scalability

These are the device's functionalities for operating as specified. Cloud computing is adapting its cost-effective approach to increasing demands. There are various scaling types available including vertical scaling, horizontal scaling and diagonal scaling [60]. To test the virtual machine (VM) scalability based on the multi-core system workloads. For the VMs message workload control protocol (TCP), scalability is constrained as compared to multiple threads. To order to improve scalability, other computing areas will require a lot of focus, such as memory architecture, network architecture and overhead computing.

C. Data Integrity

Data integrity is the framework that offers scalability, position-independent and a reliable forum for the client. To have data integrity, we need two things which are protection and performance based on public, private key and secret key generation. Confidentiality of information is secured as we encrypt data to prevent unauthorized users [61]. To verify data storage correction and prevent error a universal hash function is required. Recovery is accomplished without mistake, by maintaining confidentiality.

VI. Future Direction

A coal generating carbon emissions which are detrimental to both humans and the environment is the most recent major source of energy production. Energy consumption is a concern that has been widely recognized in the ICT sector throughout the ICT infrastructure such as a datacenter. To have an efficient cloud computing infrastructure, a scalable design was required that could support in particular the reduction of greenhouse gas (GHG) transmissions in and energy consumption. The high increase in ICT resource and its density directly impacts users' spending more on data center infrastructure as well as on cooling and energy management.

The transfer of data removes delays and reduces power consumption, and the contact pattern between CPUs is important to observe. In fact putting CPUs on the same servers, or similar to them, takes a lot of work. In addition to the energy-efficient VM placement algorithms, the application interface can provide different performance levels for end-users. However, in cloud computing, QoS-conscious VM allocation policies also play a major role. A comprehensive study is needed to identify specific patterns of behavior by cloud and distribution of workloads. Further effort is needed to find the relationship between varying workloads, while an effort should be made to create structures that can minimize SLA trade-offs and provide energy efficiency algorithms. Given the increasing deployment of large-scale, complex workflow applications, cloud computing hosts face more critical challenges in reducing consumption without infringing a certain quality of service.

VM placement has done a virtualized datacenter that can be reconfigured by live migration to preserve operational efficiency as the selection of needed VMs changes over time. It is based on the above comparison and taxonomy, it can be understood that in a cloud computing environment, there are specific holes that are wet to be filled and can open up challenges in the field of energy efficiency. The power consumption reports 55 percent in the literature, while 45 percent is reported as service quality, which demonstrates that there is still a great need to look at the power consumption field in order to be able to provide efficient service to customer needs. To maximize customer satisfaction, there is a need to match power consumption with the quality of service in order to have better service delivery. Hybrid algorithms are needed for simplification to facilitate the resolution of many multi-objective problems.

Most of the existing energy-aware resource management approaches and policies focus primarily on VM consolidation to minimize server power consumption only, without considering other resources such as networks, storage, memory, and cooling, which consumes a huge amount of energy. This is one of the big open research problems for the cloud computing community as geographic resource distribution influences network QoS. Unfortunately, the immense amount of simultaneous high-performance data can also consume large amounts of energy. SLAs and QoS t need better energy efficiency at the data center simultaneously to tackle this energy problem. The research community is called upon to do more to ensure energy efficiency and service quality of the cloud data center in order to work towards this direction.

VII. Conclusion

Energy-efficient VM positioning techniques have in years become one of the main research areas at the data center. In defining the power, quality of service, energy in hardware and software, and categorizing existing literature techniques along with a description of their characteristics and constraints. The paper provides the cloud data center with a categorization of current VM positioning strategies and algorithms. The goal is to determine the VM placement requirements for data from cloud data centers and to present a state-of - the-art potential algorithm that would provide researchers with a basis for designing enhanced solutions in a particular domain to provide flexibility, accuracy and scalability to cloud computing.

Additionally, cloud computing will support the business community due to a large number of cloud services users including mobile apps, online gaming, social media, and email. Cloud services need to be made more energy-efficient and sustainable in this respect which can meet consumer demands in a timely manner without affecting the climate. In addition, to enable energy-efficient cloud services, both energy and QoS must be handled jointly.

References

[1] Sultana, A., Using Hadoop to Support Big Data Analysis: Design and Performance Characteristics. 2015.

[2] Wijayaratne, R., et al., Flexible permission management framework for cloud attached file systems, 2016, Google Patents.

[3] Ranger, S., What is cloud computing? Everything you need to know about the cloud, explained, 2018, Retrieved from ZD Net: https://www. zdnet. com/article/what-is-cloud.

[4] Kavis, M.J., Architecting the cloud: design decisions for cloud computing service models (SaaS, PaaS, and IaaS). 2014: John Wiley & Sons.

[5] Weissman, C., et al., Method and system for pushing data to a plurality of devices in an on-demand service environment, 2015, Google Patents.

[6] Verdouw, C.N., et al., Virtualization of food supply chains with the internet of things. Journal of Food Engineering, 2016. 176: p. 128-136.

[7] Saleem, M. and J. Rajouri, Cloud computing virtualization. International Journal of Computer Applications Technology and Research, 2017. 6(7): p. 290-292.

[8] Hu, L., et al., Modeling of cloud-based digital twins for smart manufacturing with MT connect. Procedia manufacturing, 2018. 26: p. 1193-1203.

[9] Wu, X., et al., A task scheduling algorithm based on QoS-driven in cloud computing. Procedia Computer Science, 2013. 17: p. 1162-1169.

[10] Beloglazov, A., J. Abawajy, and R. Buyya, Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future generation computer systems, 2012. 28(5): p. 755-768.

[11] Jiang, L., et al. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. in International Conference on Machine Learning. 2018.

[12] Buyya, R. and S.S. Gill, Sustainable cloud computing: foundations and future directions. arXiv preprint arXiv:1805.01765, 2018.

[13] Sharma, Y., B. Javadi, and W. Si, On the reliability and energy efficiency in cloud computing. Parallel and Distributed Computing, 2015. 27: p. 111.

[14] Gizli, V. and J.M. Gómez, A Framework for Optimizing Energy Efficiency in Data Centers, in From Science to Society. 2018, Springer. p. 275-282.

[15] Pompili, D., A. Hajisami, and T.X. Tran, Elastic resource utilization framework for high capacity and energy efficiency in cloud RAN. IEEE Communications Magazine, 2016. 54(1): p. 26-32.

[16] Gill, S.S. and R. Buyya, A taxonomy and future directions for sustainable cloud computing: 360 degree view. ACM Computing Surveys (CSUR), 2018. 51(5): p. 1-33.

[17] Shuja, J., et al., Sustainable cloud data centers: a survey of enabling techniques and technologies. Renewable and Sustainable Energy Reviews, 2016. 62: p. 195-214.

[18] Ashraf, A., B. Byholm, and I. Porres, Distributed virtual machine consolidation: A systematic mapping study. Computer Science Review, 2018. 28: p. 118-130.

[19] Hameed, A., et al., A survey and taxonomy on energy efficient resource allocation techniques for cloud computing systems. Computing, 2016. 98(7): p. 751-774.

[20] Madni, S.H.H., M.S. Abd Latiff, and Y. Coulibaly, Resource scheduling for infrastructure as a service (IaaS) in cloud computing: Challenges and opportunities. Journal of Network and Computer Applications, 2016. 68: p. 173-200.

[21] Mastelic, T. and I. Brandic, Recent trends in energy-efficient cloud computing. IEEE Cloud Computing, 2015. 2(1): p. 40-47.

[22] Callau-Zori, M., et al., An experiment-driven energy consumption model for virtual machine management systems. Sustainable Computing: Informatics and Systems, 2018. 18: p. 163-174.

[23] Patel, S. and R.M. Makwana, Optimized Energy Efficient Virtual Machine Placement Algorithm and Techniques for Cloud Data Centers. JCS, 2016. 12(9): p. 448-454.

[24] Usmani, Z. and S. Singh, A survey of virtual machine placement techniques in a cloud data center. Procedia Computer Science, 2016. 78: p. 491-498.

[25] Zhang, Q., L. Cheng, and R. Boutaba, Cloud computing: state-of-the-art and research challenges. Journal of internet services and applications, 2010. 1(1): p. 7-18.

[26] Ahmad, R.W., et al., A survey on virtual machine migration and server consolidation frameworks for cloud data centers. Journal of network and computer applications, 2015. 52: p. 11-25.

[27] Pires, F.L. and B. Barán. A virtual machine placement taxonomy. in 2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing. 2015. IEEE.

[28] Peñaherrera, F. and K. Szczepaniak, Development and Application of Metrics for Evaluation of Cumulative Energy Efficiency for IT Devices in Data Centers, in Cascade Use in Technologies 2018. 2019, Springer. p. 142-153.

[29] Jin, C., X. Bai, and C. Yang, Effects of airflow on the thermal environment and energy efficiency in raised-floor data centers: A review. Science of The Total Environment, 2019. 695: p. 133801.

[30] Attaoui, W. and E. Sabir, Multi-criteria virtual machine placement in cloud computing environments: a literature review. arXiv preprint arXiv:1802.05113, 2018.

[31] Zhao, J., et al., A heuristic placement selection of live virtual machine migration for energy-saving in cloud computing environment. PloS one, 2014. 9(9): p. e108275.

[32] Beloglazov, A. and R. Buyya. Energy efficient resource management in virtualized cloud data centers. in 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing. 2010. IEEE.

[33] Liu, X.-F., et al., An energy efficient ant colony system for virtual machine placement in cloud computing. IEEE Transactions on Evolutionary Computation, 2018. 22(1): p. 113-128.

[34] Liu, X.-F., Z.-H. Zhan, and J. Zhang, An Energy Aware Unified Ant Colony System for Dynamic Virtual Machine Placement in Cloud Computing. Energies, 2017. 10(5): p. 609.

[35] Alharbi, F., et al. Profile-Based Ant Colony Optimization for Energy-Efficient Virtual Machine Placement. in International Conference on Neural Information Processing. 2017. Springer.

[36] Perumal, B. and A. Murugaiyan, A firefly colony and its fuzzy approach for server consolidation and virtual machine placement in cloud datacenters. Advances in Fuzzy Systems, 2016. 2016: p. 5.

[37] Gao, Y., et al., A multi-objective ant colony system algorithm for virtual machine placement in cloud computing. Journal of Computer and System Sciences, 2013. 79(8): p. 1230-1242.

[38] Perumal, B. and A. Murugaiyan, Virtual Machine Placement Using Hypercube Ant Colony Optimization Framework.

[39] Sarma, V.A.K., et al., An optimal ant colony algorithm for efficient VM placement. Indian Journal of Science and Technology, 2015. 8(S2): p. 156-159.

[40] Fang, W., et al., VMPlanner: Optimizing virtual machine placement and traffic flow routing to reduce network power costs in cloud data centers. Computer Networks, 2013. 57(1): p. 179-196.

[41] Li, X., et al., Energy efficient virtual machine placement algorithm with balanced and improved resource utilization in a data center. Mathematical and Computer Modelling, 2013. 58(5-6): p. 1222-1235.

[42] Mosa, A. and N.W. Paton, Optimizing virtual machine placement for energy and SLA in clouds using utility functions. Journal of Cloud Computing, 2016. 5(1): p. 17.

[43] Ferdaus, M.H., et al., Multi-objective, Decentralized Dynamic Virtual Machine Consolidation using ACO Metaheuristic in Computing Clouds. arXiv preprint arXiv:1706.06646, 2017.

[44] Li, R., et al., Multi-objective optimization for rebalancing virtual machine placement. Future Generation Computer Systems, 2017.

[45] Li, K., H. Zheng, and J. Wu. Migration-based virtual machine placement in cloud systems. in Cloud Networking (CloudNet), 2013 IEEE 2nd International Conference on. 2013. IEEE.

[46] Ferdaus, M.H., et al., An algorithm for network and data-aware placement of multi-tier applications in cloud data centers. Journal of Network and Computer Applications, 2017. 98: p. 65-83.

[47] Chowdhury, M.R., M.R. Mahmud, and R.M. Rahman, Implementation and performance analysis of various VM placement strategies in CloudSim. Journal of Cloud Computing, 2015. 4(1): p. 20.

[48] Bartók, D. and Z.A. Mann. A branch-and-bound approach to virtual machine placement. in Proceedings of the 3rd HPI Cloud Symposium “Operating the Cloud. 2015.

[49] Zhou, Z., Z. Hu, and K. Li, Virtual machine placement algorithm for both energy-awareness and SLA violation reduction in cloud data centers. Scientific Programming, 2016. 2016: p. 15.

[50] Yang, S., P. Wieder, and R. Yahyapour. Reliable virtual machine placement in distributed clouds. in Resilient Networks Design and Modeling (RNDM), 2016 8th International Workshop on. 2016. IEEE.

[51] Kansal, N.J. and I. Chana, Energy-aware virtual machine migration for cloud computing-a firefly optimization approach. Journal of Grid Computing, 2016. 14(2): p. 327-345.

[52] Tang, M. and S. Pan, A hybrid genetic algorithm for the energy-efficient virtual machine placement problem in data centers. Neural Processing Letters, 2015. 41(2): p. 211-221.

[53] Fu, X. and C. Zhou, Virtual machine selection and placement for dynamic consolidation in Cloud computing environment. Frontiers of Computer Science, 2015. 9(2): p. 322-330.

[54] Zheng, Q., et al., Virtual machine consolidated placement based on multi-objective biogeography-based optimization. Future Generation Computer Systems, 2016. 54: p. 95-122.

[55] Suseela, B.B.J. and V. Jeyakrishnan, A multi-objective hybrid ACO–PSO optimization algorithm for virtual machine placement in cloud computing. Int. J. Res. Eng. Technol, 2014. 3(4): p. 474-476.

[56] Zhou, A., et al., Cloud service reliability enhancement via virtual machine placement optimization. IEEE Transactions on Services Computing, 2017. 10(6): p. 902-913.

[57] Cohen, R., et al. Near optimal placement of virtual network functions. in Computer Communications (INFOCOM), 2015 IEEE Conference on. 2015. IEEE.

[58] Zhang, Z., C.-C. Hsu, and M. Chang. Cool cloud: A practical dynamic virtual machine placement framework for energy aware data centers. in 2015 IEEE 8th International Conference on Cloud Computing (CLOUD). 2015. IEEE.

[59] Guo, Y., S. Stolyar, and A. Walid, Shadow-routing based dynamic algorithms for virtual machine placement in a network cloud. IEEE Transactions on Cloud Computing, 2015.

[60] Hassan, S. and F. Azam. Analysis of cloud computing performance, scalability, availability, & security. in 2014 International Conference on Information Science & Applications (ICISA). 2014. IEEE.

[61] Balusamy, B., et al., Bio-inspired algorithms for cloud computing: a review. International Journal of Innovative Computing and Applications, 2015. 6(3-4): p. 181-202.